Power System Analysis

Analytical tools and structural properties

STEVEN H. LOW

California Institute of Technology
slow@caltech.edu

DRAFT:

May 21, 2025

These are *draft* lecture notes with incomplete sections, missing references, errors and inconsistencies. Corrections, comments, questions will be appreciated - please send them to slow@caltech.edu

Preface

```
The purpose of computing is insight, not numbers.
```
—- Richard W. Hamming, 1962

This book is tailored for students and researchers who are interested in both power systems and analytical tools for understanding their structural properties. It prepares students for research by equipping them with, not only power system knowledge, but also analytical techniques and a way of thinking.

It complements several excellent texts on power system analysis, e.g., [1, 2, 3, 4, 5, 6, 7]. In terms of topic, it develops from scratch basic power system concepts, single-phase and unbalanced three-phase models, and theory and algorithms for power flow optimization. It focuses on steady state modeling and analysis, as opposed to dynamics or electricity markets. In terms of style, it focuses on analytical tools and structural properties. It does not focus on computational issues or specific applications such as state estimation, unit commitment, economic dispatch, or voltage control, but uses these applications to illustrate models and techniques that are widely applicable.

Revision notes

Major changes:

Feb 5, 2022: Corrected error in external models of Δ-configured voltage source and impedance (Chapter 14.3.4)

Feb 12, 2022: Revised Chapter 16.4 on symmetrical components and sequence networks.

April 10, 2022: Revised Chapter 10 on semidefinite relaxations: BIM.

June 26, 2022:

- Added 3-phase transformer section 15.2. Split Chapters on three-phase components (later expanded into Chapter 14 on devices and Chapter 15 on line and transformers) and Chapter 16 on BIM.
- Added Chapter 8.5.2 on Newton-Raphson algorithm and Chapter 8.5.3 on interior-point method.

October 5, 2022:

- Revised Chapter 16.2 on three-phase analysis, especially the solution strategy in Chapter 16.2.3.
- Revised Chapter 16.3 on balanced networks, especially the structural result (Theorem 16.7) in Chapter 16.3.3 and per-phase analysis in Chapter 16.3.4.
- Re-organize Part I to be on Single-phase networks and Part II on Unbalanced multiphase networks.

October 21, 2022 (online version): Revise/re-organize Chapter 4.1.

November 29, 2022 (online version): Revise Chapters 4, 5, 16, 17 to expand line models in BIM and BFM, single and three-phase, to allow general transformer models where series admittances y^s_{kj} and y^s_{kj} may not be equal and admittance matrices Y may not be symmetric (single-phase) or block symmetric (3-phase). Also added a 3-phase BFS for DistFlow model.

January 3, 2023 (online version):

- Revised BFS in Chapter 17.4.2 for 3-phase DistFlow model.
- Revised three-phase OPF formulation in Chapters 9.1 and 9.2.

January 26, 2023 (online version): Revised Chapter 10 on semidefinite relaxations in BIM.

February 7, 2023 (online version)

- Added Chapter **??** semidefinite relaxation of three-phase OPF in BIM.

- Revised Chapters 11 semidefinite relaxation of single and three-phase OPF in BFM.

February 27, 2023: Clarified SVD and corrected mistakes in Takagi factorization in Chapter A.6, as well as the psdudo-inverse of admittance matrix Y in Chapter 4.2.3.

September 20, 2023 (online version):

- Revised Chapter 3.1 with the addition of T equivalent circuit and unitary voltage network models of single-phase transformer. Added Chapter 15.3 on three-phase transformer models with unitary voltage networks.
- Revised invertibility conditions and properties of admittance matrix Y, principal submatrices Y_{22}, and Schur complement Y/Y_{22} for single-phase networks (Chapters 4.2.3 and 4.2.4) and three-phase networks (Chapter 16.1.3).
- Split original chapter on three-phase Component Models into two chapters, Chapter 14 on devices (which absorbs voltage regulators originally in a separate chapter) and Chapter 15 on line and transformers.
- Added Chapter 8 on convex optimization at the beginning of Part II.

April 30, 2024 (online version):

- Chapter 15.1.3: Derive explicitly three-wire models of transmission lines from four-wire model for unbalanced three-phase lines.
- Added Chapter 14.3.2 on a case study of field data from a Southern California utility.
- Added Chapter 8.2 on convex analysis tools, expanded and re-organized Chapter 8.3 on general theory and added Chapter 8.4 on special classes of convex problems.
- Added Chapter 12 on nonsmooth convex optimization that extends theory in Chapters 8.3 and Chapter 8.4 to a nonsmooth setting.

Oct 6, 2024 (online version):

- Chapter 3.2: simplified derivation of external model and per-phase equivalent circuit of balanced three-phase transformers using Corollary 1.3 and (1.32).
- Reorganized Chapter 5 on single-phase BFM to focus on radial networks. Emphasized BFM without assumption C5.1 and allows nonzero shunt admittances (y^m_{jk}, y^m_{kj}) for transformer modeling, and added corresponding linear model in Chapter 5.4.1.
- Moved original Chapter 6 Linear models to become Chapter 4.6.
- Reorganized Chapter 17 on three-phase BFM to focus on radial networks. Added Chapter 17.1.2 on BFM without assumption C17.1 and allows nonzero shunt admittances (y^m_{jk}, y^m_{kj}) for transformer modeling; proved its equivalence and discussed the role of tree topology, cycle conditions and angle recovery in Chapter 17.2. (Still need to prove equivalence for this more general model and add the corresponding linear model.)
- Added Chapter 13 on stochastic OPF.

Oct 16, 2024 (online version): Revised Ch 4 and slides.

Oct 31, 2024 (online version):

- Revised Ch 5 BFM and slides.

- Moved voltage control (originally in BFM chapter) and tree topology identification (originally in BIM chapter) to the new Ch 7; added Ch 6.4 on Economic dispatch.
- Similarly moved the Applications section originally in Ch 13 on Stochastic OPF into the new Ch **??**; revised Ch 13.5 on uncertain Economic dispatch.

Dec 13, 2024 (online version):

- Filled in Chapters 9.3 and 9.4 on the NP-hardness and global optimality of OPF respectively.
- Added (moved from appendix to) Chapter 10.1.6 on chordal relaxation.
- Changed the order of Parts II and III, and moved three-phase OPF and its semidefinite relaxation into a new Chapter 18.
- Simplified and shortened Chapter 9.5 on techniques for scalability.
- Added Chapter 0.1 on suggested courses and Chapter 0.2 on overview.

February 7, 2025 (online version):

- Revised Ch 12 on Nonsmooth convex optimization; added slides.

February 25, 2025 (online version):

- Revised Ch 13 on stochastic OPF; added slides.

April 7, 2025 (online version):

- Re-organized Ch 6 and 7 on power system operations. Also moved security constrained economic dispatch to Ch 6.4.4, added Ch 6.4.5 on security constrained unit commitment and Ch 7.1 on state estimation.

May 21, 2025 (online version):

- Revised Ch 8 on smooth convex optimization; in particular, filled in Ch 8.5.6 on branch and bound, 8.5.7 on Benders decomposition, 8.6.5 on ADMM convergence.

Acknowledgments

Contents

0 Introduction *page* 1

0.1 How to use this book 1

0.2 Overview 3

0.3 Notation 5

0.4 Units 7

Part I Power network: models, operation, analysis 9

1 Basic concepts 11

1.1 Single-phase systems 11

1.1.1 Voltage and current phasors 12

1.1.2 Single-phase devices 13

1.1.3 KVL, KCL, Ohm's Law, Tellegen's theorem 17

1.1.4 One-line diagram and equivalent circuit 23

1.2 Three-phase systems 27

1.2.1 Y and Δ configurations 29

1.2.2 Balanced vectors and conversion matrices $\Gamma, \Gamma^{\mathsf{T}}$ 32

1.2.3 Balanced systems in Y configuration 35

1.2.4 Balanced systems in Δ configuration 42

1.2.5 Per-phase analysis for balanced systems 46

1.2.6 Example configurations and line limits 49

1.3 Complex power 53

1.3.1 Single-phase power 53
1.3.2 Three-phase power 57
1.3.3 Advantages of three-phase power 58
1.4 Bibliographical notes 61
1.5 Problems 61

2 Transmission line models 68
2.1 Line characteristics 68
2.1.1 Series resistance r and shunt conductance g 69
2.1.2 Series inductance l 69
2.1.3 Shunt capacitance c 72
2.1.4 Balanced three-phase line 73
2.2 Line models 74
2.2.1 Transmission matrix 75
2.2.2 Lumped-element Π-circuit model 79
2.2.3 Real and reactive line losses 81
2.2.4 Lossless line 82
2.2.5 Short line 84
2.3 Bibliographical notes 87
2.4 Problems 87

3 Transformer models 93
3.1 Single-phase transformer 93
3.1.1 Ideal transformer 93
3.1.2 Nonideal transformer 94
3.1.3 T equivalent circuit 97
3.1.4 Simplified model 99
3.1.5 Model with unitary voltage network 106
3.2 Balanced three-phase transformers 109
3.2.1 Ideal transformers 109

3.2.2 Nonideal transformers 115
3.3 Equivalent impedance in transformer circuit 117
3.3.1 Transmission matrix 117
3.3.2 Driving-point impedance 119
3.4 Per-phase analysis 122
3.4.1 Analysis procedure 123
3.4.2 Normal system 126
3.5 Appendix: Per-unit normalization 130
3.5.1 Kirchhoff's and Ohm's laws 131
3.5.2 Across ideal transformer 132
3.5.3 Off-nominal transformer 135
3.5.4 Three-phase quantities 136
3.5.5 Per-unit per-phase analysis 140
3.6 Bibliographical notes 141
3.7 Problems 141

4 Bus injection models 152
4.1 Component models 152
4.1.1 Single-phase sources and impedance 152
4.1.2 Single-phase line 153
4.1.3 Single-phase transformer 155
4.2 Network model: IV relation 157
4.2.1 Examples 158
4.2.2 Line model 160
4.2.3 Admittance matrix Y and its properties 162
4.2.4 Kron reduction Y/Y_{22} and its properties 172
4.2.5 Solving $I = YV$ 180
4.2.6 Radial network 182
4.2.7 Summary 185
4.3 Network models: sV relation 185

4.3.1 Complex form 186
4.3.2 Polar form 187
4.3.3 Cartesian form 188
4.3.4 Types of buses 188
4.4 Computation methods 189
4.4.1 Gauss-Seidel algorithm 189
4.4.2 Newton-Raphson algorithm 192
4.4.3 Fast decoupled algorithm 194
4.4.4 Holomorphic Embedding Load-flow Method (HELM) 195
4.5 Properties of power flow solutions 199
4.6 Linear power flow model 200
4.6.1 Laplacian matrix L 200
4.6.2 DC power flow model 205
4.6.3 Distribution factors 209
4.7 Bibliographical notes 209
4.8 Problems 210

5 Branch flow models: radial networks 217
5.1 BFM for radial networks 217
5.1.1 Line model 217
5.1.2 With shunt admittances 218
5.1.3 Without shunt admittances 221
5.1.4 Angle recovery 225
5.1.5 Power flow solutions 226
5.2 Equivalence 230
5.2.1 Extension to general networks 230
5.2.2 Equivalence of BFM and BIM 231
5.3 Backward forward sweep 235
5.3.1 General BFS 235

5.3.2 Complex form BFM 240
5.3.3 DistFlow model 242
5.4 Linear power flow models 244
5.4.1 With shunt admittances 244
5.4.2 Without shunt admittances 246
5.4.3 Linear solution and its properties 246
5.5 Bibliographical notes 250
5.6 Problems 251

6 System operation: power balance 254
6.1 Background 254
6.1.1 Overview 254
6.1.2 Basic optimization concepts 256
6.2 Unit commitment and real-time dispatch 258
6.2.1 Unit commitment 258
6.2.2 Real-time dispatch 260
6.2.3 Security constrained OPF 264
6.3 Frequency control 267
6.3.1 Assumptions and notations 268
6.3.2 Primary control 271
6.3.3 Secondary control 276
6.4 Pricing electricity and reserves 278
6.4.1 DC power flow model 278
6.4.2 Economic dispatch and LMP 279
6.4.3 LMP properties 283
6.4.4 Security constrained economic dispatch 290
6.4.5 Security constrained unit commitment 294
6.5 Bibliography 295
6.6 Problems 296

7 System operation: estimation and control 299

7.1 State estimation 299

7.2 Volt/var control on radial networks 302

7.2.1 Linear DistFlow model 302

7.2.2 Decentralized control: convergence and optimality 303

7.3 Tree topology identification 308

7.3.1 Linearized polar-form AC model 308

7.3.2 Covariance of voltage magnitudes and powers 309

7.3.3 Graphical-model method 312

7.4 Bibliographical notes 312

7.5 Problems 312

Part II Power flow optimization 317

8 Smooth convex optimization 319

8.1 Convex optimization 320

8.1.1 Affine hull and relative interior 320

8.1.2 Convex set 321

8.1.3 Derivative, directional derivative and partial derivative 323

8.1.4 Convex function 325

8.1.5 Convex program 335

8.2 Properties of convex sets and convex cones 338

8.2.1 Second-order cone K_{soc} in $\mathbb{R}^n$ 338

8.2.2 Semidefinite cone K_{psd} in $\mathbb{S}^n$ 342

8.2.3 Projection theorem 344

8.2.4 Separating hyperplanes 345

8.2.5 Farkas Lemma 347

8.3 General theory: optimality conditions 349

8.3.1 Characterization: saddle point = p-d optimality + strong duality 351

8.3.2 Characterization: KKT point = saddle point 355

8.3.3 Existence: primal optimal solutions 358
8.3.4 Existence: dual optimal solutions and constraint qualifications 359
8.3.5 Perturbed problem and local sensitivity 363
8.3.6 Envelope theorems 365
8.3.7 Equivalent representations 369
8.4 Special convex programs 371
8.4.1 Summary: general method 371
8.4.2 Linear program (LP) 374
8.4.3 Convex quadratic program (QP) 378
8.4.4 Second-order cone program (SOCP) 380
8.4.5 Semidefinite program (SDP) 384
8.5 Optimization algorithms 387
8.5.1 Steepest descent algorithm 388
8.5.2 Newton-Raphson algorithm 390
8.5.3 Interior-point algorithm 394
8.5.4 Dual and primal-dual gradient algorithms 399
8.5.5 Alternating direction method of multipliers (ADMM) 402
8.5.6 Branch and bound 404
8.5.7 Benders decomposition 408
8.6 Convergence analysis 414
8.6.1 Convergence theorems 414
8.6.2 Gauss-Seidel algorithm 417
8.6.3 Steepest descent algorithm 421
8.6.4 Interior-point algorithm 424
8.6.5 ADMM 425
8.7 Bibliographical notes 430
8.8 Problems 430

9 Optimal power flow 441
9.1 Bus injection model 441

9.1.1 Single-phase devices 442
9.1.2 Single-phase OPF 442
9.1.3 OPF as QCQP 447
9.2 Branch flow model: radial networks 453
9.3 NP-hardness 455
9.3.1 OPF feasibility on a tree network 456
9.3.2 OPF is NP-hard 457
9.3.3 Proof of Theorem 9.1 460
9.4 Global optimality: Lyapunov-like condition 462
9.4.1 Convex relaxation 462
9.4.2 Conditions for global optimality 463
9.4.3 Proof of Theorem 9.2 465
9.4.4 Application to OPF on radial network 470
9.5 Techniques for scalability: case study 474
9.5.1 SCOPF formulation 474
9.5.2 Handling nonsmoothness 479
9.5.3 Scaling computation 484
9.6 Bibliographical notes 488
9.7 Problems 489

10 Semidefinite relaxations: BIM 496
10.1 Semidefinite relaxations of QCQP 497
10.1.1 SDP relaxation 497
10.1.2 Partial matrices and rank-1 completion 498
10.1.3 Feasible sets 502
10.1.4 Semidefinite relaxations and solution recovery 504
10.1.5 Tightness of relaxations 505
10.1.6 Chordal relaxation 506
10.1.7 Strong SOCP relaxations: mesh network 510
10.1.8 Proofs 510

10.2 Application to OPF 512
10.2.1 Semidefinite relaxations 512
10.2.2 Exact relaxation: definition 517
10.3 Exactness condition: linear separability 518
10.3.1 Sufficient condition for QCQP 518
10.3.2 Application to OPF 519
10.3.3 Proofs 521
10.4 Exactness condition: small angle differences 523
10.4.1 Sufficient condition 524
10.4.2 Proof: 2-bus network 525
10.5 Other convex relaxations 530
10.6 Bibliographical notes 530
10.7 Problems 531

11 Semidefinite relaxations: BFM 537
11.1 SOCP relaxation 537
11.1.1 DistFlow model 537
11.1.2 Equivalence 539
11.1.3 General radial network 543
11.2 Exactness condition: inactive injection lower bounds 544
11.2.1 DistFlow model 544
11.2.2 General radial network 547
11.3 Exactness condition: inactive voltage upper bounds 547
11.3.1 Sufficient condition 548
11.3.2 Appendix: Proof of Theorem 11.5 551
11.4 Bibliographical notes 558
11.5 Problems 559

12 Nonsmooth convex optimization 561
12.1 Normal cones of feasible sets 562

12.1.1 Polar cone 563
12.1.2 Normal cone and tangent cone 564
12.1.3 Affine transformation 575
12.1.4 Second-order cones and SOC constraints 583
12.2 CPC functions 586
12.2.1 Extended real-valued function 587
12.2.2 Indicator function, support function and polyhedral functions 589
12.3 Gradient and subgradient 591
12.3.1 Derivative, directional derivative and partial derivative 591
12.3.2 Subgradient 592
12.3.3 Subdifferential calculus 597
12.4 Characterization: saddle point = p-d optimality + strong duality 603
12.5 Characterization: generalized KKT condition 604
12.6 Existence: primal optimal solutions 606
12.7 Existence: dual optimal solutions and strong duality 609
12.7.1 Slater Theorem 610
12.7.2 MC/MC problems 612
12.7.3 Slater Theorem 12.28: proof 617
12.8 Special convex programs 621
12.8.1 Summary: general method 621
12.8.2 Linear program (LP) 622
12.8.3 Second-order cone program (SOCP) 624
12.8.4 Conic program and convex inequality 627
12.9 Bibliographical notes 630
12.10Problems 630

13 Stochastic OPF 636

13.1 Robust optimization 637
13.1.1 General formulation 637

13.1.2 Robust linear program 643
13.1.3 Robust second-order cone program 646
13.1.4 Robust semidefinite program 652
13.1.5 Appendix: proof of S-lemma 656
13.2 Chance constrained optimization 658
13.2.1 Tractable instances: convexity, strong duality and optimality 659
13.2.2 Concentration inequalities and safe approximation 666
13.3 Convex scenario optimization 679
13.3.1 Violation probability $V(x_N^*)$ 681
13.3.2 Proof: bound on $E^N(V(x_N^*))$ 686
13.3.3 Proof: bound on $\mathbb{P}^N(V(x_N^*) > \epsilon)$ for uniformly supported problem 691
13.3.4 Proof: bound on $\mathbb{P}^N(V(x_N^*) > \epsilon)$ for general problem 696
13.3.5 Sample complexity 703
13.3.6 Optimality guarantee 704
13.4 Two-stage optimization with recourse 708
13.4.1 Stochastic linear program with fixed recourse 708
13.4.2 Stochastic nonlinear program with general recourse 718
13.5 Example application: stochastic economic dispatch 720
13.5.1 Nominal ED 721
13.5.2 Robust ED 722
13.5.3 Chance constrained ED 724
13.5.4 Scenario-based ED 724
13.5.5 Special case: no congestion 725
13.6 Example application: security constrained unit commitment 728
13.6.1 Two-stage adaptive robust formulation 728
13.6.2 Solution 729
13.7 Bibliographical notes 731
13.8 Problems 731

Part III Unbalanced three-phase networks 739

14 Component models, I: devices 741

14.1 Overview 741

14.1.1 Internal and terminal variables 742

14.1.2 Three-phase device models 744

14.1.3 Three-phase line and transformer models 745

14.1.4 Three-phase network models 746

14.1.5 Balanced operation 747

14.2 Mathematical properties of three-phase network 748

14.2.1 Pseudo-inverses of $\Gamma, \Gamma^{\mathsf{T}}$. 748

14.2.2 Similarity transformation and symmetrical components 750

14.3 Three-phase device models 753

14.3.1 Conversion rules 753

14.3.2 Case study: Riverside CA utility 758

14.3.3 Devices in *Y* configuration 762

14.3.4 Devices in Δ configuration 769

14.3.5 Δ-*Y* transformation 780

14.3.6 Comparison with single-phase devices 781

14.3.7 Summary 784

14.4 Voltage regulators 785

14.5 Bibliographical notes 785

14.6 Problems 786

15 Component models, II: line and transformers 791

15.1 Three-phase transmission or distribution line models 791

15.1.1 Review: single-phase model 791

15.1.2 Four-wire three-phase model 792

15.1.3 Three-wire three-phase model 794

15.1.4 Ideal voltage and current sources 800

15.2 Three-phase transformer models: simplified circuit 801
15.2.1 Review: single-phase transformer 802
15.2.2 General derivation method 804
15.2.3 Three-phase Π circuit, block symmetry, symmetry 809
15.2.4 YY configuration 810
15.2.5 $\Delta\Delta$ configuration 813
15.2.6 ΔY configuration 815
15.2.7 $Y\Delta$ configuration 817
15.2.8 Open transformer 818
15.2.9 Single-phase equivalent in balanced setting 821
15.3 Three-phase transformer models: unitary voltage network 824
15.3.1 Internal model: UVN per phase 824
15.3.2 Conversion rules 825
15.3.3 External model 826
15.3.4 Split-phase transformer 829
15.4 Parameter identification: examples 829
15.4.1 Simplified circuit 829
15.4.2 Unitary voltage network 833
15.5 Bibliographical notes 833
15.6 Problems 834

16 Bus injection models 836
16.1 Network models 836
16.1.1 Line model 837
16.1.2 IV relation 839
16.1.3 Invertibility of Y, Y_{22} and Y/Y_{22} 844
16.1.4 sV relation 852
16.1.5 Overall model 852
16.2 Three-phase analysis 853
16.2.1 Examples 853

16.2.2 General analysis problem 873
16.2.3 Solution strategy 876
16.3 Balanced network 884
16.3.1 Kronecker product 885
16.3.2 Three-phase analysis 885
16.3.3 Balanced voltages and currents 889
16.3.4 Phase decoupling and per-phase analysis 894
16.4 Symmetric network 899
16.4.1 Sequence impedances 900
16.4.2 Sequence voltage sources 903
16.4.3 Sequence current sources 906
16.4.4 Sequence line model 909
16.4.5 Three-phase analysis 910
16.5 Bibliographical notes 915
16.6 Problems 916

17 Branch flow models: radial networks 925
17.1 Three-phase BFM for radial networks 925
17.1.1 Line model 925
17.1.2 With shunt admittances 926
17.1.3 Without shunt admittances 928
17.2 Equivalence, cycle condition and angle recovery 931
17.2.1 Extension to general networks 931
17.2.2 Equivalence of BFM and BIM 932
17.2.3 Tree topology, cycle condition, angle recovery 934
17.3 Overall model and examples 941
17.3.1 Overall model 941
17.3.2 Examples 942
17.4 Backward forward sweep 947

17.4.1 Complex form BFM 947
17.4.2 DistFlow model 952
17.5 Linear model 952
17.5.1 Three-phase LinDistFlow 952
17.5.2 Application example 955
17.6 Bibliographical notes 955
17.7 Problems 955

18 Power flow optimization 957
18.1 Three-phase OPF 957
18.1.1 Three-phase devices 958
18.1.2 Bus injection model 960
18.1.3 Three-phase OPF as QCQP 963
18.1.4 Branch flow model: radial networks 966
18.2 Semidefinite relaxation: BIM 969
18.2.1 Reformulation 969
18.2.2 SDP relaxation 973
18.2.3 Radial network 975
18.3 Semidefinite relaxation: BFM 977
18.3.1 Reformulation 978
18.3.2 Semidefinite relaxation 978
18.4 Example applications 980
18.5 Bibliographical notes 980
18.6 Problems 981

Appendix Linear algebra preliminaries 983
A.1 Vector spaces, basis, rank, nullity 983
A.1.1 Vector spaces, subspaces, span 983
A.1.2 Basis, dimension, rank and nullity 985
A.2 Polyhedral set and extreme point 987

A.3 Schur complement and matrix inversion formula 988
A.3.1 Schur complement 988
A.3.2 Matrix inversion lemma 991
A.4 Change of basis, diagonalizability, Jordan form 992
A.4.1 Similarity transformation 992
A.4.2 Diagonalizabilty and Jordan form 993
A.5 Special matrices 995
A.6 SVD, spectral decompositions, complex symmetric matrices 998
A.6.1 Singular value decomposition for any matrix 998
A.6.2 Spectral decomposition for normal matrices 1003
A.6.3 SVD and unitary diagonalization 1005
A.6.4 Complex symmetric matrices 1006
A.7 Pseudo-inverse 1008
A.8 Norms and inequalities 1014
A.8.1 Vector norms 1014
A.8.2 Cauchy-Schwarz inequality, Hölder's inequality, dual norm 1017
A.8.3 Matrix norms 1020
A.9 Differentiability, complex differentiability, analyticity 1025
A.10 Mean value theorems 1028
A.11 Algebraic graph theory 1030
A.12 Bibliographical notes 1035
A.13 Problems 1035

Bibliography 1042

Index 1054

0 Introduction

0.1 How to use this book

This book can be used as a research reference. It can also be used as a textbook and we suggest possible courses that can be constructed from this book.

Power System Analysis I: models and operation

A 13-week course for senior undergraduate and beginning graduate students that covers Part I of the book. It develops from scratch single-phase network models and formulates optimal power flow problems. These models are then used to describe and analyze power system operation such as mechanisms for balancing power, controlling frequency, pricing electricity and reserves, estimating state, and stabilizing voltages. This course does not require prior power system knowledge or optimization theory, but does require linear algebra and interest in or exposure to mathematical analysis.

Specifically it covers

1 *Basic concepts*: Kirchhoff's laws, phasors, device models, three-phase systems, complex power (Chapter 1).
2 *Component models*: transmission line (Chapter 2), transformers (Chapters 3, possibly skipping Chapter 3.1.5).
3 *Network models*: bus injection models (Chapter 4, possibly skipping Chapter 4.4.4), and 4.5), branch flow models (Chapter 5, possibly skipping Chapter 5.3).
4 *Power system operation, I*: control mechanisms for balancing power, including unit commitment, real-time dispatch, secure operation, and primary and secondary frequency control, as well as market mechanisms for pricing electricity and reserves using locational marginal prices (Chapter 6).
5 *Power system operation, II*: state estimation, voltage control on distribution networks, and network topology identification (Chapter 7).

Power System Analysis II: power flow optimization

A 13-week graduate course that covers Part II of the book on power flow optimization. It focuses on analytical tools for and structural properties of power systems and prepares students for research.

1 *Power system basics*: Reviews models and basic operation of power systems (topics from Chapters 4, 5, 6 depending on students' prior knowledge).
2 *Convex optimization*: convex analysis, optimality conditions, special convex programs, optimization algorithms, convergence analysis (Chapters 8).
3 *Optimal power flow*: OPF in BIM and BFM, NP-hardness, global optimality, techniques for scalability (Chapter 9).
4 *Semidefinite relaxations of OPF*: SDP, chordal, SOCP relaxations of OPF, exactness conditions (Chapters 10 and 11).
5 *Nonsmooth convex optimization*: normal cones and feasible sets, CPC functions and subgradients, optimality conditions, special convex programs (Chapter 12).
6 *Stochastic OPF*: robust optimization, chance constrained optimization, convex scenario program, two-stage optimization with recourse (Chapter 13).

Unbalance Three-phase Power System.

A 10-week undergraduate/graduate course that covers Part III of the book on unbalanced three-phase networks. It develops from scratch three-phase component and network models, three-phase optimal power flow and its semidefinite relaxations. It shows how models and analysis for single-phase networks extend directly to a three-phase setting. Prior knowledge of single-phase power networks or optimization theory will be helpful but not absolutely necessary.

1 *Review*: Single-phase power networks (topics from Chapters 4 and 5 depending on students' prior knowledge).
2 *Component models*: mathematical properties of three-phase systems, three-phase devices in Y and Δ configurations, three-phase transmission or distribution lines, three-phase transformers (Chapters 14 and 15).
3 *Bus injection model*: network model, three-phase analysis, balanced network (Chapter 16, possibly skipping Chapter 16.4).
4 *Branch flow model*: network model, equivalence, examples, linear model (Chapter 17, possibly skipping Chapter 17.4).
5 *Review*: basic convex optimization theory and algorithms (topics from Chapter 8 depending on students' prior knowledge).
6 *Power flow optimization*: three-phase OPF, semidefinite relaxations, example applications (Chapter 18).

0.2 Overview

The book consists of three parts and an appendix.

Part I Power network: models, operation, analysis

1 *Chapter 1* introduces basic concepts in modeling the steady-state behavior of an alternating current (AC) power system, including circuit models, Kirchhoff's laws, phasor representation, balanced three-phase systems, per-phase equivalent, and complex power.
2 *Chapter 2* develops circuit models for the terminal behavior of a balanced three-phase transmission line that map the voltage and current at one end of the line to those at the other end.
3 *Chapter 3* develops models for balanced three-phase transformers and their per-phase equivalent and analysis techniques for circuits containing transformers, including per-unit normalization.
4 *Chapter 4* uses the component models of previous chapters to construct a class of network models we call the *bus injection model* (BIM). It introduces the network admittance matrix Y that relates linearly bus voltages and current injections, its Kron reduction, and their analytical properties. It also introduces power flow equations that relate nonlinearly bus voltages and power injections and presents iterative algorithms for solving these equations. Finally it introduces a linearized power flow model called the DC power flow model that is widely used for electricity market operation.
5 *Chapter 5* introduces the *branch flow model* (BFM) for radial networks with a tree topology and proves its equivalence to the bus injection model. It presents a fast iterative algorithm called the backward forward sweep for solving power flow equations for radial networks. Finally it introduces a linearized model that admits an explicit solution and bounds nonlinear power flow solutions.
6 *Chapter 6* overviews three control mechanisms at different timescales, unit commitment, real-time dispatch and frequency control, that balance power supply and demand. It also studies pricing of electricity and reserves using locational marginal prices and optimality properties of these prices.
7 *Chapter 7* illustrates the models and tools developed in earlier chapters through three applications: state estimation, voltage control on distribution networks, and topology identification.

Part II Power flow optimization

1 *Chapter 8* formulates convex optimization problems and introduces some of the most useful tools for convex analysis. We develop a general theory to characterize optimal solutions and provide sufficient conditions for their existence, and then apply the general theory to special classes of convex optimization problems

widely used in applications. We describe iterative algorithms for solving convex optimization problems and basic techniques for analyzing their convergence.

2 *Chapter 9* formulates optimal power flow (OPF) problems that underly numerous power system applications, in both the bus injection model and the branch flow model. It proves that OPF is NP-hard but a subclass characterized by a Lyapunov-like condition can be solved efficiently to global optimality. Finally it describes common techniques for scaling OPF solutions.

3 *Chapter 10* studies the semidefinite relaxation of the nonconvex OPF problem formulated in BIM as a quadratically constrained quadratically program. It develops the concept of partial matrices and their positive semidefinite rank-1 completion to exploit the sparsity of large networks. Finally it proves two sufficient conditions for exact second-order cone (SOCP) relaxations of OPF on single-phase radial networks. Convex relaxation complements linear approximation and local iterative algorithms as one of the main tools for dealing with the nonconvexity of OPF.

4 *Chapter 11* studies the semidefinite relaxation of OPF in BFM for radial networks. It formulates SOCP relaxation and proves its equivalence to the SOCP relaxation in BIM. Finally it proves two sufficient conditions for exact SOCP relaxation for single-phase radial networks.

5 *Chapter 12* generalizes the structural results of Chapter 8.3 to a convex but nonsmooth setting, motivated by stochastic OPF studied in Chapter 13. It shows that convexity is fundamental, but not smoothness, and, once the basic framework is established, the more abstract approach here that relies only on convexity is both more natural and simpler conceptually.

6 *Chapter 13* studies basic methods for stochastic optimization, robust optimization, chance constrained optimization, scenario programming, and two-stage optimization with recourse. A focus is on problems (e.g., two-stage optimization) that are convex, but often nonsmooth, to which optimality conditions studied in Chapter 12 are applicable and computation algorithms studied in Chapter 8 can be adapted by replacing gradients with subgradients. Finally we present examples to illustrate concepts of stochastic OPF.

Part III: Unbalanced three-phase networks

1 *Chapter 14* studies the mathematical properties that underly the behavior of unbalanced three-phase systems and derive models of three-phase voltage sources, current sources, power sources, and impedances in Y and Δ configurations.

2 *Chapter 15* derives models of three-phase lines and transformers.

3 *Chapter 16* uses the component models of Chapters 14 and 15 to extend the bus injection model to the unbalanced three-phase setting. It also introduces the sequence coordinate in which sequence networks become decoupled when there is a certain symmetry in the original phase coordinate.

4 *Chapter 17* extends the branch flow model to the unbalanced three-phase setting.

5 *Chapter 18* extends OPF and its semidefinite relaxations (studied in Chapters 9, 10, 11) from single-phase to unbalanced three-phase networks.

Appendix: Linear algebra preliminaries

Appendix A collects mathematical preliminaries used in the rest of the book.

0.3 Notation

Let $\mathbb{C}$ denote the set of complex numbers, $\mathbb{R}$ the set of real numbers, $\mathbb{R}_+$ the set of nonnegative real numbers, $\mathbb{R}_-$ the set of nonpositive real numbers, $\mathbb{N}$ the set of integers and $\mathbb{N}_+$ the set of positive integers. We use $\mathbf{i}$ to denote $\sqrt{-1}$. For $a \in \mathbb{C}$, its real and imaginary parts are denoted by $\mathrm{Re}\, a$ and $\mathrm{Im}\, a$ respectively. Its complex conjugate is usually denoted by $\bar{a}$ or a^{H} (though $\bar{x}$ also denotes a particular vector in $\mathbb{R}^n$ when it is clear from the context that $\bar{x}$ is a real quantity). For any set $A \subseteq \mathbb{C}^n$, $\mathrm{conv}\, A$ denotes the convex hull of A. For $a \in \mathbb{R}$, $[a]^+ := \max\{a, 0\}$. For $a, b \in \mathbb{C}$, $a \le b$ means $\mathrm{Re}\, a \le \mathrm{Re}\, b$ and $\mathrm{Im}\, a \le \mathrm{Im}\, b$. We sometimes abuse notation to use the same symbol a to denote either a complex number $\mathrm{Re}\, a + \mathbf{i}\,\mathrm{Im}\, a$ or a size 2 real vector $a = (\mathrm{Re}\, a, \mathrm{Im}\, a)$ depending on the context. The empty set is denoted $\emptyset$.

In general scalar or vector variables are in small letters, e.g. u, w, x, y, z. Most power system quantities however are in capital letters, e.g. $S_{jk}, P_{jk}, Q_{jk}, I_j, V_j$. Unless otherwise specified, a vector is a column vector and is written interchangeably as

$$V = \begin{bmatrix} V_a \\ V_b \\ V_c \end{bmatrix} \text{ or } V = (V_a, V_b, V_c)$$

A variable without a subscript usually denotes a vector with appropriate components, e.g. $s := (s_j, j = 0, \ldots, n)$, $S := (S_{jk}, (j,k) \in E)$. For a vector $a = (a_1, \ldots, a_k)$, a_{-i} denotes $(a_1, \ldots, a_{i-1}, a_{i+1}, a_k)$ without the a_i entry. For a subset $A \subseteq \{1, \ldots, k\}$, $a_{-A} := (a_i, i \notin A)$. For vectors x, y, $x \le y$ denotes componentwise inequality. We freely refer to x as singular if we mean the vector x or as plural if we mean its components $x_1, \ldots, x_n$. For example we may refer to λ^* as a locational marginal price or locational marginal prices.

Matrices are usually in capital letters. Let M, N be index sets with $m := |M|$, $n := |N|$. An $m \times n$ matrix with $a_{ij} \in \mathbb{C}$ as its (i,j)-th entry for $i \in M, j \in N$, can be written as $A = (a_{ij}, i \in M, j \in N)$. Given $k := \min\{m, n\}$ and scalars $a_1, \ldots, a_k$, $\mathrm{diag}(a_1, \ldots, a_k)$ is a $k \times k$ diagonal matrix with a_i on its diagonal. Given an $m \times n$ matrix A, $\mathrm{diag}(A) := \mathrm{diag}(A_{11}, \ldots, A_{kk})$. We use $\bar{A}$ to denote the componentwise complex conjugate of a matrix A. The transpose of a matrix A is denoted by A^{T} and its Hermitian (or conjugate) transpose by $A^{\mathsf{H}} := \bar{A}^{\mathsf{T}}$. If a is a scalar then $a^{\mathsf{H}} = \bar{a}$ is its complex conjugate. We use interchangeably $(y^s)^{\mathsf{H}}$ and $y^{s\mathsf{H}}$. A matrix A is Hermitian if $A = A^{\mathsf{H}}$. A complex matrix A is positive semidefinite (or psd), denoted by $A \succeq 0$, if A is Hermitian *and* $x^{\mathsf{H}} A x \ge 0$ for all $x \in \mathbb{C}^n$. A real matrix A is positive semidefinite (or psd), denoted by $A \succeq 0$, if A is symmtric *and* $x^{\mathsf{T}} A x \ge 0$ for all $x \in \mathbb{R}^n$. In particular if

$A \succeq 0$ then by definition $A = A^{\mathsf{H}}$ if A is complex and $A = A^{\mathsf{T}}$ if A is real. [1] A is negative semidefinite (nsd) if $-A$ is psd. For matrices A, B, $A \succeq B$ means $A - B$ is psd. Let $\mathbb{S}^n$ be the set of all $n \times n$ Hermitian matrices, $\mathbb{S}^n_+$ the set of $n \times n$ psd matrices, and $\mathbb{S}^n_-$ the set of $n \times n$ nsd matrices.

A graph $G = (N, E)$ consists of a set N of nodes and a set $E \subseteq N \times N$ of edges. If G is undirected then $(j,k) \in E$ if and only if $(k,j) \in E$. If G is directed then $(j,k) \in E$ only if $(k,j) \notin E$; in this case we will use (j,k) and $j \to k$ interchangeably to denote an edge pointing from j to k. Therefore, for an undirected graph, $\sum_{(j,k)\in E} x_{jk}$ includes both x_{jk} and x_{kj} for each edge $(j,k) \in E$, whereas, for a directed graph, $\sum_{(j,k)\in E} x_{jk}$ includes a single term x_{jk} for each directed edge $j \to k$. Sometimes, we write $\sum_{(j,k)\in E} \left(x_{jk} + x_{kj}\right)$ instead of $\sum_{(j,k)\in E} x_{jk}$ to emphasize the undirected nature of the graph. By "$j \sim k$" we mean an edge (j,k) if G is undirected and either $j \to k$ or $k \to j$ if G is directed. Sometimes we write $j \in G$ or $(j,k) \in G$ to mean $j \in N$ or $(j,k) \in E$ respectively. A *path* $p := (j_1, \ldots, j_K)$ is an ordered set of nodes $j_k \in N$ so that $(j_k, j_{k+1}) \in E$ for $k = 1, \ldots, K-1$. In that case we refer to a link or a node in the cycle by $(j_k, j_{k+1}) \in p$ or $j_k \in p$ respectively. A *cycle* is a path where $j_K = j_1$. A *simple cycle* is a cycle that visits every node at most once. Unless specified otherwise, we refer to j interchangeably as a node or a bus and $j \sim k$ interchangeably as a link, an edge, or a line.

Given a function $f : \mathbb{R}^n \to \mathbb{R}^m$, $\frac{\partial f}{\partial x}$ is the $m \times n$ matrix whose (j,k) entry is

$$\left[\frac{\partial f}{\partial x}\right]_{jk} := \frac{\partial f_j}{\partial x_k}(x), \qquad j = 1, \ldots, m, \;\; k = 1, \ldots, n$$

and $\nabla f(x) := \left(\frac{\partial f}{\partial x}\right)^{\mathsf{T}}$ is its transpose. In particular if $m = 1$ then $\frac{\partial f}{\partial x}$ is a row vector and $\nabla f(x)$ is a column vector.

We use e to denote the constant $\lim_n (1 + 1/n)^n$ and $e_j \in \{0,1\}^n$ the unit vector of appropriate size n with a single 1 in the jth position. We use $\ln = \log_e$ to denote the natural log. When there is no confusion we may also use log to denote ln. The vector $\mathbf{1}_n$ usually denotes the vector of all 1s of size n and $\mathbb{I}_n$ usually denotes the identity matrix of size n. Without the subscript, the vector $\mathbf{1}$ and the identity matrix $\mathbb{I}$ either denote the corresponding vector and matrix of size 3 (in unbalanced three-phase systems) or a generic size depending on context. We overload notation and use the same letter to denote different things depending on the context; e.g., I may mean current or the identity matrix, G may mean a graph or the real part of an admittance matrix $Y = G + \mathbf{i}B$, and x may mean a generic variable or the imaginary part (reactance) of an impedance $z = r + \mathbf{i}x$.

For the study of three-phase power systems, both balanced and unbalanced, $e^a := (1,0,0)$, $e^b := (0,1,0)$, $e^c := (0,0,1)$, and $e^\phi_j \in \{0,1\}^{3n}$ is the unit vector with a single 1 in the $j\phi$th position. We often use $\alpha := e^{-\mathbf{i}2\pi/3}$. The standard balanced vector in

[1] As explained in Definition A.2 and Remark A.1 of Chapter A.5, for a complex matrix, $x^{\mathsf{H}} A x \geq 0$ for all $x \in \mathbb{C}^n$ implies that A is Hermitian, so including Hermitian in the definition of psd is redundant and only for uniformity, because for a real matrix, $x^{\mathsf{T}} A x \geq 0$ for all $x \in \mathbb{R}^n$ does not imply A is symmetric.

positive sequence is $\alpha_+ := (1, \alpha, \alpha^2)$ and that in negative sequence is $\alpha_- := (1, \alpha^2, \alpha)$. The following conversion matrices are key to the understanding of three-phase power systems:

$$\Gamma := \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{bmatrix}, \qquad \Gamma^{\mathsf{T}} := \begin{bmatrix} 1 & 0 & -1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}$$

Its properties are explained in Theorems 1.2 and 14.2. The similarity transformation to obtain symmetrical components due to Fortescue is defined by the eigenvectors $(\mathbf{1}, \alpha_+, \alpha_-)$ of Γ.

0.4 Units

The unit of a quantity is specified usually the first time the quantity is introduced. Commonly used units in this book are collected here for convenience. We often overload notations so that the same symbol may refer to different quantities depending on the context, e.g., I may denote a vector of current phasors $I = (I_i, i = 1, \ldots, n)$ or the identity matrix of appropriate size, V may denote a vector of voltage phasors $V = (V_i, i = 1, \ldots, n)$ or their unit volt.

1. voltage $v(t), V$: volt (V).
2. current $i(t), I$: ampere (A).
3. real power P : watt (W); reactive power Q : volt-ampere reactive (var); complex power $S := P + \mathbf{i}Q$, apparent power $|S|$: volt-ampere (VA).
4. resistance r, reactance $x = \mathbf{i}\omega l$ or $1/\mathbf{i}\omega c$, impedance $z := r + \mathbf{i}x$: ohm (Ω).
5. conductance $g := r/(r^2 + x^2)$, susceptance $b := x/(r^2 + x^2)$, admittance $y := z^{-1} =: g + \mathbf{i}b$: Siemen (S) or mho ($\Omega^{-1}$).
6. inductance l: henry (H); magnetic flux linkage $\lambda(t) = li(t)$: weber-turn (Wb-turn).
7. capacitance c: farad (F); electric charge $q(t) = cv(t)$: coulomb (C)

We will sometimes overload notation, e.g., l is used sometimes to denote inductance, sometimes inductance per unit length, some times a line index. The meaning should be clear from the context.

Part I

Power network: models, operation, analysis

1 Basic concepts

This chapter introduces basic concepts in modeling the steady-state behavior of an alternating current (AC) power system where voltages and currents are sinusoidal functions of time. For us, steady state means that the frequencies of voltages and currents in the entire network are at their nominal value (e.g., 60 Hz in the US, 50 Hz in China and Europe). In Chapter 1.1 we describe phasor representation of sinusoidal voltages and currents, and introduce circuit models of devices that make up a single-phase system. In Chapter 1.2 we explain balanced three-phase systems and how to simplify their analysis using per-phase models. In Chapter 1.3 we define the concept of complex power for single-phase and three-phase systems, and illustrate through an example that a three-phase system saves power and conductors compared with a single-phase system serving the same load.

1.1 Single-phase systems

An AC system consists of generators and loads connected by transmission or distribution lines and transformers. Their behavior can be described using quantities such as voltages, currents, and power which are sinusoidal functions of time. These quantities obey laws of physics. For our purposes they are the Kirchhoff's current law (KCL), Kirchhoff's voltage law (KVL), and Ohm's law. These laws allow us to analyze or simulate system behavior in the time domain. For steady-state behavior it is often easier to transform these quantities to the phasor domain, apply the corresponding physical laws in the phasor domain to analyze the steady state of a power network, and then translate the results back to the time domain, as illustrated in Figure 1.1.

In this section we define voltage and current phasors, present simple models of generators, loads, and lines using voltage sources, current sources, and impedances. We also summarize KCL, KVL and Ohm's law in the phasor domain. They can be used to analyze a network of these circuit elements.

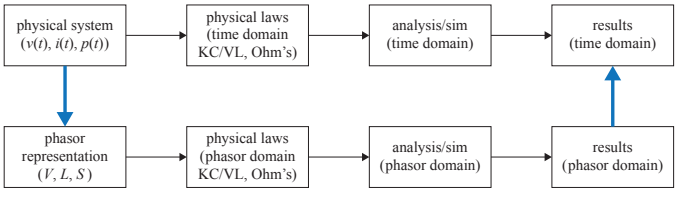


Figure 1.1 Phasor representation and analysis.

1.1.1 Voltage and current phasors

The quantities of interest, voltage $v(t)$, current $i(t)$, and power $p(t)$, are physical and can be empirically measured. The potential energy gained in moving a unit of charge from point k to point j is called the *voltage*, or *electric potential difference*, between j and k, denoted by v_{jk}. Its SI unit (International Systems of Units) is volt (V), or equivalently, joule/coulomb. Usually we arbitrarily fix a reference point 0 for all voltages in the system under study. In that case we refer to the voltage at point j with respect to the reference point simply as the *voltage at* j and denote v_{j0} simply by v_j. Then the voltage between two points j and k is $v_{jk} := v_j - v_k$ and represents the energy required to move a unit of charge from point k to point j. The flow rate of electric charge through a point is called the *current* through that point. Its SI unit is *ampere* (A), or equivalently, coulomb/second. The rate of energy transfer when a unit of charge is moved through an electric potential difference (voltage) between two points is called electric *power*. Its SI unit is watt (W), or equivalently, joule/second. It is equal to the product of voltage and current between these two points.

A sinusoidal voltage function is

$$v(t) = V_{\max}\cos(\omega t + \theta_V) \;=\; \mathrm{Re}\left\{V_{\max}e^{\mathbf{i}\theta_V}\cdot e^{\mathbf{i}\omega t}\right\}$$

where $V_{\max}$ is the amplitude (i.e., maximum magnitude) of the voltage $v(t)$, ω is the steady-state frequency in radian, and θ_V is the phase angle. In steady state, ω is assumed fixed systemwide, and hence a voltage function is fully specified by two parameters $(V_{\max}, \theta_V)$. This motivates the definition of voltage *phasor*

$$V := \frac{V_{\max}}{\sqrt{2}}\, e^{\mathbf{i}\theta_V} \qquad \text{volt (V)}$$

such that

$$v(t) = \mathrm{Re}\left(\sqrt{2}|V|\cdot e^{\mathbf{i}(\omega t+\theta_V)}\right) \tag{1.1}$$

The period of $v(t)$ is $T := 2\pi/\omega$. The magnitude of the voltage phasor

$$|V| := \frac{V_{\max}}{\sqrt{2}}$$

is equal to the root-mean-square (RMS) value of the voltage, defined as

$$\sqrt{\frac{1}{T}\int_0^{\mathsf{T}} v^2(t)dt} = \sqrt{\frac{1}{T}\int_0^{\mathsf{T}} V_{\max}^2 \cos^2(\omega t+\theta_V)\,dt} \;=\; \frac{V_{\max}}{\sqrt{2}}$$

where we have used $\cos^2\phi = (1+\cos 2\phi)/2$.

Similarly let the sinusoidal current function be

$$i(t) = I_{\max}\cos(\omega t+\theta_I) \qquad \text{ampere (A)}$$

with the corresponding current phasor

$$I := \frac{I_{\max}}{\sqrt{2}}\, e^{\mathbf{i}\theta_I}$$

such that

$$i(t) = \mathrm{Re}\left(\sqrt{2}|I|\cdot e^{\mathbf{i}(\omega t+\theta_I)}\right) \tag{1.2}$$

The RMS value of the current is $|I| := I_{\max}/\sqrt{2}$.

1.1.2 Single-phase devices

Basic building blocks of an AC power system are generators that generate power, loads that consume power, transmission and distribution lines, and transformers that connect generators and loads. These devices can be modeled by circuit elements such as impedances, voltage sources, current sources, and (later) power sources, as we now explain.

Impedance z.

The voltage and current across a resistor r in ohm (Ω), an ideal inductor l in henry (H), or an ideal capacitor c in farad (F) satisfy a linear relation, both in the time domain and in the phasor domain. We now derive Ohm's law in the phasor domain from its representation in the time domain.

Consider the circuit in Figure 1.2. The voltage $v(t)$ across the resistor r and the

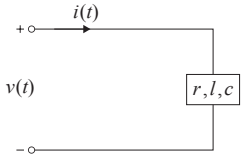


Figure 1.2 In phasor domain the voltage V and current I across a linear circuit element z are related by $V = zI$ where the impedances for resistor r, inductor l, capacitor c are $z = r, \mathbf{i}\omega l, (\mathbf{i}\omega c)^{-1}$ respectively.

current $i(t)$ through it are related by Ohm's law:

$$v(t) = r\, i(t)$$

Using (1.1)(1.2), this is equivalent to:

$$\mathrm{Re}\left\{V \cdot \sqrt{2} e^{\mathbf{i}\omega t}\right\} = \mathrm{Re}\left\{r\, I \cdot \sqrt{2} e^{\mathbf{i}\omega t}\right\}$$

Hence Ohm's law in the phasor domain for a resistor is:

$$V = r\, I$$

The current across a resistor is called *in phase* with the voltage.

An ideal inductor l is characterized by

$$v(t) = l \frac{di(t)}{dt}$$

Substituting (1.1) and

$$\frac{di(t)}{dt} = -\omega\, I_{\max} \sin(\omega t + \theta_I) \quad = \quad \omega\, I_{\max} \cos(\omega t + \theta_I + \pi/2)$$

we have

$$\mathrm{Re}\left\{V \cdot \sqrt{2} e^{\mathbf{i}\omega t}\right\} = \mathrm{Re}\left\{\mathbf{i}\omega\, l\, I \cdot \sqrt{2} e^{\mathbf{i}\omega t}\right\}$$

or in the phasor domain:

$$V = (\mathbf{i}\omega\, l)\, I$$

The current across an inductor is said to *lag* the voltage by $\pi/2$ radian.

Similarly an ideal capacitor c is characterized by

$$i(t) = c \frac{dv(t)}{dt}$$

Substituting (1.2) and

$$\frac{dv(t)}{d(t)} = -\omega\, V_{\max} \sin(\omega t + \theta_V) \quad = \quad \omega\, V_{\max} \cos(\omega t + \theta_V + \pi/2)$$

we have

$$\mathrm{Re}\left\{I \cdot \sqrt{2} e^{\mathbf{i}\omega t}\right\} \quad = \quad \mathrm{Re}\left\{\mathbf{i}\omega\, c\, V \cdot \sqrt{2} e^{\mathbf{i}\omega t}\right\}$$

or in the phasor domain:

$$V = \frac{1}{\mathbf{i}\omega\, c} I$$

The current across a capacitor is said to *lead* the voltage by $\pi/2$ radian.

In summary we define the *impedances* of these elements, a resistor r, an ideal inductor l, and an ideal capacitor c in the phasor domain as respectively:

$$z_r := r, \qquad z_l := \mathbf{i}\omega l, \qquad z_c := \frac{1}{\mathbf{i}\omega c}$$

Instead of impedance z, sometimes it is convenient to use its inverse, called the *admittance* $y := z^{-1}$. The voltage V across an impedance z (or admittance y) and the current I through it are related in the phasor domain by

$$V = zI \text{ and } I = yV$$

An important advantage of phasor representation of an AC circuit is that circuit analysis involves only algebraic operations rather than differential equations in the time domain.

Example 1.1. A voltage $v(t)$ is applied to a resistor r and an inductor l in series and the current through these devices is $i(t)$. Derive the dynamic equation that relates $(v(t), i(t))$ in the time domain and the corresponding equation that relates their phasors (V, I).

Solution. Let $v_1(t) = ri(t)$ denote the voltage drop across the resistor and $v_2(t)$ the voltage drop across the inductor that satisfies $v_2(t) = l\frac{d}{dt}i(t)$. Then the relation between $(v(t), i(t))$ is given by KVL: $v(t) = v_1(t) + v_2(t)$ or

$$v(t) = ri(t) + l\frac{d}{dt}i(t)$$

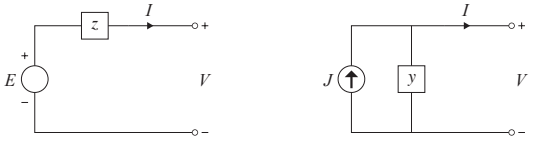
Noting that $v(t) = \text{Re}\left\{\sqrt{2}V e^{\mathbf{i}\omega t}\right\}$ and $i(t) = \text{Re}\left\{\sqrt{2}I e^{\mathbf{i}\omega t}\right\}$, we multiply both sides of the equation above by $e^{\mathbf{i}\omega t}$ to get

$$\begin{aligned} \sqrt{2}V e^{\mathbf{i}\omega t} &= r\sqrt{2}I e^{\mathbf{i}\omega t} + l\left(\mathbf{i}\omega\sqrt{2}I e^{\mathbf{i}\omega t}\right) \\ V &= (r + \mathbf{i}\omega l) I \end{aligned}$$

Hence the resistor and inductor in series can be modeled in the phasor domain by an impedance $z := r + \mathbf{i}\omega l$. □

Voltage source (E, z).

In the phasor domain, a voltage source is a circuit model with a constant *internal voltage* E in series with an impedance z, as shown in Figure 1.3(a). Its external behavior is



(a) Voltage source (b) Current source

Figure 1.3 A voltage source (E, z) and a current source (J, y). An ideal voltage source has $z = 0$ and an ideal current source has $y = 0$.

described by the relation between its *terminal voltage and terminal* (V, I):

$$V = E - zI$$

Hence the open-circuit (terminal) voltage V equals the internal voltage E. We often adopt an ideal voltage source with $z = 0$. In this case $V = E$.

Current source (J, y).

In the phasor domain, a current source is a circuit model with a constant *internal current* J in parallel with an admittance y, as shown in Figure 1.3(b). Its external behavior is described by the relation between its terminal voltage and current (V, I):

$$I = J - yV$$

Hence the closed-circuit (terminal) current I equals the internal current J. We often adopt an ideal current source with $y = 0$. In this case $I = J$.

Remark 1.1. 1 A nonideal voltage source (E, z) and a current source (J, y) are equivalent, i.e., have the same terminal voltage and current relationship if their parameters satisfy

$$J = \frac{E}{z} \qquad \text{(closed-circuit equivalent)}$$

$$y := z^{-1} \qquad \text{(open-circuit equivalent)}$$

2 Ideal voltage or current sources are reasonable models as their series impedances or shunt admittances can be combined with the series impedance and shunt admittances of a transmission or distribution line to which they are connected, as we will see in Chapter 2. We will therefore often use ideal voltage and current sources in this book with series series impedances and shunt admittances.

□

Single-phase devices.

Basic devices in a power system are generators, loads, transmission and distribution lines, transformers, and other control devices. A generator can be modeled by a voltage source or current source. A load can be modeled by an impedance (or admittance), a voltage source, or a current source. A transmission or distribution line can be modeled by a series impedance and a shunt admittance at each end of the line; the details are described in Chapter 2. A transformer can be modeled by a series impedance and a shunt admittance followed by voltage and current gains; the details are described in Chapter 3. We will introduce in Chapter 1.3 the concept of complex power. This leads to a device model that we will call a *power source* that generates or draws a constant power. These are summarized in Table 1.1. They are abstract models of physical devices. For relation to a common load model, called ZIP, that describes how power consumed by a load depends on the voltage magnitude $|V|$ across the load, see Exercise 1.1. This book develops techniques for analyzing power system models constructed from these circuit elements.

Device	Circuit model
Generator	Voltage source, current source, power source
Load	Impedance, voltage source, current source, power source
Line	Impedance (Chapter 2)
Transformer	Impedance, voltage/current gain (Chapter 3)

Table 1.1 Circuit elements commonly used for modeling generators, loads, lines, and transformers.

1.1.3 KVL, KCL, Ohm's Law, Tellegen's theorem

Consider a circuit consisting of an interconnection of resistors, inductors, capacitors, and voltage and current sources. An ideal voltage source between two points enforces a given voltage between these two points. An ideal current source between two points enforces a given current between them. We now describe Kirchhoff's current law (KCL), Kirchhoff's voltage law (KVL), Ohm's law for a general circuit and derive a result called Tellegen's theorem.

We represent a circuit by a connected *directed* graph $\hat{G} := (\hat{N}, \hat{E})$ with an arbitrary orientation where $\hat{N}$ is a set of nodes and $\hat{E} \subseteq \hat{N} \times \hat{N}$ is a set of links. We sometimes abuse notation and use $\hat{N}$ to denote both the set of nodes and the number of nodes in $\hat{N}$ when the meaning should be clear from the context. We allow multiple links between two nodes j and k (see Figure 1.4). A link l that points from node j to node k is represented by $l = (j,k)$ or $l = j \to k$. Multiple links $l_1, \ldots, l_k$ between nodes j and k may have different orientations, e.g., $l_1 = j \to k$ and $l_2 = k \to j$. There are two variables associated with each link $l = (j,k)$ between nodes j and k. The voltage across link l is denoted by U_l in the direction of l and the branch current over link l from j to k is denoted by J_l.

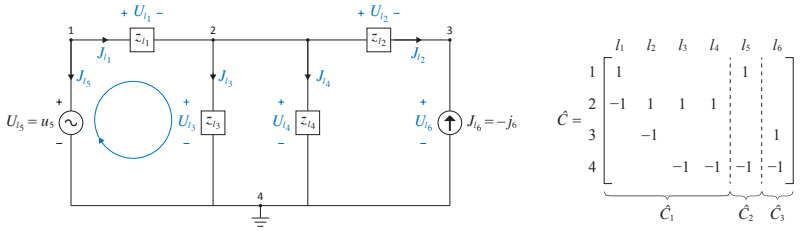
A link l represents either an impedance, a voltage source, or a current source. If link l represents an impedance then its value z_l is given and the voltage U_l and branch current J_l across link l satisfies $U_l = z_l J_l$ (Ohm's law). If link l represents a voltage source then $U_l = u_l$ is given, and if it represents a current source then $J_l = j_l$ is given. These notations are illustrated in Figure 1.4a.

KCL, KVL.

Kirchhoff's current law (KCL) states that the incident currents at any node j sum to zero:

$$-\sum_{i:i\to j\in\hat{E}} J_{ij} + \sum_{k:j\to k\in\hat{E}} J_{jk} = 0 \tag{1.3a}$$

For the example in Figure 1.4 this means $-J_{l_1} + J_{l_2} + J_{l_3} + J_{l_4} = 0$ at node 2. Kirchhoff's voltage law (KVL) states that voltage drops around any cycle c sum to zero. Consider



(a) Circuit (b) Incidence matrix

Figure 1.4 A circuit represented as a directed graph where each link l is either an impedance z_l, a voltage source U_l, or a current source J_l. The voltage source $U_{l_5} = u_5$ and current source $J_{l_6} = -j_6$ are given. Its incidence matrix $\hat{C}$ is partitioned into $\hat{C}_1$ corresponding to the impedances, $\hat{C}_2$ corresponding to the voltage source, and $\hat{C}_3$ corresponding to the current source.

a cycle c in the graph with an arbitrary orientation, say, clockwise. A link l in the cycle that is in the same direction as c is denoted by $l \in c$ and a link l that is in the opposite direction to c is denoted by $-l \in c$. Then KVL states that the voltage drops around any cycle c sum to zero:

$$\sum_{l \in c} U_l \; - \; \sum_{-l \in c} U_l = 0 \tag{1.3b}$$

For the cycle indicated in Figure 1.4(a) we have $U_{l_1} + U_{l_3} - U_{l_5} \; = \; 0$.

We can represent (1.3) compactly in vector notation. Let $U := \left(U_l, l \in \hat{E}\right)$ and $J := \left(J_l, l \in \hat{E}\right)$ denote the vectors of voltages and currents respectively across these lines. Let $\hat{C} \in \{-1, 0, 1\}^{|\hat{N}| \times |\hat{E}|}$ be the node-by-link *incidence matrix* defined by:

$$\hat{C}_{jl} := \begin{cases} 1 & \text{if } l = j \to k \text{ for some bus } k \\ -1 & \text{if } l = i \to j \text{ for some bus } i \\ 0 & \text{otherwise} \end{cases}, \qquad j \in \hat{N},\ l \in \hat{E}$$

See Figure 1.4 (properties of general incidence matrices are summarized in Appendix A.11). Then Kirchhoff's current law (1.3a) states that

KCL: $$\hat{C}\, J = 0 \tag{1.4a}$$

Kirchhoff's voltage law is equivalent to the condition that there exist *nodal* voltages $V \in \mathbb{C}^{|\hat{N}|}$ (with respect to the common reference point node 0) such that

KVL: $$U = \hat{C}^{\mathsf{T}} V \tag{1.4b}$$

i.e., given line voltages U, there must exist nodal voltages such that $U_l = V_j - V_k$ where $l = j \to k$, from which (1.3b) follows. This seems intuitive and can be proved mathematically using concepts in algebraic graph theory (Exercise 1.2). Without loss of generality we use node $\hat{N}$ as the common reference point for all voltages, i.e., we have by definition

$$V_{\hat{N}} \; := \; 0 \tag{1.4c}$$

Circuit analysis.

Consider a circuit represented by an incidence matrix $\hat{C}$. The $|\hat{N}|\times|\hat{E}|$ incidence matrix $\hat{C}$ is of rank $|\hat{N}|-1$ since $\hat{G}$ is connected, with span($\mathbf{1}$) as its null space (see Chapter A.11 for more details). Therefore (1.4) consists of $|\hat{N}|+|\hat{E}|+1$ complex equations in $|\hat{N}|+2|\hat{E}|$ complex variables (V,J,U), of which $|\hat{N}|+|\hat{E}|$ equations are linearly independent. To obtain another $|\hat{E}|$ equations we note that across every link l is exactly one of the following devices:

1 *impedance* with a given z_l: Its behavior is described by Ohm's law

$$U_l = z_l J_l \tag{1.5a}$$

2 *ideal voltage source* with a given u_l: Its behavior is described by

$$U_l = u_l \tag{1.5b}$$

3 *ideal current source* with a given j_l: Its behavior is described by

$$J_l = j_l \tag{1.5c}$$

Partition the set $\hat{E}$ of links into three disjoint sets $\hat{E} =: \hat{E}_1 \cup \hat{E}_2 \cup \hat{E}_3$ where $\hat{E}_1$ is the set of impedances, $\hat{E}_2$ voltage sources, and $\hat{E}_3$ current sources. Then (1.4)(1.5) specify $|\hat{N}|+2|\hat{E}|+1$ equations in $|\hat{N}|+2|\hat{E}|$ variables (V,J,U), of which at most $|\hat{N}|+2|\hat{E}|$ equations are linearly independent:

$$\begin{bmatrix} 0 & \hat{C} & 0 \\ 0 & -Z & \mathbb{I}_{|\hat{E}_1|} \\ 0 & 0 & \mathbb{I}_{|\hat{E}_2|} \\ 0 & \mathbb{I}_{|\hat{E}_3|} & 0 \\ \hat{C}^{\mathsf{T}} & 0 & -\mathbb{I}_{|\hat{E}|} \\ e^{\mathsf{T}}_{|\hat{N}|} & 0 & 0 \end{bmatrix} \begin{bmatrix} V \\ J \\ U \end{bmatrix} = \begin{bmatrix} 0_{|\hat{N}|} \\ 0_{|\hat{E}_1|} \\ u \\ j \\ 0_{|\hat{E}|} \\ 0_1 \end{bmatrix} \tag{1.6}$$

where $Z := \operatorname{diag}\left(z_l, l \in \hat{E}_1\right)$ is the diagonal matrix of impedances, $u := (u_l, l \in \hat{E}_2)$ and $j := (j_l, l \in \hat{E}_3)$ are vectors of voltage and current sources respectively, 0_m is the zero vector of size m, $\mathbb{I}_m$ is the identity matrix of size m, and $e_n \in \{0,1\}^{|\hat{N}|}$ is the unit vector with a single 1 in the nth entry. A circuit analysis problem is to solve (1.4)(1.5), or equivalently (1.6), for these variables. A sufficient condition is given in Theorem 1.1 for the existence and uniqueness of solution. A necessary condition for the existence of a solution is that the given voltage and current vectors (v,j) are consistent, e.g., if only current sources are incident on a node k, then these given currents must satisfy KCL at node k, or if a set of voltage sources form a cycle c then these given voltages must satisfy KVL on c.

The system of equations (1.6) can be simplified, as follows. Order the links such

that the incidence matrix decomposes into submatrices $\hat{C}_1, \hat{C}_2, \hat{C}_3$ corresponding to impedances, voltage sources, and current sources respectively (see Figure 1.4b):

$$\hat{C} =: \begin{bmatrix} \hat{C}_1 & \hat{C}_2 & \hat{C}_3 \end{bmatrix}$$

Partition the branch voltages U and branch currents J accordingly:

$$U := \begin{bmatrix} U_1 \\ u \\ U_3 \end{bmatrix}, \quad J := \begin{bmatrix} J_1 \\ J_2 \\ j \end{bmatrix}$$

where v and j are the given vectors of voltage and current sources respectively. Then KCL and KVL are

$$\begin{gathered} \hat{C}_1 J_1 + \hat{C}_2 J_2 = -\hat{C}_3 j \\ U_1 = \hat{C}_1^{\mathsf{T}} V, \quad u = \hat{C}_2^{\mathsf{T}} V, \quad U_3 = \hat{C}_3^{\mathsf{T}} V \end{gathered}$$

for some nodal voltages V. Use Ohm's law $U_1 = Z J_1$ to eliminate U_1 to obtain

$$\begin{bmatrix} 0 & \hat{C}_1 & \hat{C}_2 & 0 \\ \hat{C}_1^{\mathsf{T}} & -Z & 0 & 0 \\ \hat{C}_2^{\mathsf{T}} & 0 & 0 & 0 \\ \hat{C}_3^{\mathsf{T}} & 0 & 0 & -\mathbb{I}_{|\hat{E}_3|} \end{bmatrix} \begin{bmatrix} V \\ J_1 \\ J_2 \\ U_3 \end{bmatrix} = \begin{bmatrix} -\hat{C}_3 j \\ 0 \\ u \\ 0 \end{bmatrix} \tag{1.7}$$

The desired quantities (V, J_1, J_2, U_3) are solutions of (1.7) if they exist. Given J_1, U_1 is given by $U_1 = Z J_1$.

Recall that we take without loss of generality node $\hat{N}$ as the common reference point for nodal voltages and assign $V_{\hat{N}} := 0$. We can consider the $(|\hat{N}| - 1) \times |\hat{E}|$ *reduced incidence matrix* C obtained from $\hat{C}$ by deleting the last row corresponding to the reference node $\hat{N}$. The advantage of using C is that it has a full row rank of $|\hat{N}| - 1$. Let $V_{-\hat{N}} := \left(V_j, j \neq \hat{N}\right)$ be the vector of all non-reference nodal voltages. Similarly partition C into $C =: [C_1 \;\; C_2 \;\; C_3]$. Then (1.7) is equivalent to the following equation:

$$\underbrace{\begin{bmatrix} 0 & C_1 & C_2 & 0 \\ C_1^{\mathsf{T}} & -Z & 0 & 0 \\ C_2^{\mathsf{T}} & 0 & 0 & 0 \\ C_3^{\mathsf{T}} & 0 & 0 & -\mathbb{I}_{|\hat{E}_3|} \end{bmatrix}}_{M} \begin{bmatrix} V_{-\hat{N}} \\ J_1 \\ J_2 \\ U_3 \end{bmatrix} = \begin{bmatrix} -C_3 j \\ 0 \\ u \\ 0 \end{bmatrix} \tag{1.8}$$

The key feature of this model, compared with (1.7), is that it does not contain the reference node $\hat{N}$.

Example 1.2. Consider the circuit in Figure 1.4 represented by the directed graph $\hat{G} = (\hat{N}, \hat{E})$ with

$$\begin{aligned} \hat{N} &:= \{1,2,3,4\} \\ \hat{E} &:= \{l_1 := 1 \to 2,\ l_2 := 2 \to 3,\ l_3 := 2 \to 4,\ l_4 := 2 \to 4,\ l_5 := 1 \to 4,\ l_6 := 3 \to 4\} \end{aligned}$$

The incidence matrix $\hat{C}$ can be partitioned into submatrices

$$\hat{C}_1 := \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 1 & 1 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & -1 \end{bmatrix}, \qquad \hat{C}_2 := \begin{bmatrix} 1 \\ 0 \\ 0 \\ -1 \end{bmatrix}, \qquad \hat{C}_3 := \begin{bmatrix} 0 \\ 0 \\ 1 \\ -1 \end{bmatrix}$$

The reduced incidence submatrices are then

$$C_1 := \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 1 & 1 \\ 0 & -1 & 0 & 0 \end{bmatrix}, \qquad C_2 := \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \qquad C_3 := \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

The equation (1.8) becomes:

$$\left[\begin{array}{ccc|cccc|cc} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ \hline 1 & -1 & 0 & -z_{l_1} & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & -z_{l_2} & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -z_{l_3} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & -z_{l_4} & 0 & 0 \\ \hline 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 \end{array}\right] \left[\begin{array}{c} V_1 \\ V_2 \\ V_3 \\ \hline J_{l_1} \\ J_{l_2} \\ J_{l_3} \\ J_{l_4} \\ \hline J_{l_5} \\ \hline U_{l_6} \end{array}\right] = \left[\begin{array}{c} 0 \\ 0 \\ j_6 \\ \hline 0 \\ 0 \\ 0 \\ 0 \\ \hline u_5 \\ \hline 0 \end{array}\right]$$

□

We now discuss the existence and uniqueness of solution to (1.8).

Theorem 1.1. The matrix M in (1.8) is invertible if the following square matrices of sizes $|\hat{N}| - 1$ and $|\hat{E}_2|$ respectively are invertible:

$$C_1 Z^{-1} C_1^{\mathsf{T}}, \qquad C_2^{\mathsf{T}} \left(C_1 Z^{-1} C_1^{\mathsf{T}} \right)^{-1} C_2$$

where $\hat{E}_2$ is the set of voltage sources. □

A necessary condition for $C_1 Z^{-1} C_1^{\mathsf{T}}$ to be nonsingular is that the graph $\hat{G}$ is connected. In that case, if z_l are real and positive (i.e., resistive network) then $C_1 Z^{-1} C_1^{\mathsf{T}}$ is nonsingular since $Z := \mathrm{diag}(z_l)$ is positive definite and C and hence its submatrix C_1 are both of full row rank. When Z is complex, however, $C_1 Z^{-1} C_1^{\mathsf{T}}$ may be singular even if z_l are all nonzero and C_1 is of full row rank (see discussions in Chapter 4.2.3). The matrix C_2^{T} is of full row rank if and only if no voltage sources form a cycle in the circuit.

The proof of Theorem 1.1 relies on the following fact. Let $M \in \mathbb{C}^{n \times n}$ and partition it into blocks:

$$M = \begin{bmatrix} A_1 & B \\ D & A_2 \end{bmatrix}$$

where $A_1 \in \mathbb{C}^{k\times k}$, $k < n$, and the other submatrices are of matching dimensions. If A_2 is invertible then the $k\times k$ matrix $M/A_2 := A_1 - BA_2^{-1}D$ is called the *Schur complement of block* A_2 of matrix M. In that case M is nonsingular if and only if M/A_2 is nonsingular. Similarly if A_1 is invertible then the $(n-k)\times(n-k)$ matrix $M/A_1 := A_2 - DA_1^{-1}B$ is called the *Schur complement of block* A_1 of matrix M, and M is nonsingular if and only if M/A_1 is nonsingular; see Theorem A.4 in Appendix A.3.

Proof of Theorem 1.1 We can interchange the second and third rows and interchange the second and third column and write (1.8) equivalently in terms of the matrix

$$\tilde{M} = \left[\begin{array}{cc|cc} 0 & C_2 & C_1 & 0 \\ C_2^{\mathsf{T}} & 0 & 0 & 0 \\ \hline C_1^{\mathsf{T}} & 0 & -Z & 0 \\ C_3^{\mathsf{T}} & 0 & 0 & -\mathbb{I}_{|\hat{E}_3|} \end{array}\right]$$

The matrix M is nonsingular if and only if $\tilde{M}$ is. Since Z and $\mathbb{I}_{|\hat{E}_3|}$ are both nonsingular, $\tilde{M}$ is nonsingular if and only if the Schur complement of $\operatorname{diag}\left(-Z, -\mathbb{I}_{|\hat{E}_3|}\right)$:

$$S := \begin{bmatrix} 0 & C_2 \\ C_2^{\mathsf{T}} & 0 \end{bmatrix} + \begin{bmatrix} C_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Z^{-1} & 0 \\ 0 & \mathbb{I}_{|\hat{E}_3|} \end{bmatrix} \begin{bmatrix} C_1^{\mathsf{T}} & 0 \\ C_3^{\mathsf{T}} & 0 \end{bmatrix} = \begin{bmatrix} C_1 Z^{-1} C_1^{\mathsf{T}} & C_2 \\ C_2^{\mathsf{T}} & 0 \end{bmatrix}$$

is nonsingular. The Schur complement S is a square matrix of size $(\hat{N}-1)+|\hat{E}_2|$ where $\hat{E}_2$ is the set of voltage sources. If $C_1 Z^{-1} C_1^{\mathsf{T}}$ is nonsingular then M is nonsingular if and only if the Schur complement

$$S/\left(C_1 Z^{-1} C_1^{\mathsf{T}}\right) := -C_2^{\mathsf{T}}\left(C_1 Z^{-1} C_1^{\mathsf{T}}\right)^{-1} C_2$$

is nonsingular. □

Tellegen's theorem

An important result in circuit theory is Tellegen's theorem that expresses a relation between voltage drops across links and currents on these links. It is a simple consequence of Kirchhoff's laws and algebraic graph theory (see Chapter A.11 for more details). Since the rank of the $|\hat{N}|\times|\hat{E}|$ incidence matrix $\hat{C}$ is $|\hat{N}|-1$ assuming $\hat{G}$ is connected, $\operatorname{rank}\left(\hat{C}^{\mathsf{T}}\right) = \operatorname{rank}(\hat{C}) = |\hat{N}|-1$ and the dimension of the null space $\operatorname{null}(\hat{C})$ is $|\hat{E}|-|\hat{N}|+1$. Recall that the subspaces $\operatorname{null}(\hat{C})$ and $\operatorname{range}\left(\hat{C}^{\mathsf{T}}\right)$ are orthogonal complements of each other and they span $\mathbb{C}^{|\hat{E}|}$, i.e., $\mathbb{C}^{|\hat{E}|} = \operatorname{null}(\hat{C}) + \operatorname{range}\left(\hat{C}^{\mathsf{T}}\right)$. The KCL and KVL (1.4a)(1.4b) say that the branch current (vector) J is in $\operatorname{null}(\hat{C})$ and the branch voltage (vector) U is in $\operatorname{range}\left(\hat{C}^{\mathsf{T}}\right)$ respectively. Therefore

Tellegen's theorem: $$J^{\mathsf{H}}U = 0$$

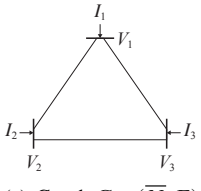
It is remarkable that this relation holds for any branch current J and branch voltage U, even if they are from different networks as long as these networks have the same incidence matrix $\hat{C}$.

1.1.4 One-line diagram and equivalent circuit

A power system is often not specified as a circuit of the form we study in Chapter 1.1.3. Instead it is usually specified by what is called a *one-line diagram*. A one-line diagram is equivalent to a circuit that includes the common reference point for nodal voltages as an addition node. Each line in the one-line diagram may represent a transmission line, a distribution line or a transformer, single or multi-phased. As we will see below if a single-phase line has a equivalent Π circuit then the line translates into three links in the equivalent circuit. In this subsection we formally define one-line diagram and derive its equivalent circuit. A one-line diagram can be analyzed by applying the method of Chapter 1.1.3 to its equivalent circuit.

One-line diagram.

A one-line diagram specifies a network topology and admittance parameters associated with the lines; see an example in Figure 1.5 for a three-bus network. Formally we define a one-line diagram as a pair $(G, \mathbb{Y})$ where $G := (\overline{N}, E)$ is a graph and $\mathbb{Y} := \left(y^s_{jk}, y^m_{jk}, y^m_{kj},\ l = (j,k) \in E\right)$ is a set of line parameters for every line $l \in E$ (we assume here a single-phase system and $y^s_{jk} = y^s_{kj}$). Each node $j \in \overline{N}$ represents a bus in the power system. We will therefore refer to j as a bus or a node interchangeably. Each link $l \in E$ represents a transmission or distribution line or a transformer. We will therefore refer to l as a line, a link or a branch interchangeably. The line parameter $y^s_{jk} \in \mathbb{C}$ is called the *series admittance* associated with line (j,k) and $\left(y^m_{jk}, y^m_{kj}\right) \in \mathbb{C}^2$ is called its *shunt admittances*. We will see below how these parameters determine the equivalent circuit of the line. There can be multiple lines between two buses, though for notational simplicity we often assume there is a single line between each pair of buses in which case a line l between buses j and k can be identified by (j,k).



(a) Graph $G = (\overline{N}, E)$

(b) Line parameters $\mathbb{Y}$

Figure 1.5 One-line diagram for a three-bus network $(G, \mathbb{Y})$. It is not a circuit but has an equivalent Π circuit model.

There can be a *nodal* device at each node $j \in \overline{N}$. The device can be an impedance z_j, an ideal voltage source v_j, or an ideal current source i_j. The interpretation is that these devices are connected between node j and the common voltage reference point and behave according to (1.5). (We will introduce later the nodal device called a power source.)

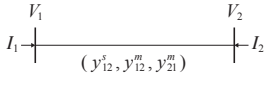
The behavior of the network specified by a one-line diagram is described in terms of its equivalent circuit.

Equivalent circuit.

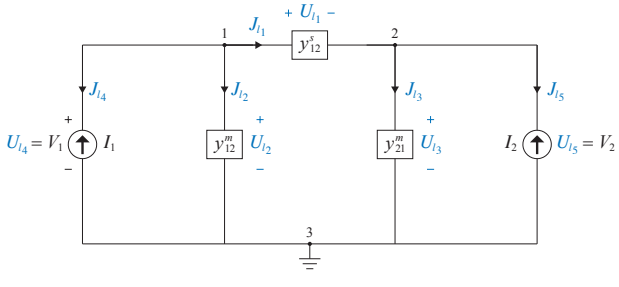
Associated with each node j are a nodal voltage $V_j \in \mathbb{C}$ with respect to an arbitrary but common reference point and a nodal current injection $I_j \in \mathbb{C}$. To derive the relation between the vectors (V, I) of nodal voltages and currents specified by the one-line diagram, we first derive its equivalent circuit and then apply the method of Chapter 1.1.3 to the circuit.

We illustrate this with a simple 2-bus network. The method and the conclusion extend directly to general networks.

Example 1.3 (Equivalent Π circuit of a single line)**.** Figure 1.6(a) specifies a one-line diagram $(G, \mathbb{Y})$ for a network consisting of two nodes 1 and 2 connected by a line $l = (1,2)$. Suppose there is an ideal current source at each node with given current injections (I_1, I_2). The nodal voltages are (V_1, V_2). The line parameter $\left(y^s_{12}, y^m_{12}, y^m_{21}\right)$



(a) One-line diagram $(G, \mathbb{Y})$



(b) Equivalent Π circuit

Figure 1.6 One-line diagram $(G, \mathbb{Y})$ with two nodes $1, 2$ connected by a line $l = (1,2)$ and its equivalent Π circuit. The nodal current injections (I_1, I_2) and the nodal voltages (V_1, V_2) in the one-line diagram become current sources and branch voltages respectively between nodes $1, 2$ and the reference node 3 in the Π circuit.

defines the equivalent circuit in Figure 1.6(b) called the Π circuit of line $l = (1,2)$. (We will explain the origin of the equivalent circuit in Chapter 2.) The application of KVL,

KCL, and Ohm's law on the Π circuit leads to a relation between (I_1, I_2) and (V_1, V_2), as we now explain.

Let the *directed* graph $\hat{G} := (\hat{N}, \hat{E})$ represent the Π circuit where

$$\hat{N} := \{1,2,3\}$$
$$\hat{E} := \{l_1 := 1 \to 2,\ l_2 := 1 \to 3,\ l_3 := 2 \to 3,\ l_4 := 1 \to 3,\ l_5 := 2 \to 3\}$$

as shown in Figure 1.6(b). Note that the graph G of the one-line diagram has 2 nodes while the graph $\hat{G}$ of its equivalent circuit has 3 nodes with node 3 being the voltage reference point. For each link $l \in \hat{E}$ let U_l and J_l denote the voltage and current across link l in the direction of l. Let $U := (U_l, l \in \hat{E})$ and $J := (J_l, l \in \hat{E})$. The devices on the links l_1, l_2, l_3 are admittances with

$$l_1 : J_{l_1} = y^s_{12} U_{l_1}, \qquad l_2 : J_{l_2} = y^m_{12} U_{l_2}, \qquad l_3 : J_{l_3} = y^m_{21} U_{l_3}$$

Since the nodal devices at nodes 1 and 2 are ideal current sources with given currents I_1 and I_2 respectively, we have

$$l_4 : J_{l_4} = -I_1, \qquad l_5 : J_{l_5} = -I_2$$

The node-by-link incidence matrix $\hat{C}$ of the Π circuit is

$$\hat{C} := \begin{bmatrix} 1 & 1 & 0 & 1 & 0 \\ -1 & 0 & 1 & 0 & 1 \\ 0 & -1 & -1 & -1 & -1 \end{bmatrix}$$

The KCL, KVL and Ohm's law in terms of $\hat{C}, U, J$ for the Π circuit in Figure 1.6(b) are:

$$\text{KCL}: \quad \hat{C}J = 0 \tag{1.9a}$$
$$\text{KVL}: \quad \exists\, V := (V_1, V_2, V_3) \quad \text{s.t.} \quad U = \hat{C}^{\mathsf{T}} V \tag{1.9b}$$
$$\text{Ohm's law}: \quad J_{l_1} = y^s_{12} U_{l_1},\ J_{l_2} = y^m_{12} U_{l_2},\ J_{l_3} = y^m_{21} U_{l_3} \tag{1.9c}$$
$$\text{nodal current sources}: \quad J_{l_4} = -I_1,\ J_{l_5} = -I_2 \tag{1.9d}$$

We will set the nodal voltage $V_3 := 0$, i.e., node 3 in $\hat{N}$ is chosen to be the voltage reference point. This allows us to eliminate branch variables (U, J) from (1.9) to obtain a relation between the nodal currents $I := (I_1, I_2)$ and voltages $V := (V_1, V_2)$:

$$I_1 = y^s_{12}(V_1 - V_2) + y^m_{12} V_1, \qquad I_1 = y^s_{12}(V_2 - V_1) + y^m_{21} V_2$$

In vector form this is $I = YV$ with

$$Y := \begin{bmatrix} y^s_{12} + y^m_{12} & -y^s_{12} \\ -y^s_{12} & y^s_{12} + y^m_{21} \end{bmatrix}$$

The matrix Y is called the *admittance matrix* of the network, a single-line in this example. The admittance matrix Y can be expressed using the submatrix $C_{1\text{line}} := \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ of $\hat{C}$ corresponding to link l_1 with the series admittance y^s_{12}. Note that $C_{1\text{line}}$ includes every node in the equivalent circuit except the reference node 3, i.e., C describes the

connectivity between exactly the set of nodes in the original one-line diagram. If we let $Y^s := [y^s_{12}]$ and $Y^m := \begin{bmatrix} y^m_{12} \\ y^m_{21} \end{bmatrix}$ then

$$Y \;:=\; C_{1\text{line}} Y^s C_{1\text{line}}^{\mathsf{T}} + \text{diag}(Y^m)$$

□

For a general network specified by a one-line diagram $(G = (\overline{N}, E), \mathbb{Y})$ let $V := (V_j, j \in \overline{N})$ and $I := (I_j, j \in \overline{N})$ denote the vectors of nodal voltages and current injections from the nodal devices respectively. We interpret the line parameter $\left(y^s_{jk}, y^m_{jk}, y^m_{kj}\right)$ of each line (j,k) as defining a Π circuit model for the line, as explained in Example 1.3. This induces an equivalent circuit for the entire network that can be described by a directed graph $\hat{G} = (\hat{N}, \hat{E})$ constructed from $G = (\overline{N}, E)$, as follows. The set $\hat{N}$ of nodes in the equivalent circuit is

$$\hat{N} \;:=\; \overline{N} \cup \{|\overline{N}| + 1\}$$

where the additional node $\hat{N} := |\overline{N}| + 1$ is the reference point for all voltages, i.e., $V_{\hat{N}} := 0$.

For each node $j \in \overline{N}$ in the one-line diagram, there is a link $l = j \to \hat{N}$ in the equivalent circuit corresponding to the nodal device at j. The voltage U_l across line $l = j \to \hat{N}$ is $U_l = V_j$ and the current J_l across link l in the direction of l is $J_l = -I_j$. If the nodal device at node j is an impedance z_j, then $V_j = U_l = z_j J_l = -z_l I_j$; if it is an ideal voltage source V_j, then $U_l = V_j$ is given; if it is an ideal current source I_j, then $J_l = -I_j$ is given. If there is no nodal device at node j, then we set $J_l = -I_j := 0$.

For each line $(j,k) \in E$ parametrized by $\left(y^s_{jk}, y^m_{jk}, y^m_{kj}\right)$ in the one-line diagram, there are 3 links $\left(l_{jk}, l_{j\hat{N}}, l_{k\hat{N}}\right)$ in $\hat{E}$ in the equivalent circuit. The currents $\left(J_{jk}, J_{j\hat{N}}, J_{k\hat{N}}\right)$ and voltages $\left(U_{jk}, U_{j\hat{N}}, U_{k\hat{N}}\right)$ across these links satisfy:

$$\begin{aligned} l_{jk} = j \to k \;&:\; J_{jk} = y^s_{jk} U_{jk} \\ l_{j\hat{N}} = j \to \hat{N} \;&:\; J_{j\hat{N}} = y^m_{jk} U_{j\hat{N}} \\ l_{k\hat{N}} = k \to \hat{N} \;&:\; J_{k\hat{N}} = y^m_{kj} U_{k\hat{N}} \end{aligned}$$

The set of links $l = j \to k$ corresponding to series admittances is the set E in the one-line diagram. Let $\hat{E}_{\hat{N}}$ denote the set of links $l = j \to \hat{N}$ connecting nodes $j \in \overline{N}$ to the reference node $\hat{N}$. They correspond to the shunt admittances on each line $(j,k) \in E$ and the nodal device at each node $j \in \overline{N}$. The set $\hat{E}$ in the equivalent circuit is the disjoint union of these two types of links:

$$\hat{E} \;=\; E \cup \hat{E}_{\hat{N}}$$

See the two-bus network in Figure 1.6 and its equivalent Π circuit for an example. If bus $j \in \overline{N}$ is connected to m_j other buses $k \in \overline{N}$ in the one-line diagram, then there will be

m_j links $l_{j_k\hat{N}} = j \to \hat{N}$ in the equivalent circuit, for $k = 1, \ldots, m_j$, all between node j and $\hat{N}$, representing shunt admittances y^m_{jk} on these lines. Therefore $|\hat{E}_{\hat{N}}| = |\overline{N}| + 2|E|$.

Let $C_{1\text{line}}$ be the incidence matrix for the subgraph of the circuit consisting of non-reference nodes $\overline{N}$ and links in E connecting them, i.e., $C_{1\text{line}}$ describes the connectivity between exactly the nodes in the one-line diagram:

$$[C_{1\text{line}}]_{jl} := \begin{cases} 1 & \text{if } l = j \to k \text{ in } E \\ -1 & \text{if } l = i \to j \text{ in } E \\ 0 & \text{otherwise} \end{cases}, \qquad j \in \overline{N},\ l \in E$$

Let $Y^s := \text{diag}\left(y^s_{jk},\ (j,k) \in E\right)$ denote the diagonal matrix of series admittances on the lines. Let $Y^m := \text{diag}\left(y^m_{jj},\ j \in \overline{N}\right)$ denote the diagonal matrix of total shunt admittances $y^m_{jj} := \sum_{k:(j,k)\in E} y^m_{jk}$ incident on each bus j. Then the linear relation between nodal current injections and voltages found in Example 1.3:

$$I = YV \tag{1.10a}$$

holds for the general network with the admittance matrix Y given by (Exercise 1.5)

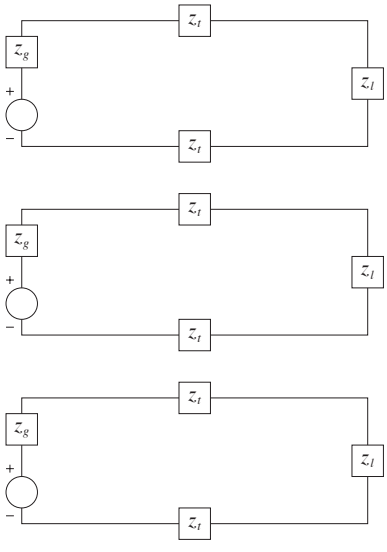
$$Y = C_{1\text{line}} Y^s C^{\mathsf{T}}_{1\text{line}} + Y^m \tag{1.10b}$$

The relation (1.10) serves as a formal identification of a one-line diagram $(G, \mathbb{Y})$ with an equivalent Π circuit. Moreover given $(G, \mathbb{Y})$ we can directly write down the admittance matrix Y without going through the circuit analysis conducted above. We therefore often refer to the one-line diagram itself as a circuit model. This relation will be studied in detail in Chapter 4.

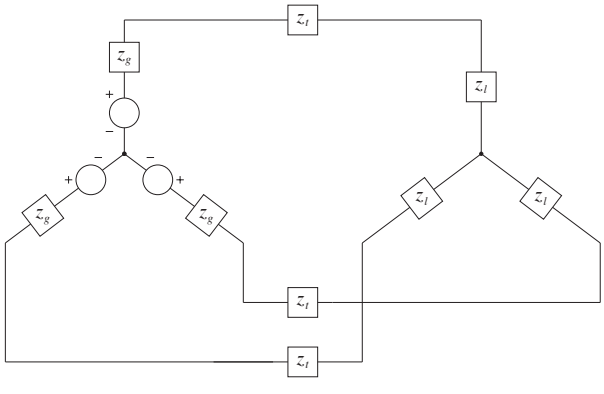
1.2 Three-phase systems

To motivate three-phase systems, consider the single-phase system in Figure 1.7(a) composed of three identical circuits each consisting of a generator modeled as a voltage source in series with an impedance z_g, a forward conductor and a return conductor each modeled as an impedance z_t, and a load modeled as an impedance z_l. The same loads can also be supplied by a three-phase system shown in Figure 1.7(b). As we will illustrate in Chapter 1.3.3, such a three-phase system needs half as much the conductor and incurs half as much the thermal loss as the single-phase system. In this section we explain the operation of three-phase systems.

Three-phase sources and loads can be arranged in Y (Wye) or Δ (Delta) configurations. This is explained in Chapter 1.2.1. A three-phase system is balanced if all the sources are balanced, loads are identical, and transmission lines are identical and have symmetric geometry. A balanced three-phase system has several simplifying properties. In Chapter 1.2.2 we prove a theorem that summarizes the mathematical structure of balanced three-phase systems that underlies these properties. We apply this theorem



(a) Single-phase system



(b) Balanced three-phase system

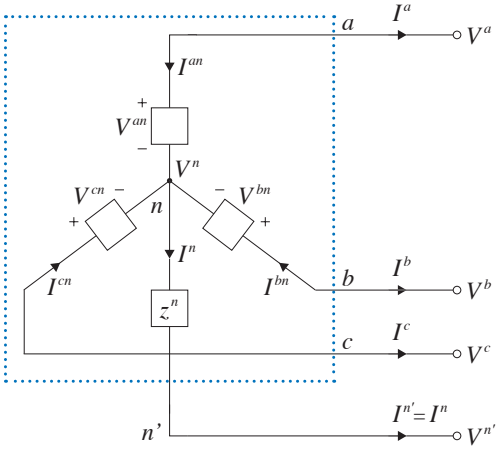
Figure 1.7 A single-phase system and a balanced three-phase system that transfer power from generators through transmission lines to loads.

to balanced system in Y configuration (Chapter 1.2.3) and Δ configuration (Chapter 1.2.4). This leads to per-phase analysis of a balanced system described in Chapter 1.2.5. Finally we present in Chapter 1.2.6 example configurations common in a power distribution system.

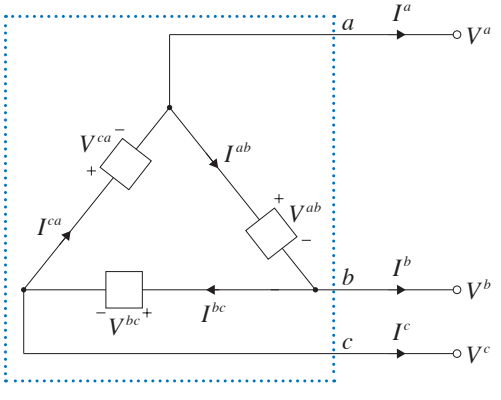
Even though power systems are generally multiphased, single-phase models are widely used as per-phase models of balanced three-phase systems, especially for transmission system applications. Unbalanced three-phase systems are studied in Part III of this book.

1.2.1 Y and Δ configurations

Three single-phase devices can be arranged in either an Y or a Δ configuration as shown in Figure 1.8. They can be three voltage sources, three current sources, or



(a) Y configuration



(b) Δ configuration

Figure 1.8 Three-phase systems, not necessarily balanced, in Y and Δ configurations.

three impedances and they may not be identical, e.g., the three impedances may have different values.

Y configuration.

For the Y configuration, the internal voltage (vector) is $V^Y := (V^{an}, V^{bn}, V^{cn})$. These voltages are called *phase-to-neutral* or *phase* voltages. The internal current (vector) $I^Y := (I^{an}, I^{bn}, I^{cn})$ is defined to flow from each terminal to the neutral as shown in Figure 1.8(a). The external behavior of a three-phase device is described by what is measurable on the terminal of the device. The *terminal* (or *nodal* or *bus*) voltage $V := (V^a, V^b, V^c)$ are voltages with respect to an arbitrary but common reference point, and the *terminal* (or *line*) current $I := \left(I^a, I^b, I^c\right)$ is defined to be the current coming out of the device as shown in the figure. If the common reference point is taken to be the neutral of this device then $V = V^Y$, i.e., the terminal voltage is the same as the phase voltage for Y configuration. Otherwise $V = V^Y - V^n\mathbf{1}$ where $\mathbf{1}$ is three-dimensional vector of all 1s. As we will see in Chapters 1.2.3 and 1.2.4, for a balanced systems, the neutrals of all Y-configured devices are at the same voltage and therefore can serve as the common reference point. This is not necessarily the case for an unbalanced system, which we will study in Part III of this book.

Hence, for Y configuration, the terminal voltage and current (V, I) are determined by the internal voltage and current $\left(V^Y, I^Y\right)$ according to (when the common reference point for V is the neutral so that $V^n := 0$):

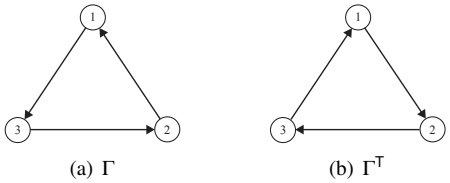
$$V = V^Y, \qquad I = -I^Y \tag{1.11}$$

When the common reference is not the neutral of this device, we have $V = \left(V^Y + V^n\mathbf{1}\right)$.

Instead of the terminal voltage V it is also common to describe the behavior of the three-phase device in terms of its *line-to-line* or *line* voltage $V^{\text{line}} := \left(V^{ab}, V^{bc}, V^{ca}\right)$. To relate V^{line} to V or to V^Y, define the matrices Γ and its transpose Γ^T:

$$\Gamma := \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{bmatrix}, \qquad \Gamma^\mathsf{T} := \begin{bmatrix} 1 & 0 & -1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \tag{1.12}$$

We call Γ and Γ^T *conversion matrices*. They can be interpreted as the bus-by-line incidence matrices of the directed graphs shown in Figure 1.9. Then



(a) Γ (b) Γ^T

Figure 1.9 Directed graphs of which Γ and Γ^T are incidence matrices.

$$\begin{bmatrix} V^{ab} \\ V^{bc} \\ V^{ca} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{bmatrix}}_{\Gamma} \begin{bmatrix} V^{a} \\ V^{b} \\ V^{c} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{bmatrix}}_{\Gamma} \begin{bmatrix} V^{an} \\ V^{bn} \\ V^{cn} \end{bmatrix}$$

or in vector form:

$$V^{\text{line}} = \Gamma V = \Gamma V^{Y} \tag{1.13}$$

This holds for both Y and Δ configurations and whether or not the common reference point for V is the neutral of a Y configured device (since $\Gamma \mathbf{1} = 0$).

Δ configuration.

For the Δ configuration in Figure 1.8(b), the internal voltage (vector) is the line-to-line voltage $V^{\Delta} := (V^{ab}, V^{bc}, V^{ca}) = V^{\text{line}}$, and the internal current $I^{\Delta} := (I^{ab}, I^{bc}, I^{ca})$ is the line-to-line current. As for the Y configuration, the terminal voltage $V := (V^{a}, V^{b}, V^{c})$ are voltages with respect to an arbitrary but common reference point. The terminal current is $I := \left(I^{a}, I^{b}, I^{c}\right)$ as shown in Figure 1.8(b). The terminal voltage and current (V, I) is determined by the internal voltage and current $\left(V^{\Delta}, I^{\Delta}\right)$ according to

$$\underbrace{\begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{bmatrix}}_{\Gamma} \begin{bmatrix} V^{a} \\ V^{b} \\ V^{c} \end{bmatrix} = \begin{bmatrix} V^{ab} \\ V^{bc} \\ V^{ca} \end{bmatrix}, \qquad \begin{bmatrix} I^{a} \\ I^{b} \\ I^{c} \end{bmatrix} = -\underbrace{\begin{bmatrix} 1 & 0 & -1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}}_{\Gamma^{\mathsf{T}}} \begin{bmatrix} I^{ab} \\ I^{bc} \\ I^{ca} \end{bmatrix}$$

or in vector form (for arbitrary common reference point for V):

$$\Gamma V = V^{\Delta}, \qquad I = -\Gamma^{\mathsf{T}} I^{\Delta} \tag{1.14}$$

Equivalent Y configuration.

For any Δ configuration with given internal voltage $V^{\Delta} := (V^{ab}, V^{bc}, V^{ca})$ and current $I^{\Delta} := (I^{ab}, I^{bc}, I^{ca})$, an equivalent Y configuration is one that has the same external behavior. This means that, if $V^{Y} := (V^{an}, V^{bn}, V^{cn})$ and $I^{Y} := (I^{an}, I^{bn}, I^{cn})$ are the internal voltage and current of the Y-equivalent then they are related to $\left(V^{\Delta}, I^{\Delta}\right)$ according to (from (1.13) (1.14)):

$$\Gamma V^{Y} = V^{\Delta}, \qquad I^{Y} = \Gamma^{\mathsf{T}} I^{\Delta} \tag{1.15}$$

Summary.

The external behavior (1.11) and (1.14) for Y and Δ configurations respectively as well as their equivalence (1.15) hold for any three-phase system whether or not it is balanced. The relation (1.13) between line-to-line voltage V^{line} and terminal voltage V holds for Y and Δ configurations whether or not the system is balanced.

The behavior of a three-phase system is determined by the mathematical properties of the conversion matrices Γ and Γ^{T}. When a system is balanced the conversion becomes particularly simple because the transformation of balanced vectors under Γ and Γ^{T} preserves their balanced nature (Corollary 1.3). We now explain these mathematical properties and then apply them to the analysis of balanced systems in Chapters 1.2.3 and 1.2.4.

1.2.2 Balanced vectors and conversion matrices $\Gamma, \Gamma^{\mathsf{T}}$

Definition 1.1 (Balanced vector)**.** A vector $x := (x_1, x_2, x_3)$ with $x_j = |x_j| e^{\mathbf{i}\theta_j} \in \mathbb{C}$, $j = 1, 2, 3$, is called *balanced* if x_j have the same magnitude and they are separated by 120°, i.e.,

$$|x_1| \;=\; |x_2| \;=\; |x_3|$$

and either

$$\theta_2 - \theta_1 = -\frac{2\pi}{3} \;\text{ and }\; \theta_3 - \theta_1 = \frac{2\pi}{3} \qquad \text{(positive sequence)} \tag{1.16a}$$

or

$$\theta_2 - \theta_1 = \frac{2\pi}{3} \;\text{ and }\; \theta_3 - \theta_1 = -\frac{2\pi}{3} \qquad \text{(negative sequence)} \tag{1.16b}$$

In this chapter we focus on single-phase equivalent circuits of balanced systems. In Part III of this book we study unbalanced systems and generalize the definition of balance to allow a nonzero bias (see Definition 14.1), i.e., we will call $\hat{x}$ a (generalized) balanced vector if it is of the form $\hat{x} = x + \gamma \mathbf{1}$ and x is balanced according to Definition 1.1, for some possibly nonzero $\gamma \in \mathbb{C}$. The bias γ may models a common reference voltage or the internal loop flow in a Δ configuration. We assume $\gamma = 0$ in this chapter which amounts to the assumption that loop flows are zero and that all neutrals are grounded directly and voltages are defined with respect to the ground.

A balanced vector x is said to be in a *positive sequence* if x satisfies (1.16a) and in a *negative sequence* set if x satisfies (1.16b). Let

$$\alpha := e^{-\mathbf{i}2\pi/3}$$

Clearly $\alpha^2 = e^{\mathbf{i}2\pi/3}$, $\alpha^3 = 1$; see Figure 1.10. (Also see Exercise 1.6 for more properties of α.) Define the vectors

$$\alpha_+ := \begin{bmatrix} 1 \\ \alpha \\ \alpha^2 \end{bmatrix}, \qquad \alpha_- := \begin{bmatrix} 1 \\ \alpha^2 \\ \alpha \end{bmatrix} \tag{1.17a}$$

Then α_+ is a balanced vector in a positive sequence and α_- is a balanced vector in a negative sequence. Moreover the set of all balanced positive-sequence vectors is

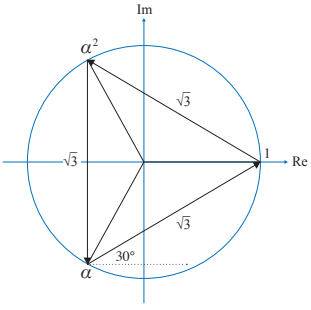


Figure 1.10 Phase shift $\alpha := e^{-\mathbf{i}2\pi/3}$ in Theorem 1.2.

span(α_+) and the set of all balanced negative-sequence vectors is span(α_-), i.e., x is a balanced vector in a positive sequence and y a balanced vector in a negative sequence if and only if

$$x = x_1\alpha_+, \qquad y = y_1\alpha_-, \qquad x_1, y_1 \in C \tag{1.17b}$$

Note that $\overline{\alpha}_+ = \alpha_-$ where for any vector x, $\overline{x}$ is its complex conjugate componentwise. Define the matrix F whose columns are α_+, α_- as well as $\mathbf{1}$ normalized:

$$F := \frac{1}{\sqrt{3}}\begin{bmatrix}\mathbf{1} & \alpha_+ & \alpha_-\end{bmatrix} = \frac{1}{\sqrt{3}}\begin{bmatrix}1 & 1 & 1\\ 1 & \alpha & \alpha^2\\ 1 & \alpha^2 & \alpha\end{bmatrix} \tag{1.18}$$

All main properties of balanced three-phase systems originate from the mathematical properties of the vectors α_+, α_- and their transformation under the matrices $\Gamma, \Gamma^\mathsf{T}$ defined in (1.12), summarized in Theorem 1.2. Its proof is left as Exercise 1.7. The theorem implies in particular that the transformations Γ and Γ^T preserve the balanced nature of a vector and hence ensures that the entire network stays balanced. The key enabling property is that the voltages and currents from balanced sources are in span(α_+) or span(α_-) and (α_+, α_-) are eigenvectors of $\Gamma, \Gamma^\mathsf{T}$ (according to (1.19a)(1.20a)).

Theorem 1.2 (Transformation of balanced vectors by $\Gamma, \Gamma^\mathsf{T}$)**.** Let $\alpha := e^{-\mathbf{i}2\pi/3}$. Recall the balanced vectors (α_+, α_-) defined in (1.17a), the matrices F in (1.18) and $\Gamma, \Gamma^\mathsf{T}$ in (1.12).

1 Suppose the entries x_j of $x := (x_1, x_2, x_3) \in \mathbb{C}^3$ have the same magnitude. Then x is balanced if and only if $x_1 + x_2 + x_3 = 0$.
2 The columns of F are orthonormal. Both F and $\overline{F}$ are complex symmetric, i.e., $F^\mathsf{T} = F$ and $\overline{F}^\mathsf{T} = \overline{F}$, where $\overline{F}$ is the complex conjugate of F componentwise. Hence

$$F^{-1} = F^\mathsf{H} = \overline{F} = \frac{1}{\sqrt{3}}\begin{bmatrix}\mathbf{1} & \alpha_- & \alpha_+\end{bmatrix}$$

3 Γ is a normal matrix, $\Gamma\Gamma^\mathsf{T} = \Gamma^\mathsf{T}\Gamma$. (Note that $\Gamma\Gamma^\mathsf{T} = \Gamma^\mathsf{T}\Gamma$ are Laplacian matrices of the graphs in Figure 1.9.)

4 *Spectral decomposition of* Γ:
   1 The eigenvalues and eigenvectors of Γ are

$$\Gamma\mathbf{1} = 0, \qquad \Gamma\alpha_+ = (1-\alpha)\alpha_+, \qquad \Gamma\alpha_- = (1-\alpha^2)\alpha_- \tag{1.19a}$$

   where $1-\alpha = \sqrt{3}e^{\mathbf{i}\pi/6}$ and $1-\alpha^2 = \sqrt{3}e^{-\mathbf{i}\pi/6}$.
   2 Therefore the spectral decomposition of Γ is:

$$\Gamma = F\begin{bmatrix}0 & & \\ & 1-\alpha & \\ & & 1-\alpha^2\end{bmatrix}\overline{F} \tag{1.19b}$$

5 *Spectral decomposition of* Γ^{T}:
   1 The eigenvalues and eigenvectors of Γ^{T} are

$$\Gamma^{\mathsf{T}}\mathbf{1} = 0, \qquad \Gamma^{\mathsf{T}}\alpha_- = (1-\alpha)\alpha_-, \qquad \Gamma^{\mathsf{T}}\alpha_+ = (1-\alpha^2)\alpha_+ \tag{1.20a}$$

   where $1-\alpha = \sqrt{3}e^{\mathbf{i}\pi/6}$ and $1-\alpha^2 = \sqrt{3}e^{-\mathbf{i}\pi/6}$.
   2 Therefore the spectral decomposition of Γ^{T} is:

$$\Gamma^{\mathsf{T}} = \overline{F}\begin{bmatrix}0 & & \\ & 1-\alpha & \\ & & 1-\alpha^2\end{bmatrix}F \tag{1.20b}$$

The following corollary of the theorem is repeatedly used in the analysis of balanced systems. It says that the transformation of a balanced vector x under Γ and Γ^{T} reduces to a scaling by $(1-\alpha)$ and $(1-\alpha^2)$ respectively.

Corollary 1.3. For any balanced positive-sequence vector $x \in \mathbb{C}^3$ and $\gamma \in \mathbb{C}$, we have

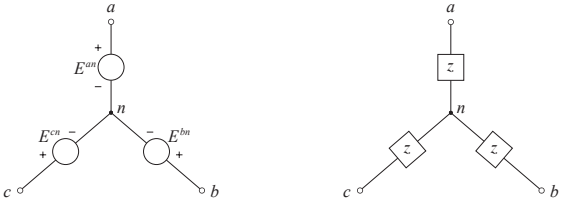
1 $\Gamma(x+\gamma\mathbf{1}) = (1-\alpha)x$.
2 $\Gamma^{\mathsf{T}}(x+\gamma\mathbf{1}) = (1-\alpha^2)x$.
3 $\Gamma\Gamma^{\mathsf{T}}(x+\gamma\mathbf{1}) = \Gamma^{\mathsf{T}}\Gamma(x+\gamma\mathbf{1}) = 3x$.

Informally a three-phase system is called *balanced* if all voltages and currents are balanced vectors in, say, positive-sequence sets. The main consequence of the corollary is the following. A three-phase system consists of voltage sources, current sources, and impedances connected by lines. The voltage and current at any point in the system are induced by the internal voltages of voltage sources and the internal currents of current sources. When these sources are balanced positive-sequence sets, their internal voltages and currents are in $\text{span}(\alpha_+)$ and α_+ is an eigenvector of Γ and Γ^{T}. This means that the transformation of balanced voltages and currents under $\Gamma, \Gamma^{\mathsf{T}}$ reduces to a scaling of these variables by their eigenvalues $1-\alpha$ and $1-\alpha^2$ respectively. Since the voltage and current at every point in the system are linear combinations of transformed source voltages and source currents, transformed by Γ, Γ^T and line admittance matrices, they remain in $\text{span}(\alpha_+)$ when the sources are balanced and the lines are identical and

phase-decoupled. This is the key property that enables balanced sources to induce balanced voltages and currents throughout the network, leading to per-phase analysis of three-phase systems. A formal statement and its proof have to wait till Chapter 16 (Theorem 16.7) when we develop a general model of unbalanced three-phase system. In this chapter we will use the corollary to analyze example circuits to build intuition.

1.2.3 Balanced systems in Y configuration

Figure 1.11 shows the Y configuration of voltage sources and impedance loads. The loads are said to be *balanced* if their impedances z are identical. An ideal three-



(a) Balanced sources (b) Balanced loads

Figure 1.11 Balanced three-phase (a) voltage source E^Y and (b) impedance $z^Y := \text{diag}(z,z,z)$ in Y configuration.

phase voltage source in Y configuration is specified by its internal voltage (vector) $E^Y := \left(E^{an}, E^{bn}, E^{cn}\right)$ in the phasor domain between the terminals a,b,c and the neutral n respectively. It is called *balanced* if E^Y is a balanced vector according to Definition 1.1, i.e.,

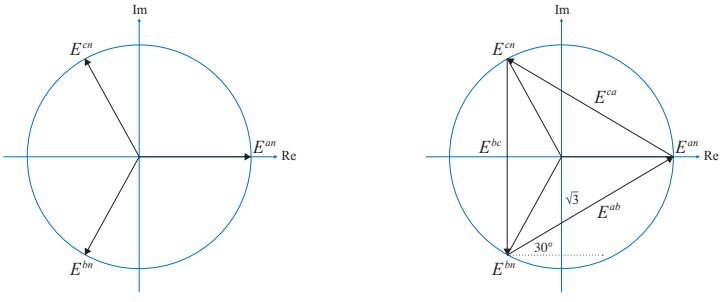
$$\text{positive sequence:} \qquad E^{an} = 1\angle\theta, \quad E^{bn} = 1\angle\theta - 120^\circ, \quad E^{cn} = 1\angle\theta + 120^\circ$$

or

$$\text{negative sequence:} \qquad E^{an} = 1\angle\theta, \quad E^{bn} = 1\angle\theta + 120^\circ, \quad E^{cn} = 1\angle\theta - 120^\circ$$

where their magnitudes are normalized to 1. See Figure 1.12(a) where $\theta = 0$. For a balanced voltage source in a positive sequence, the instantaneous voltages in the time domain reach their maximum values in the order abc. We sometimes call abc in such an order a *positive sequence* and the voltages $\left\{E^{an}, E^{bn}, E^{cn}\right\}$ a (balanced) positive-sequence set. Whether a voltage source is in a positive or negative sequence depends only on how one labels the wires. Therefore, unless otherwise specified, we will always consider abc to be a positive sequence. If there are multiple three-phase sources connected to the same network their phase sequences must be the same.

Theorem 1.2 implies the following properties of a balanced positive-sequence voltage source:



(a) Phase voltages (b) Phase and line voltages

Figure 1.12 A balanced three-phase source in Y configuration. (a) Its phase voltage (vector) $E^Y := \left(E^{an}, E^{bn}, E^{cn}\right)$ is a balanced vector. (b) Its line voltage $E^{\text{line}} = \Gamma E^Y = (1-\alpha)E^Y$.

1 Sum to zero: $E^{an} + E^{bn} + E^{cn} = 0$
2 All voltages and currents are in a balanced positive sequence, i.e., all are in $\text{span}(\alpha_+)$.
3 Phases are decoupled.

Sum to zero.

The first property follows from Theorem 1.2.1, or more directly, $E^Y = \alpha_+ E^{an}$ and hence $\mathbf{1}^{\mathsf{T}} E^Y = \left(\mathbf{1}^{\mathsf{T}} \alpha_+\right) E^{an} = 0$.

Line voltage V^{line} is balanced.

The second properties is due to the fact that α_+ is an eigenvector of $\Gamma, \Gamma^{\mathsf{T}}$. Specifically the line voltage $E^{\text{line}} := \left(E^{ab}, E^{bc}, E^{ca}\right)$ across the terminals is given by $E^{\text{line}} = \Gamma E^Y$ from (1.13)). This implies $\mathbf{1}^{\mathsf{T}} E^{\text{line}} = E^{ab} + E^{bc} + E^{ca} = 0$. Moreover Corollary 1.3 implies

$$E^{\text{line}} = \Gamma E^Y \quad = \quad (1-\alpha)E^Y$$

Hence E^{line} is in a balanced positive sequence if E^Y is, i.e., $E^{bc} = e^{-\mathbf{i}2\pi/3} E^{ab}$ and $E^{ca} = e^{\mathbf{i}2\pi/3} E^{ab}$. Since $1-\alpha = \sqrt{3} e^{\mathbf{i}\pi/6}$ we have

$$E^{ab} \quad = \quad \sqrt{3}\, e^{\mathbf{i}\pi/6} E^{an}, \qquad E^{bc} \quad = \quad \sqrt{3}\, e^{\mathbf{i}\pi/6} E^{bn}, \qquad E^{ca} \quad = \quad \sqrt{3}\, e^{\mathbf{i}\pi/6} E^{cn}$$

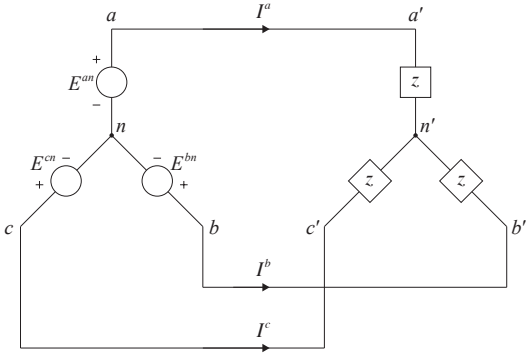
This is illustrated in Figure 1.12(b).

Balanced systems are phase-decoupled.

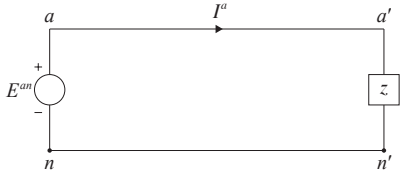
We start by analyzing the simple circuit in Figure 1.13(a) when a balanced three-phase load is connected to a balanced three-phase positive-sequence voltage source in Y configuration. We will show that

1 The neutral-to-neutral voltage is zero, $V_{nn'} = 0$.
2 The internal voltage and current across the impedances are in a balanced positive sequence.

The most important implication is that the phases are decoupled, i.e., the variables in each phase depend on quantities only in that phase, and can be analyzed separately. We will illustrate through examples that these conclusions hold in more general balanced systems than the simple circuit in Figure 1.13(a). A full understanding of phase decoupling and per-phase analysis is postponed till Part III of this book where a balanced system is studied in the context of general unbalanced systems.



(a) Balanced three-phase system



(b) Equivalent per-phase system

Figure 1.13 Balanced three-phase source and load in Y configuration and its per-phase model.

Referring to Figure 1.13(a) let

- $E^Y := \left(E^{an}, E^{bn}, E^{cn}\right)$ and $V'^Y := \left(V^{a'n'}, V^{b'n'}, V^{c'n'}\right)$ denote the internal voltages from terminals to neutrals, and $I'^Y := \left(I^{a'n'}, I^{b'n'}, I^{c'n'}\right)$ denote the internal current between the terminals a', b', c' and the neutral n' across the identical impedances z.

- $V := \left(V^a, V^b, V^c\right)$ denote the terminal voltage (vector), with respect to an arbitrary and common reference point, not necessarily the neutral n or n';
- V^n and $V^{n'}$ denote the neutral voltages with respect to the common reference point.

Given the balanced positive-sequence voltage E^Y and balanced impedances z, we wish to show that $V^n = V^{n'}$, that V'^Y, I'^Y are in a balanced positive sequence, and that phases are decoupled.

Solution. KVL, KCL, and Ohm's law imply

$$E^Y = V - V^n\mathbf{1}, \quad V'^Y = V - V^{n'}\mathbf{1}, \quad V'^Y = zI'^Y, \quad \mathbf{1}^\mathsf{T} I'^Y = 0 \tag{1.22}$$

Therefore $E^Y - V'^Y = \left(V^{n'} - V^n\right)\mathbf{1}$ and hence (since $\mathbf{1}^\mathsf{T} E^Y = 0$)

$$\mathbf{1}^\mathsf{T}\left(E^Y - V'^Y\right) = \left(V^{n'} - V^n\right)\mathbf{1}^\mathsf{T}\mathbf{1} \implies 3\left(V^{n'} - V^n\right) = -\mathbf{1}^\mathsf{T} V'^Y = -z\left(\mathbf{1}^\mathsf{T} I'^Y\right) = 0$$

showing that the voltage across the neutrals $V_{nn'} = 0$. Substituting it into (1.22) yields (denoting $y := z^{-1}$)

$$V'^Y = E^Y + \left(V^n - V^{n'}\right)\mathbf{1} = E^Y, \qquad I'^Y = yV'^Y = yE^Y$$

Hence both V'^Y and I'^Y are in a balanced positive sequence. Moreover the phases are decoupled in that $V_{\phi n'}$ and $I_{\phi n'}$, $\phi = a', b', c'$, depend only on $E_{\phi n}$ but not on voltages on other phases.

In view of Theorem 1.2.1, the terminal voltage V is not balanced unless $V^n = V^{n'} = 0$, i.e., the neutral is taken as the common reference point for voltages, because

$$\mathbf{1}^\mathsf{T} V = \mathbf{1}^\mathsf{T}\left(E^Y + V^n\mathbf{1}\right) = 3V^n$$

□

Remark 1.2. 1 Since $V_{nn'} = 0$, even if n and n' are connected, the current on that wire will be zero. We can therefore either assume n and n' are connected or disconnected in our analysis, whichever is more convenient.

2 Since the currents are balanced, $I^a + I^b + I^c = 0$ or $i^a(t) + i^b(t) + i^c(t) = 0$ at all times t, the currents flow from and return to the sources only via the wires connecting the sources to the loads, and no additional physical wires are necessary for return currents. This halves the amount of required wire compared with three separate single-phase circuits; see Chapter 1.3.3.

As a consequence, each phase of the balanced system is decoupled and equivalent to the circuit in Figure 1.13(b). We can therefore analyze the phase a equivalent circuit; see Chapter 1.2.5. The voltages and currents in phase b and phase c circuits will be the corresponding phase a quantities shifted by $-120°$ and $120°$ respectively, assuming the three-phase source is of positive sequence.

These conclusions hold for more general circuits than that in Figure 1.13(a), as Example 1.4 shows.

Example 1.4 (Balanced three-phase system in Y configuration)**.** Figure 1.14 shows a balanced three-phase source of positive sequence supplies two sets of balanced three-phase loads in parallel through balanced transmission lines. The transmission lines have a common admittance t and all loads have a constant admittance l, as shown in the figure. Suppose the neutrals are connected by lines with a common admittance y. De-

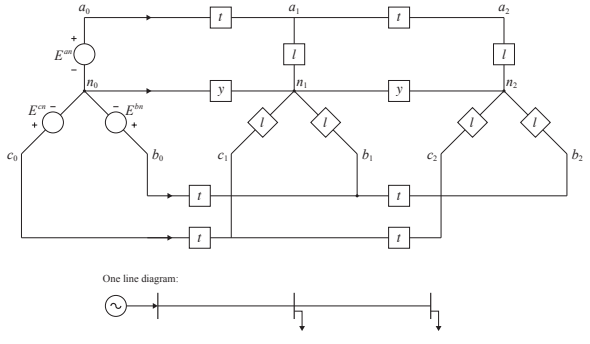


Figure 1.14 Balanced three-phase system in Y configuration (Example 1.4).

note the internal voltages and currents in stage $k = 1, 2$, by $V_k^Y := (V^{a_k n_k}, V^{b_k n_k}, V^{c_k n_k})$ and $I_k^Y := (I^{a_k n_k}, I^{a_k n_k}, I^{a_k n_k})$ respectively. Denote the terminal voltages and currents from stage $k-1$ to stage k, $k = 1, 2$, by $V_k := (V^{a_{k-1} a_k}, V^{b_{k-1} b_k}, V^{c_{k-1} c_k})$ and $I_k := (I^{a_{k-1} a_k}, I^{a_{k-1} b_k}, I^{a_{k-1} c_k})$ respectively.

Suppose $y \neq 0$, $t = y/\mu$, and $l = y/\mu^2$ for some real number $\mu \neq 0$. Prove that

1 $V_{n_0 n_1} = V_{n_1 n_2} = 0$.
2 For $k = 1, 2$, all voltages and currents V_k^Y, V_k, I_k^Y, I_k are balanced positive-sequence sets.
3 The phases are decoupled, i.e.,

$$E_0^Y = V_1 + V_1^Y$$
$$V_1^Y = V_2 + V_2^Y$$

where $E_0^Y := (E^{a_0 n_0}, E^{b_0 n_0}, E^{c_0 n_0})$.

This implies that the three phases of the balanced system in Figure 1.14 are decoupled and can be studied by analyzing the per-phase circuit shown in Figure 1.15 where the line admittances connecting the neutrals are set to zero.

Solution:

1 We will apply Ohm's law and Kirchhoff's current and voltage laws (KCL and KVL) to derive two linear equations in $(V_{n_0 n_1}, V_{n_1 n_2})$ and show that $V_{n_0 n_1} = V_{n_1 n_2} = 0$ is

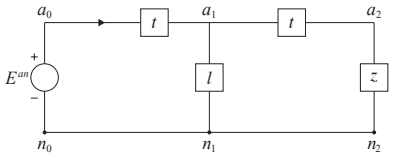


Figure 1.15 The per-phase equivalent circuit of the balanced system in Figure 1.14 in Y configuration.

the only solution to these equations. By Ohm's law across each admittance, the currents are in terms of voltages:

$$I_k^Y = lV_k^Y,\ I_k = tV_k, \qquad k = 1,2 \tag{1.23}$$

This allows us to eliminate currents I_k^Y, I_k and express KCL and KVL in the following in terms only of voltages V_k^Y, V_k.

Making use of (1.23), apply KCL at node (a_1, b_1, c_1) to obtain

$$tV^{a_0a_1} = lV^{a_1n_1} + tV^{a_1a_2},\ tV^{b_0b_1} = lV^{b_1n_1} + tV^{b_1b_2},\ tV^{c_0c_1} = lV^{c_1n_1} + tV^{c_1c_2}$$

and similarly for KCL at nodes (a_2, b_2, c_2). This in vector form is

$$tV_1 = lV_1^Y + tV_2 \tag{1.24a}$$
$$tV_2 = lV_2^Y \tag{1.24b}$$

Apply KCL at nodes (n_0, n_1, n_2) to obtain

$$t\left(\mathbf{1}^{\mathsf{T}}V_1\right) + yV^{n_0n_1} = 0$$
$$l\left(\mathbf{1}^{\mathsf{T}}V_1^Y\right) + yV^{n_0n_1} = yV^{n_1n_2}$$
$$l\left(\mathbf{1}^{\mathsf{T}}V_2^Y\right) + yV^{n_1n_2} = 0$$

where $\mathbf{1} := (1,1,1)$ is the column vector of all 1's. Hence, since $y/t = \mu$ and $y/l = \mu^2$, we have

$$\mathbf{1}^{\mathsf{T}}V_1 = -\mu V^{n_0n_1},\ \mathbf{1}^{\mathsf{T}}V_1^Y = -\mu^2 V^{n_0n_1} + \mu^2 V^{n_1n_2},\ \mathbf{1}^{\mathsf{T}}V_2^Y = -\mu^2 V^{n_1n_2} \tag{1.25}$$

Finally, apply KVL around the loops from stage 0 to stage 1 to obtain

$$E^{a_0n_0} = V^{a_0a_1} + V^{a_1n_1} - V^{n_0n_1},\ E^{b_0n_0} = V^{b_0b_1} + V^{b_1n_1} - V^{n_0n_1},\ E^{c_0n_0} = V^{c_0c_1} + V^{c_1n_1} - V^{n}$$

and similarly for loops from stage 1 to stage 2. This in vector form is

$$E_0^Y = V_1 + V_1^Y - V^{n_0n_1}\mathbf{1} \tag{1.26a}$$
$$V_1^Y = V_2 + V_2^Y - V^{n_1n_2}\mathbf{1} \tag{1.26b}$$

where $E_0^Y := (E^{a_0n_0}, E^{b_0n_0}, E^{c_0n_0})$. Substitute (1.24b) into the last equation to eliminate V_2:

$$V_1^Y = \left(\frac{1}{\mu} + 1\right)V_2^Y - V^{n_1n_2}\mathbf{1} \tag{1.26c}$$

To obtain a system of equations that involves only $(V^{n_0n_1}, V^{n_1n_2})$, multiply (1.26) by $\mathbf{1}^\mathsf{T}$ and apply (1.25) to obtain (using $\mathbf{1}^\mathsf{T}E_0 = 0$ since the sources are balanced):

$$\begin{bmatrix} \mu^2+\mu+3 & -\mu^2 \\ -\mu^2 & 2\mu^2+\mu+3 \end{bmatrix} \begin{bmatrix} V^{n_0n_1} \\ V^{n_1n_2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{1.27}$$

We now argue that the determinant of the matrix in (1.27) is nonzero, and hence $V^{n_0n_1} = V^{n_1n_2} = 0$. Let $B := \mu^2+\mu+3$. Then

$$\text{determinant} = B(B+\mu^2) - \mu^4$$

If determinant is zero then

$$B = -\frac{\mu^2}{2}\left(1 \pm \sqrt{5}\right)$$

By the definition of $B := \mu^2+\mu+3$ we therefore have

$$(3 \pm \sqrt{5})\mu^2 \; + \; 2\mu \; + \; 6 = 0$$

It is easy to check that no real number μ satisfies this equation, and hence $V^{n_0n_1} = V^{n_1n_2} = 0$.

2 We now prove that (V_k^Y, V_k) are balanced positive-sequence sets. Since $V^{n_1n_2} = 0$, (1.26c) implies

$$V_2^Y = \frac{\mu}{\mu+1} V_1^Y \tag{1.28}$$

Substitute this and (1.24b) into (1.24a) to obtain

$$V_1 = \frac{1}{\mu} V_1^Y \; + \; \frac{1}{\mu} V_2^Y \; = \; \frac{2\mu+1}{\mu(\mu+1)} V_1^Y$$

Substitute into (1.26a) to get

$$E_0^Y = \frac{2\mu+1}{\mu(\mu+1)} V_1^Y \; + \; V_1^Y$$

Hence

$$V_1^Y \; = \; \frac{\mu(\mu+1)}{\mu^2+3\mu+1} E_0 \;\text{ and }\; V_1 \; = \; \frac{\mu(2\mu+1)}{\mu^2+3\mu+1} E_0$$

Hence V_1, V_1^Y are balanced positive-sequence sets since E_0 is. Furthermore V_2, V_2^Y are balanced positive-sequence sets from (1.28) and (1.24b). Then (1.23) implies that all currents (I_k^Y, I_k) are balanced positive-sequence sets.

3 To show that the phases are decoupled, substitute $V^{n_0n_1} = V^{n_1n_2} = 0$ in (1.26a)(1.26b).

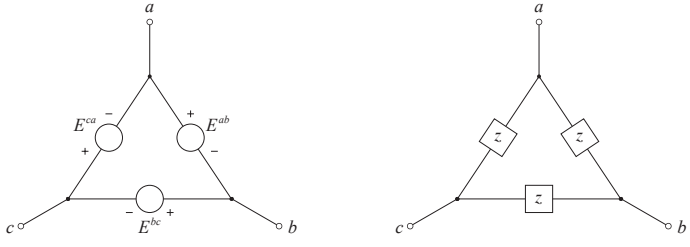
This completes the proof. □

Remark 1.3 (Phase-decoupling of lines)**.** 1 A key enabling property that allows the balanced nature of voltages and currents to propagate from one node to the next is the assumption that three-phase lines are phase-decoupled (see Example 1.4 and Exercise 1.10). This assumption is valid only if the lines are symmetric and the sources and loads are balanced such that currents and charges both sum to zero in these lines across phases; see Chapter 2.1.4. Otherwise an unbalanced three-phase model of transmission lines should be used; see Part III of this book.

2 If the lines are symmetric but the sources or loads are unbalanced then variables of different phases are coupled. A similarity transformation can be used to transform the system to a so called sequence coordinate in which the lines become decoupled and single-phase analysis can then be applied in the sequence coordinate; see Chapter 16 in Part III of this book.

1.2.4 Balanced systems in Δ configuration

Figure 1.16 shows the Δ configuration of a balanced voltage source and a balanced impedance. An ideal voltage source in Δ configuration is specified by its line voltage



(a) Balanced source (b) Balanced load

Figure 1.16 Balanced three-phase (a) voltage source E^{Δ} and (b) impedance z^{Δ} in Δ configuration.

$E^{\Delta} := \left(E^{ab}, E^{bc}, E^{ca}\right)$. It is *balanced* if E^{Δ} is a balanced vector according to Definition 1.1, i.e., assuming positive sequence:

$$E^{bc} = e^{-\mathbf{i}2\pi/3}\, E^{ab}, \quad E^{ca} = e^{\mathbf{i}2\pi/3}\, E^{ab}$$

A balanced three-phase system in Δ configuration enjoys the same properties as such a system in Y configuration in Chapter 1.2.3 does. In particular the line voltages sum to zero (see Figure 1.12(b)):

$$E^{ab} + E^{bc} + E^{ca} = 0$$

The three-phase voltages and currents in a balanced system in Δ configuration driven by balanced three-phase positive-sequence sources are balanced positive sequences. Moreover the phases are decoupled. We illustrate this in the next example.

Example 1.5 (Balanced three-phase system in Δ configuration)**.** Figure 1.17 shows a balanced three-phase source connected to a balanced three-phase load through balanced transmission lines in Δ configuration. The transmission lines have identical admittance $t \neq 0$ and the loads are of constant admittance $l \neq 0$. Suppose the internal voltage

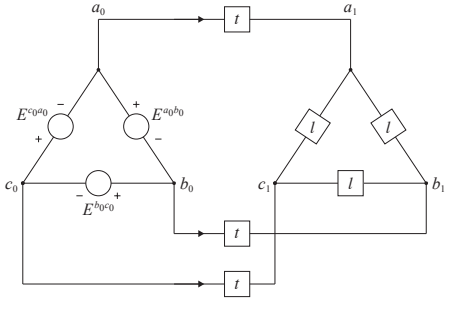


Figure 1.17 Example 1.5.

$E^\Delta := (E^{a_0b_0}, E^{b_0c_0}, E^{c_0a_0})$ is in a positive sequence. Denote the terminal current by $I := (I^{a_0a_1}, I^{b_0b_1}, I^{c_0c_1})$, the terminal voltage by $V := (V^{a_0a_1}, V^{b_0b_1}, V^{c_0c_1})$, and the line-to-line voltage by $U := (V^{a_1b_1}, V^{b_1c_1}, V^{c_1a_1})$. We will show that I, V, U are in balanced positive sequences, provided the ratio

$$\mu \;:=\; \frac{t}{l} \neq -3$$

Solution. Apply KCL at nodes a_1, b_1, c_1 to get (cf. (1.14)):

$$I \;=\; l\Gamma^{\mathsf{T}} U \;=\; tV \tag{1.29}$$

where Γ^{T} is defined in (1.12). Apply KVL to get

$$E^\Delta = U \,+\, \Gamma V \tag{1.30}$$

where Γ is defined in (1.12). Eliminate V from (1.29) and (1.30) to get

$$E^\Delta = \frac{1}{\mu}\left(\mu\mathbb{I} \,+\, \Gamma\Gamma^{\mathsf{T}}\right) U \;=\; \frac{1}{\mu}\begin{bmatrix} \mu+2 & -1 & -1 \\ -1 & \mu+2 & -1 \\ -1 & -1 & \mu+2 \end{bmatrix} U \tag{1.31}$$

where $\mu := t/l$ and $\mathbb{I}$ is the identity matrix of size 3. The matrix $\mu\mathbb{I} + \Gamma\Gamma^{\mathsf{T}}$ has a determinant of $\mu(\mu+3)^2$ and hence is nonsingular provided $\mu \neq 0, -3$. Since E^Δ is a balanced positive-sequence matrix we have

$$\left(\mu\mathbb{I} \,+\, \Gamma\Gamma^{\mathsf{T}}\right) U = \mu\, E^{ab}\, \alpha_+$$

It therefore suffices to show that α_+ is an eigenvector of $\mu\mathbb{I} + \Gamma\Gamma^{\mathsf{T}}$ with an associated eigenvalue λ, for then

$$U = \mu\, E^{ab} \left(\mu\mathbb{I} \,+\, \Gamma\Gamma^{\mathsf{T}}\right)^{-1} \alpha_+ \;=\; \frac{\mu\, E^{ab}}{\lambda}\, \alpha_+$$

showing that U is also a balanced positive-sequence voltage (note that if $Ax = \lambda x$ for a nonsingular matrix A then $A^{-1}x = \frac{1}{\lambda}x$). To show that α_+ is an eigenvector of $\mu\mathbb{I} + \Gamma\Gamma^{\mathsf{T}}$, we apply Theorem 1.2 to get

$$\left(\mu\mathbb{I} + \Gamma\Gamma^{\mathsf{T}}\right)\alpha_+ = \mu\alpha_+ + \Gamma(1-\alpha^2)\alpha_+ = \left(\mu + (1-\alpha)\left(1-\alpha^2\right)\right)\alpha_+ = \underbrace{(\mu+3)}_{\lambda}\alpha_+$$

as desired. This shows that U is indeed a balanced positive-sequence voltage. Indeed

$$U = \frac{\mu}{\mu+3}E^{\Delta}$$

To show that phase voltages V are also a balanced positive sequence and decoupled, use (1.29) and Corollary 1.3 to get

$$V = \frac{1}{\mu}\Gamma^{\mathsf{T}}U = \frac{1}{\mu}\left(1-\alpha^2\right)U = \frac{1-\alpha^2}{\mu+3}E^{\Delta}$$

Hence V is in a balanced positive sequence. The expression $I = tV$ from (1.29) then implies that the phase current I is also in a balanced positive sequence and that the phases are decoupled. □

Δ and Y transformation.

A balanced Δ-configured system also has a per-phase equivalent circuit. We now explain how to transform between Δ and Y configuration. This is the first step in per-phase analysis of balanced three-phase system described in Chapter 1.2.5 where all balanced devices in Δ configuration are transformed into their equivalent Y configuration, the per-phase circuit of the Y-equivalent network is then analyzed and the result translated back to the original system with Δ-configured devices. The validity of this procedure is formally proved in Chapter 16.3.4.

As explained in Chapter 1.2.1, given any balanced internal voltage $V^{\Delta} := (V^{ab}, V^{bc}, V^{ca})$ and current $I^{\Delta} := (I^{ab}, I^{ac}, I^{aa})$ in Δ configuration, an equivalent Y configuration is one that has the same external behavior, i.e., the internal voltage $V^{Y} := (V^{an}, V^{bn}, V^{cn})$ and current $I^{Y} := (I^{an}, I^{an}, I^{an})$ of the Y-equivalent satisfy (1.15) reproduced here

$$\Gamma V^{Y} = V^{\Delta}, \qquad I^{Y} = \Gamma^{\mathsf{T}} I^{\Delta}$$

Assume the neutral of the Y equivalent voltage source is the reference for all voltages and $V^{n} = 0$. Since V^{Y} and I^{Δ} are balanced vectors, Corollary 1.3 implies

$$(1-\alpha)V^{Y} = V^{\Delta}, \qquad I^{Y} = (1-\alpha^2)I^{\Delta}$$

Hence the Y-equivalent of $\left(Y^{\Delta}, I^{\Delta}\right)$ is

$$V^{Y} = \frac{1}{1-\alpha}V^{\Delta} = \frac{1}{\sqrt{3}e^{\mathbf{i}\pi/6}}V^{\Delta}, \quad I^{Y} = \left(1-\alpha^2\right)I^{\Delta} = \frac{\sqrt{3}}{e^{\mathbf{i}\pi/6}}I^{\Delta} \tag{1.32a}$$

This implies in particular that a voltage source E^Δ in Δ configuration has an equivalent Y-configured voltage source with $E^Y := (1-\alpha)^{-1}E^\Delta$. It also implies that a current source J^Δ in Δ configuration has an equivalent Y-configured current source with $J^Y := (1-\alpha^2)J^\Delta = \sqrt{3}e^{-\mathbf{i}\pi/6}J^\Delta$.

Consider a balanced three-phase impedance $z^\Delta \in \mathbb{C}$ in Δ configuration as shown in Figure 1.18(a). An Y-equivalent is a balanced impedance $z^Y \in \mathbb{C}$ as shown in Figure 1.18(b) so that their external behavior is the same, i.e., the terminal currents I are the same when the same line-to-line voltage V^{line} is applied to both impedances. Let

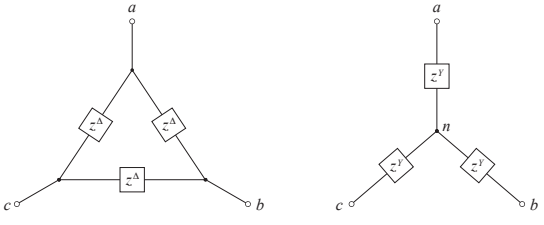


Figure 1.18 Δ-Y transformation of balanced loads: $z^Y = z^\Delta/3$.

$V^\Delta \in \mathbb{C}^3$ and $I^\Delta \in \mathbb{C}^3$ be the internal voltage and current across the impedance z^Δ in Δ configuration. Let $Z^\Delta := \text{diag}\left(z^\Delta, z^\Delta, z^\Delta\right)$. Then $V^\Delta = Z^\Delta I^\Delta$ and

$$V^{\text{line}} \;=\; V^\Delta \;=\; Z^\Delta I^\Delta, \qquad\qquad I \;=\; -\Gamma^{\mathsf{T}} I^\Delta \;=\; -(1-\alpha^2)I^\Delta$$

where the last equality follows from Corollary 1.3. Hence, for Δ-configured impedance, the line-to-line voltage V^{line} is related to the terminal current I according to

$$V^{\text{line}} \;=\; -\frac{1}{1-\alpha^2}\, Z^\Delta I$$

For the Y-equivalent, let $V^Y \in \mathbb{C}^3$ and $I^Y \in \mathbb{C}^3$ be its internal voltage and current across the impedance z^Y in Y configuration. Let $Z^Y := \text{diag}\left(z^Y, z^Y, z^Y\right)$. Then $V^Y = Z^Y I^Y$ and Corollary 1.3 implies

$$V^{\text{line}} \;=\; \Gamma V^Y \;=\; (1-\alpha)\, Z^Y I^Y, \qquad\qquad I \;=\; -I^Y$$

Hence, for Y-configured impedance, the line-to-line voltage V^{line} is related to the terminal current I according to

$$V^{\text{line}} \;=\; -(1-\alpha) Z^Y I$$

The relationships between the line-to-line voltage V^{line} and the terminal current I for both the Δ-configured impedance and its Y-equivalent will be identical if and only if

$$z^Y = \frac{z^\Delta}{(1-\alpha)(1-\alpha^2)} \;=\; \frac{z^\Delta}{3} \tag{1.32b}$$

The corresponding admittances $y^Y := \left(z^Y\right)^{-1}$ and $y^\Delta := \left(z^\Delta\right)^{-1}$ are related by $y^Y = 3y^\Delta$.

1.2.5 Per-phase analysis for balanced systems

A balanced three-phase system consists of balanced three-phase sources and loads connected by balanced (identical) transmission lines. Given a balanced three-phase system with all sources and loads *in Y configuration*, assuming there is no mutual inductance between phases, then

- all the neutrals are at the same potential;
- all phases are decoupled;
- all corresponding network variables are in balanced sets of the same sequence as the sources.

These properties lead to equivalent per-phase circuits, as explained in Chapter 1.2.3. Even though we have only illustrated these properties for simple systems, they hold more generally. They allow us to study such a system by analyzing a single phase, say, phase a. The corresponding variables in phases b and c lags those in phase a by 120° and 240° respectively when abc is a positive sequence, and by 240° and 120° respectively when abc is a negative sequence.

When some or all of the sources and loads are in Δ configuration, the phases are still decoupled and can be analyzed separately. To obtain the equivalent per-phase circuit, however, we first transform each Δ-configured device into an equivalent Y-configured device using the transformation (1.32a) for voltage sources and (1.32b) for impdances. We then analyze the equivalent circuit that consists of only Y-configured devices. Finally we translate the results for equivalent Y configuration back to the corresponding quantities in Δ configuration.

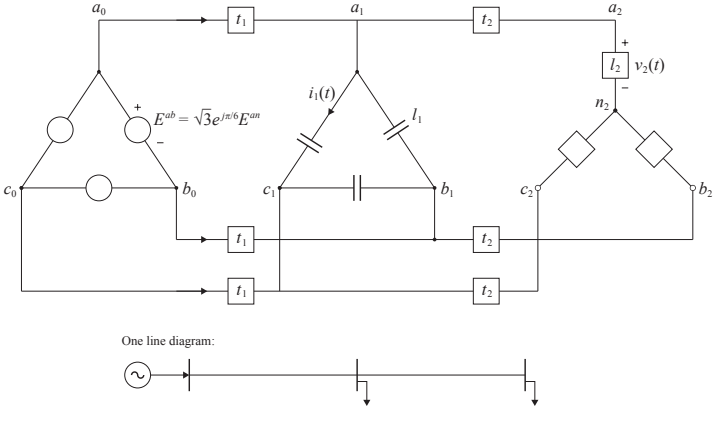
We emphasize that these transformations hold only in the balanced case with balanced sources, identical impedances, and symmetric transmission lines. Moreover the equivalence of these two configurations is with respect to their external behavior (V^{ab}, I^a, etc); for internal behavior, we have to analyze the original circuit; see Example 1.6.

In summary, the procedure for per-phase analysis is:

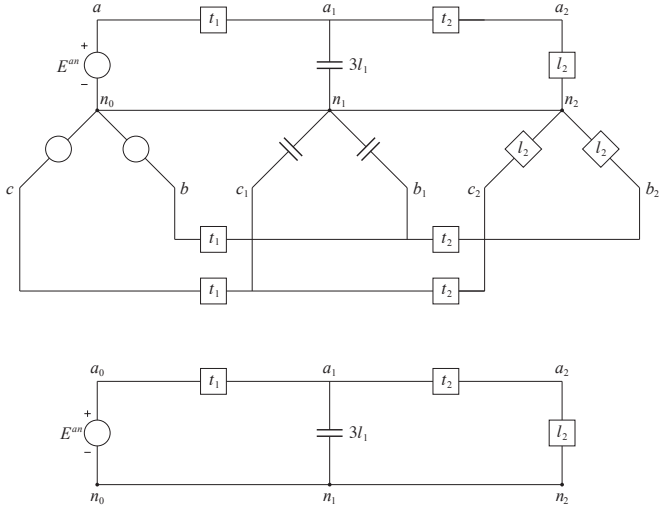
1 Convert all sources and loads in Δ configuration into their equivalent Y configurations using (1.32a) for sources and (1.32b) for loads.
2 Solve for the desired phase a variables using phase a circuit with all neutrals connected.
3 For positive-sequence sources, the phase b and c variables are determined by subtracting 120° and 240° respectively from the corresponding phase a variables. For negative-sequence sources, add 120° and 240° instead.
4 If variables in the internal of a Δ configuration are desired, derive them from the original circuits.

This procedure is formally justified in Chapter 16.3.4. We illustrate it with an example.

Example 1.6 (Per-phase analysis)**.** Consider the balanced three-phase system shown in Figure 1.19. The three-phase sources are a balanced positive sequence in Δ configuration with line voltage $E^{ab} = \sqrt{3}e^{\mathbf{i}\pi/6}E^{an}$, etc. The Δ-configured loads are balanced with identical admittances l_1, and the Y-configured loads are balanced with identical admittances l_2. The transmission lines are modeled by admittances t_1 and t_2. Find the current $i_1(t)$ and voltage $v_2(t)$ in the diagram. Assume $3l_1l_2 + 3l_1t_2 + l_2(t_1+t_2) + t_1t_2 \neq 0$.



(a) Balanced three-phase system



(b) Equivalent per-phase system

Figure 1.19 Balanced three-phase system and its per-phase equivalent circuit. The balanced three-phase loads have admittances l_1 and l_2, and the transmission lines have admittances t_1 and t_2.

Solution. First we convert the Δ sources to their equivalent Y sources using (1.32a) and Δ loads to their equivalent Y loads using (1.32b). The result is shown in the upper panel of Figure 1.19(b). Then we construct the equivalent per-phase circuit with all neutrals n, n_1, n_2 connected, as shown in the lower panel of Figure 1.19(b).

We analyze the per-phase circuit to solve for voltages

$$V_1 := V^{a_1 n_1} \text{ and } V_2 := V^{a_2 n_2}$$

Applying KCL to nodes a_1 and a_2 we get

$$\begin{aligned} t_1 (E^{an} - V_1) &= 3 l_1 V_1 + t_2 (V_1 - V_2) \\ t_2 (V_1 - V_2) &= l_2 V_2 \end{aligned}$$

Hence

$$\begin{bmatrix} 3l_1 + t_1 + t_2 & -t_2 \\ t_2 & -(l_2 + t_2) \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} t_1 E^{an} \\ 0 \end{bmatrix}$$

By assumption, the determinant

$$\Delta := -(3 l_1 l_2 + 3 l_1 t_2 + l_2 (t_1 + t_2) + t_1 t_2)$$

is nonzero. Hence

$$\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \frac{1}{\Delta} \begin{bmatrix} -(l_2 + t_2) & t_2 \\ -t_2 & 3l_1 + t_1 + t_2 \end{bmatrix} \begin{bmatrix} t_1 E^{an} \\ 0 \end{bmatrix} = \frac{-t_1 E^{an}}{\Delta} \begin{bmatrix} l_2 + t_2 \\ t_2 \end{bmatrix} \tag{1.33}$$

Since $V^{a_2 n_2} = V_2$, we get:

$$v_2(t) = \sqrt{2}\, |V_2| \cos(\omega t + \angle V_2)$$

where ω is the steady-state system frequency and V_2 is given by (1.33). To calculate

$$i_1(t) = \sqrt{2}\, |I^{a_1 c_1}| \cos(\omega t + \angle I^{a_1 c_1}) \tag{1.34}$$

we use (1.32a) to first get

$$V^{a_1 b_1} = \sqrt{3}\, e^{\mathbf{i}\pi/6} V_1$$

where V_1 is given by (1.33). Hence

$$I^{a_1 b_1} = l_1 V^{a_1 b_1} = \sqrt{3}\, l_1\, e^{\mathbf{i}\pi/6} V_1$$

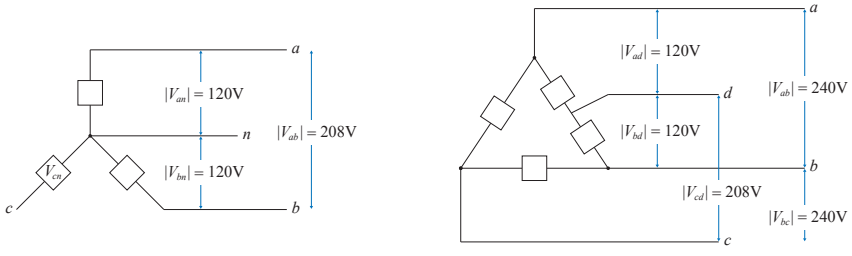
Since the sources are a positive sequence we have

$$I^{a_1 c_1} = -I^{a_1 a_1} = -I^{a_1 b_1} e^{\mathbf{i}2\pi/3} = -\sqrt{3}\, e^{\mathbf{i}5\pi/6} 3 l_1 V_1 = 3\sqrt{3}\, e^{-\mathbf{i}\pi/6} l_1 V_1$$

where V_1 is given by (1.33). Substituting $I^{a_1 c_1}$ into (1.34) yields $i_1(t)$. □

1.2.6 Example configurations and line limits

The secondary sides of three-phase distribution transformers in the US are commonly configured as shown in Figure 1.20. For our purposes we can treat them as balanced three-phase sources. Figure 1.20(a) shows the secondary side of a typical 5-wire three-



(a) $208Y/120V$ 3-phase Y (b) $240V$ split phase Δ

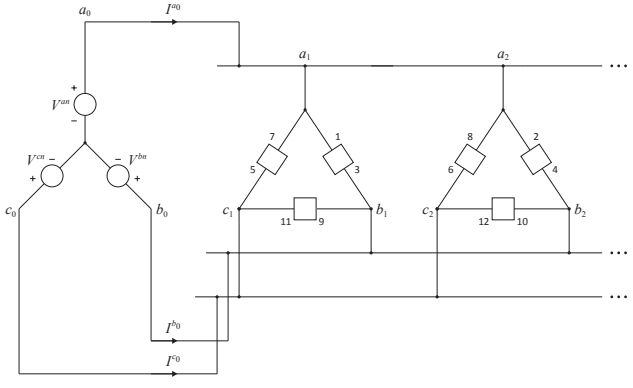
Figure 1.20 Common distribution transformer configurations.

phase transformer in Y configuration. Three phase wires (labeled a, b, c) and a neutral wire (labeled n) are shown. The fifth wire, not shown, is the earth ground wire, typically connected to neutral. A different voltage magnitude can be supplied to a load depending on how it is connected. The voltage magnitude between a phase wire and the neutral is 120V and that between a pair of phase wires is $120\sqrt{3}\text{V} = 208\text{V}$.

Figure 1.20(b) shows a 5-wire transformer in Δ configuration with one of the phases center-tapped to provide three voltage levels. Four phase wires (labeled a, b, c, d) are shown but an earth ground wire is not shown. The voltage magnitude between wires ad or bd is 120V, whereas that between wire cd is 208V (derive this). The line-to-line voltage magnitude is 240V.

Line limits. Figure 1.21(a) shows a Y-configured voltage source connected to a set of loads in Δ configuration. The voltage source is the secondary side of a three-phase $208Y/120\text{V}$ transformer shown in Figure 1.20(a). The voltage magnitude across each load is the line-to-line voltage 208V. Figure 1.21(b) shows the electric panel arrangement to connect the loads to the voltage source. The dot in the first row indicates that the wires numbered 1 and 2 are connected to phase a, the dot in the second row indicates that the wires numbered 3 and 4 are connected to phase b, the dot in the third row indicates that the wires numbered 5 and 6 are connected to phase c, and so on. Therefore the load connected between wires 1 and 3 is connected between phase a and phase b lines (see the corresponding labels on the loads in Figure 1.21(a)). Similarly for the load connected between wires 2 and 4, and other loads connected between different phases.

We are interested in the currents $J_0 := (I^{a_0a_1}, I^{b_0b_1}, I^{c_0c_1})$ supplied by the three-



(a) Voltage source supplying loads in parallel

(b) Panel arrangement

Figure 1.21 (a) Three-phase voltage source connected to loads in parallel. (b) Three-phase panel used to connect loads in parallel to the voltage source.

phase source to the loads. Suppose the wires connecting the three-phase source to the loads are rated at $I^{\max}$. Then we require that the current magnitude in each phase be bounded by $I^{\max}$:

$$|I^{p_0 p_1}| \leq I^{\max}, \qquad p = a, b, c \tag{1.35}$$

Suppose the loads are not impedance loads, but constant current loads that draw specified currents. Let the current drawn by the load in Figure 1.21(a) between wires 1 and 3 be $I^{a_1 b_1}$, that between wires 9 and 11 be $I^{b_1 c_1}$, that between wires 5 and 7 be $I^{c_1 a_1}$. In general, let the load currents in the kth three-phase load be $I_k := (I^{a_k b_k}, I^{b_k c_k}, I^{c_k a_k})$. We now derive bounds on the load currents $(I_k, k = 1, \ldots, K)$ that enforce the line limits (1.35).

Before proceeding, we mention as an example application the smart charging of electric vehicles where each load is a vehicle. We are to design an algorithm that determines the charging rate, i.e., current magnitude $|I^{p_k q_k}|$, for each vehicle to optimize certain objective subject to capacity constraints such as (1.35) and other constraints. Such an algorithm can be applied periodically, e.g., every minute, to update the charging rates. Note that in this kind of applications, the system is unbalanced since the loads $|I^{p_k q_k}|$ are generally not identical across phases, but here we ignore the effect of wires connecting these devices.

Applying KCL at nodes (a_1, b_1, c_1) we have

$$\underbrace{\begin{bmatrix} I^{a_0a_1} \\ I^{b_0b_1} \\ I^{c_0c_1} \end{bmatrix}}_{J_0} = \underbrace{\begin{bmatrix} 1 & 0 & -1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}}_{\Gamma^{\mathsf{T}}} \underbrace{\begin{bmatrix} I^{a_1b_1} \\ I^{b_1c_1} \\ I^{c_1a_1} \end{bmatrix}}_{I_1} + \underbrace{\begin{bmatrix} I^{a_1a_2} \\ I^{b_1b_2} \\ I^{c_1c_2} \end{bmatrix}}_{J_1}$$

where $J_k := (I^{a_k a_{k+1}}, I^{b_k b_{k+1}}, I^{c_k c_{k+1}})$, $k = 0, \ldots, K-1$, are the line currents from stage k to stage $k+1$. In general we have

$$J_k = \Gamma^{\mathsf{T}} I_k \;+\; J_{k+1}, \qquad k = 0, \ldots, K-1$$

Hence the total supply currents are given by

$$J_0 = \Gamma^{\mathsf{T}} (I_0 + I_1 + \cdots + I_K) \tag{1.36}$$

when there are K three-phase constant current loads. Note that this expression does not require that the loads are balanced. In particular, if a load (say) $I^{a_k b_k}$ is absent, then we set $I^{a_k b_k} = 0$ in (1.36).

Let the total load current in each leg of the Δ configuration be denoted by

$$I^{ab} := \sum_{k=1}^{K} I^{a_k b_k}, \quad I^{bc} := \sum_{k=1}^{K} I^{a_k c_k}, \quad I^{ca} := \sum_{k=1}^{K} I^{a_k a_k} \tag{1.37}$$

Then (1.36) can be written in terms of the total load currents as:

$$\begin{bmatrix} I^{a_0a_1} \\ I^{b_0b_1} \\ I^{c_0c_1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & -1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} I^{ab} \\ I^{bc} \\ I^{ca} \end{bmatrix}$$

The line limits (1.35) are therefore

$$\begin{aligned} |I^{a_0a_1}| = |I^{ab} - I^{ca}| &\le I^{\max} \\ |I^{b_0b_1}| = |I^{bc} - I^{ab}| &\le I^{\max} \\ |I^{c_0c_1}| = |I^{ca} - I^{bc}| &\le I^{\max} \end{aligned}$$

Enforcing line limits requires one to know not just the magnitudes of the load currents, but also their phases in order to compute their sums. As explained in the caption of

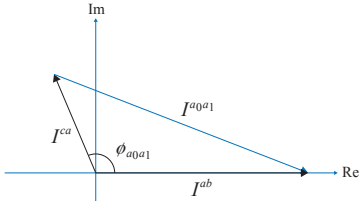


Figure 1.22 $I^{a_0a_1} = I^{ab} - I^{ca}$. Hence by the cosine rule $|I^{a_0a_1}|^2 = |I^{ab}|^2 + |I^{ca}|^2 - 2\,|I^{ab}|\,|I^{ca}| \cos\phi$ where $\phi_{a_0a_1} := \angle I^{ca} - \angle I^{ab}$ is the angle between I^{ab} and I^{ca}.

Figure 1.22, these inequalities are equivalent to:

$$|I^{ab}|^2+|I^{ca}|^2 \;-\; 2|I^{ab}||I^{ca}|\cos\phi_{a_0a_1} \le (I^{\max})^2 \tag{1.39a}$$

$$|I^{bc}|^2+|I^{ab}|^2 \;-\; 2|I^{bc}||I^{ab}|\cos\phi_{b_0b_1} \le (I^{\max})^2 \tag{1.39b}$$

$$|I^{ca}|^2+|I^{bc}|^2 \;-\; 2|I^{ca}||I^{bc}|\cos\phi_{c_0c_1} \le (I^{\max})^2 \tag{1.39c}$$

If we know the angles $\phi_{p_0p_1}$, $p=a,b,c$, between the total load currents (I^{ab},I^{bc},I^{ca}) in each leg of the Δ configuration, then (1.39) are convex quadratic constraints on the magnitudes of (I^{ab},I^{bc},I^{ca}). We next consider several special cases and derive simple bounds on the magnitudes $(|I^{a_kb_k}|,|I^{b_kc_k}|,|I^{c_ka_k}|)$ of the individual load currents that will enforce (1.39).

Assumption 1: Current phasors $I^{a_kb_k}$ have the same, and known, phase angle θ_{ab} for all k; similarly for $I^{b_kc_k}$ and $I^{c_ka_k}$. From (1.37) we have

$$I^{ab} := e^{\mathbf{i}\theta_{ab}}\sum_{k=1}^{K}\left|I^{a_kb_k}\right|, \quad I^{bc} := e^{\mathbf{i}\theta_{bc}}\sum_{k=1}^{K}\left|I^{b_kc_k}\right|, \quad I^{ca} := e^{\mathbf{i}\theta_{ca}}\sum_{k=1}^{K}|I^{c_ka_k}|$$

and constraints (1.39a) become

$$\left(\sum_{k=1}^{K}\left|I^{a_kb_k}\right|\right)^2+\left(\sum_{k=1}^{K}|I^{a_ka_k}|\right)^2-2\left(\sum_{k=1}^{K}\left|I^{a_kb_k}\right|\right)\left(\sum_{k=1}^{K}|I^{a_ka_k}|\right)\cos\phi_{a_0a_1}\le(I^{\max})^2 \tag{1.40}$$

where $\cos\phi_{a_0a_1}:=\theta_{ca}-\theta_{ab}$ is known. Similarly for constraints (1.39b) and (1.39c). These are quadratic constraints in the magnitudes $(|I^{a_kb_k}|,|I^{a_kc_k}|,|I^{a_ka_k}|)$ of the individual load currents that will enforce (1.39), given the angles $\phi_{p_0p_1}$, $p=a,b,c$, between the load currents in different legs of the Δ configuration.

Assumption 2: In addition to Assumption 1, the angles $\phi_{p_0p_1}=120^\circ$, for $p=a,b,c$. Then $\cos\phi_{p_0p_1}=-1/2$ and (1.40) becomes

$$\left(\sum_{k=1}^{K}\left|I^{a_kb_k}\right|\right)^2+\left(\sum_{k=1}^{K}|I^{a_ka_k}|\right)^2+\left(\sum_{k=1}^{K}\left|I^{a_kb_k}\right|\right)\left(\sum_{k=1}^{K}|I^{a_ka_k}|\right)\le(I^{\max})^2 \tag{1.41}$$

Similarly for constraints (1.39b) and (1.39c).

Assumption 3 (balanced case): All load currents have the same magnitude and the phases of currents on different legs of the Δ differ by 120°. That is, assuming positive sequence, for all $k=1,\ldots,K$, we have

$$I^{a_kb_k} \;=\; I\,e^{\mathbf{i}\theta_{ab}}, \quad I^{b_kc_k} \;=\; I\,e^{\mathbf{i}\theta_{bc}}, \quad I^{c_ka_k} \;=\; I\,e^{\mathbf{i}\theta_{ca}}$$

where I is the common magnitude of the load currents, and

$$\theta_{ab}-\theta_{bc} \;=\; 120^\circ, \quad \theta_{bc}-\theta_{ca} \;=\; 120^\circ, \quad \theta_{ca}-\theta_{ab} \;=\; 120^\circ$$

Then the constraint (1.41) reduces to $3K^2I^2\le(I^{\max})^2$, or a bound on the common

magnitude I of individual load currents

$$I \leq \frac{I^{\max}}{\sqrt{3}K} \tag{1.42}$$

Linear bounds. Many applications operate in unbalanced conditions, e.g., adaptive electric vehicle charging where the magnitudes $|I^{p_k q_k}|$ of the load currents are to be determined and generally different. In these cases there are two difficulties with the line limits (1.40) and (1.41). First the angles $(\theta_{ab}, \theta_{bc}, \theta_{ca})$ may not be known. Second even when these angles are known, the constraints are quadratic which can be computationally too expensive to implement in real time in inexpensive devices. In this case, we can impose linear constraints which are simpler but more conservative.

Take phase a as an example. Since $|I^{a_0 a_1}| = |I^{ab} - I^{ca}| \leq |I^{ab}| + |I^{ca}|$, a simple limit on the load currents that enforces $|I^{a_0 a_1}| \leq I^{\max}$ is to require

$$|I^{ab}| + |I^{ca}| \leq I^{\max}$$

i.e., the sum of the magnitudes of the total load currents in legs ab and ca should be less than the current rating $I^{\max}$. From (1.37) we have $\left|I^{ab}\right| = \left|\sum_k I^{a_k b_k}\right| \leq \sum_k |I^{a_k b_k}|$. Hence a simple linear bound on the load current magnitudes is:

$$\sum_{k=1}^{K} \left(|I^{a_k b_k}| + |I^{a_k a_k}| \right) \leq I^{\max} \tag{1.43}$$

The constraints on phases b and c are similar.

For a balanced system we can easily assess how conservative the bound (1.43) is compared with the exact limit (1.42) on the load currents. In the balanced case the bound (1.43) reduces to

$$I \leq \frac{I^{\max}}{2K}$$

Hence it is $\sqrt{3}/2 \sim 87\%$ of that in (1.42), i.e., it is conservative by $\sim 13\%$ for a balanced system.

1.3 Complex power

1.3.1 Single-phase power

Instantaneous power.

When a voltage $v(t)$ is applied across two ports and a current $i(t)$ flows between them, as shown in Figure 1.23(a), energy is delivered to the network that connects the ports.

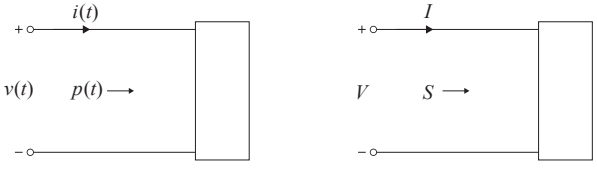
We define the *instantaneous power* supplied as:

$$p(t) := v(t)i(t) = \frac{V_{\max} I_{\max}}{2}(\cos(\theta_V - \theta_I) + \cos(2\omega t + \theta_V + \theta_I)) \tag{1.44}$$

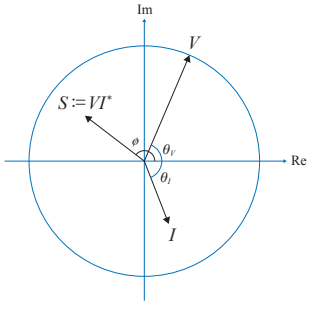
Since the last term inside the bracket of (1.44) is sinusoidal with twice the nominal frequency ω the average power delivered is

$$\frac{1}{T}\int_0^T p(t)dt = \frac{V_{\max} I_{\max}}{2}\cos(\theta_V - \theta_I)$$

where $T := 2\pi/\omega$.



(a) Instantaneous power (b) Complex power



(c) $S = V I^*$

Figure 1.23 Definition of power

Complex power.

Define the *complex power* in terms of the voltage and current phasors as:

$$S := VI^* = \frac{V_{\max} I_{\max}}{2} e^{\mathbf{i}(\theta_V - \theta_I)} = |V||I|e^{\mathbf{i}\phi} \tag{1.45}$$

where I^* denotes the complex conjugate of I. See Figures 1.23(b) and (c). Here $\phi := \theta_V - \theta_I$ is called the *power factor angle* and $\cos\phi$ is called the *power factor (PF)*. Power engineers often says *leading* or *lagging* power factor: here *lagging* means

current I lags voltage V so that $\phi > 0$. A leading power factor has $\phi < 0$. A unity power factor means $\phi = 0$. Figure 1.24 shows four complex powers

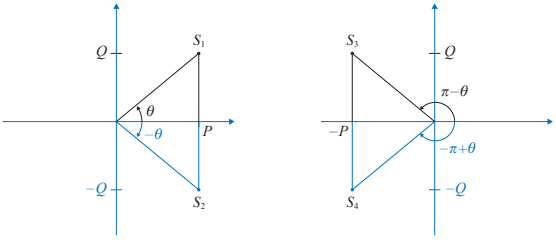


Figure 1.24 Power factor angles ϕ and power factor $\cos\phi$.

$$S_1 := P + \mathbf{i}Q, \qquad S_2 := P - \mathbf{i}Q, \qquad S_3 := -P + \mathbf{i}Q, \qquad S_2 := -P - \mathbf{i}Q$$

with power factor angles $\phi_1 := \theta$, $\phi_2 := -\theta$, $\phi_3 := \pi - \theta$, and $\phi_4 := -\pi + \theta$ respectively. Here $P, Q > 0$ and $\theta \in [0, \pi]$. Their power factors are

$$\cos\phi_1 = \frac{P}{\sqrt{P^2+Q^2}} = \cos\phi_2, \qquad \cos\phi_3 = \frac{-P}{\sqrt{P^2+Q^2}} = \cos\phi_4$$

Therefore power factor $\cos\phi_i$ does not differentiate between S_1 and S_2. Power engineers specify S_1 as power factor $\cos\theta$ lagging ($\phi_1 > 0$ and therefore $Q_1 := Q > 0$) and S_2 as power factor $\cos\theta$ leading ($\phi_2 < 0$ and $Q_2 := -Q < 0$). Similarly S_3 has a power factor $-\cos\theta$ lagging ($\phi_3 > 0$ and $Q_3 := Q > 0$) and S_4 has a power factor $-\cos\theta$ leading ($\phi_4 < 0$ and $Q_4 := -Q < 0$). For example "a load draws 100kW at a power factor of 0.707 leading" means that the real power $\mathrm{Re}(S) = 100$ kW and $\cos\phi = \frac{1}{\sqrt{2}}$. Since the power factor is leading, $\phi = -45°$ and $S = 100 - j100$ kVA .

Note that S is *not* a phasor because $\sqrt{2}|S|\cos(\omega t + \phi)$ is not the instantaneous power in the time domain. This complex quantity is important in power flow analysis in the phasor domain, as we will see. The real part of S

$$P := |V||I|\cos\phi$$

is called the *active* or *real power* and its unit is W (watt). The imaginary part of S

$$Q := |V||I|\sin\phi$$

is called the *reactive power* and its unit is var (volt-ampere reactive). We write both $S = P + jQ$ and $S = |V||I|e^{\mathbf{i}\phi}$. The magnitude $|S| = |V||I|$ is called the *apparent power* and its unit is VA (volt-ampere). Given an active power P and a power factor $\cos\phi$, the complex power S is given by (since $P = |S|\cos\phi$)

$$S = \frac{P}{\cos\phi}\, e^{\mathbf{i}\phi}$$

i.e. the complex power is completely determined by the active power P and the power

factor angle ϕ. Power is balanced at every node in a network. If I_{jk} and S_{jk} are sending-end current and power respectively from node j to node k, then power balance at node j means $\sum_k S_{jk} = 0$. This is a consequence of KCL $\sum_k I_{jk} = 0$ and the definition of branch power $S_{jk} := V_j I_{jk}^*$.

Relation between instantaneous and complex power.

The complex power S in the phasor domain is related to the instantaneous power in the time domain as follows. We can use (1.44) to express the instantaneous power $p(t)$ in terms of active power P and reactive power Q as (Problem 1.11):

$$p(t) = P + P\cos 2(\omega t + \theta_I) - Q\sin 2(\omega t + \theta_I) \tag{1.46}$$

It is then clear that the active power P is equal to the average power delivered (in the time domain):

$$P = \frac{1}{T}\int_0^T p(t)dt$$

as the last two terms in (1.46) average to zero over a cycle T. The reactive power Q determines the magnitude of the instantaneous power $p(t)$.

Power delivered to an impedance.

The current and voltage across an impedance z is related by Ohm's law, $V = zI$ and hence

$$|z| = \frac{|V|}{|I|} \text{ and } \angle z = \theta_V - \theta_I =: \phi$$

Therefore from (1.45)

$$S = z|I|^2 = |z||I|^2 e^{\mathbf{i}\phi}$$

and

$$P = |z||I|^2\cos\phi \text{ and } Q = |z||I|^2\sin\phi$$

The active and reactive power for the three passive elements are given in Table 1.2.

	$\lvert z\rvert$	$\phi = \angle z$	P	Q
Resistor $z = r$	r	0	$r\lvert I\rvert^2$	0
Inductor $z = \mathbf{i}\omega l$	ωl	$\pi/2$	0	$\omega l\lvert I\rvert^2$
Capacitor $z = (\mathbf{i}\omega c)^{-1}$	$(\omega c)^{-1}$	$-\pi/2$	0	$-(\omega c)^{-1}\lvert I\rvert^2$

Table 1.2 Power delivered to RLC elements.

Therefore the power delivered to a resistor is active ($Q = 0$). The instantaneous power $p(t) := v(t)i(t)$ is

$$p(t) := ri^2(t) \;=\; rI_{\max}^2 \cos^2(\omega t + \theta_I) \;=\; P(1 + \cos 2(\omega t + \theta_I))$$

which is (1.46). Table 1.2 also implies that the complex power delivered to an inductor or a capacitor is reactive. Substituting into (1.46), the instantaneous power $p(t)$ to a purely reactive load depends only on the reactive power Q:

$$p(t) = \begin{cases} -Q \sin 2(\omega t + \theta_I) & \text{for inductor } z = j\omega l \\ Q \sin 2(\omega t + \theta_V) & \text{for capacitor } z = (j\omega c)^{-1} \end{cases}$$

i.e., the net (average) power delivered to the load is zero and the instantaneous power is sinusoidal with twice the frequency and has an amplitude Q.

Example 1.7. Suppose $z = j\omega l$ (inductance) or $z = (j\omega c)^{-1}$ (capacitance). Prove directly in time domain that the average delivered power is 0 and the amplitude of the instantaneous power is Q.

Solution: Suppose power is delivered to an inductor $z = j\omega l$. Let the current be $i(t) = I_{\max} \cos(\omega t + \theta_I)$. Then the voltage $v(t)$ across the inductor is given by

$$v(t) = l\frac{di}{dt}(t) \;=\; -\omega l\, I_{\max} \sin(\omega t + \theta_I)$$

and therefore

$$\begin{aligned} p(t) = v(t)i(t) \;&=\; -\omega l\, I_{\max}^2 \sin(\omega t + \theta_I)\cos(\omega t + \theta_I) \\ &= -\omega l \frac{I_{\max}^2}{2} \sin 2(\omega t + \theta_I) \;=\; -\omega l\, |I|^2 \sin 2(\omega t + \theta_I) \\ &= -Q \sin 2(\omega t + \theta_I) \end{aligned}$$

where the last equality follows from $Q = |z||I|^2 \sin \angle z = \omega l |I|^2$ since $\angle z = \frac{\pi}{2}$. Moreover the average power delivered is

$$P = \frac{1}{T}\int_0^{\mathsf{T}} p(t)dt \;=\; 0$$

The case of capacitor load $z = (j\omega c)^{-1}$ is similar and omitted (see Exercise 1.13). □

1.3.2 Three-phase power

Under balanced three-phase operation, the total instantaneous power delivered is constant and the total complex power is 3 times the per-phase complex power.

Indeed, for a balanced three-phase positive-sequence source, we have

$$V^{bn} = V^{an} e^{-\mathrm{i}2\pi/3}, \;\; I^{an} = I^{an} e^{-\mathrm{i}2\pi/3} \;\text{ and }\; V^{cn} = V^{an} e^{\mathrm{i}2\pi/3}, \;\; I^{an} = I^{an} e^{\mathrm{i}2\pi/3}$$

Hence

$$S_{3\phi} = V^{an} I^{an\mathsf{H}} + V^{bn} I^{an\mathsf{H}} + V^{cn} I^{an\mathsf{H}} \;=\; 3V^{an} I^{an\mathsf{H}} \;=\; 3S$$

where $S := V^{an} I^{an\mathsf{H}}$ is the per-phase complex power.

For instantaneous power, we have from (1.44), for a balanced three-phase positive-sequence source,

$$\begin{aligned} p_{3\phi}(t) &:= v^a(t)i^a(t) + v^b(t)i^b(t) + v^c(t)i^c(t) \\ &= |V^a||I^a|\left(\cos\phi + \cos(2\omega t + \theta_V + \theta_I)\right) \\ &\quad + |V^a||I^a|\left(\cos\phi + \cos(2\omega t + (\theta_V - 2\pi/3) + (\theta_I - 2\pi/3))\right) \\ &\quad + |V^a||I^a|\left(\cos\phi + \cos(2\omega t + (\theta_V + 2\pi/3) + (\theta_I + 2\pi/3))\right) \\ &= 3|V^a||I^a|\cos\phi \;+\; |V^a||I^a|\left(\cos\theta(t) + \cos(\theta(t) - 4\pi/3) + \cos(\theta(t) + 4\pi/3)\right) \\ &= 3P \end{aligned}$$

where $\theta(t) := 2\omega t + \theta_V + \theta_I$ and P is the per-phase active power. Here the last equality follows from

$$\cos x + \cos(x - 4\pi/3) + \cos(x + 4\pi/3) = \mathrm{Re}\left(e^{\mathbf{i}x} + e^{\mathbf{i}(x-4\pi/3)} + e^{\mathbf{i}(x+4\pi/3)}\right)$$

and

$$\left(e^{\mathbf{i}x} + e^{\mathbf{i}(x-4\pi/3)} + e^{\mathbf{i}(x+4\pi/3)}\right) = \left(e^{\mathbf{i}x} + e^{\mathbf{i}(x+2\pi/3)} + e^{\mathbf{i}(x-2\pi/3)}\right) \;=\; 0$$

where the last equality follows from Theorem 1.2.

1.3.3 Advantages of three-phase power

There are two main advantages of balanced three-phase systems over a system with a single phase or that with other polyphases.

First it offers several benefits to motor operation. The total instantaneous power $p_{3\phi}(t) = 3P$ delivered is constant over time in a balanced three-phase system. On a generator or motor this produces a constant mechanical torque, reducing vibrations, noise, wear and tear, and other mechanical issues. A three-phase system can also self-start an induction motor.

In contrast, the instantaneous power

$$p_{1\phi}(t) = P \;+\; |V||I|\cos(2\omega t + \theta_V + \theta_I) \;\;=:\;\; P + |V||I|\cos\theta(t)$$

in a single-phase system, where $\theta(t) := 2\omega t + \theta_V + \theta_I$, is a sinusoidal signal with twice the system frequency. This is the case also with a two-phase system where the

instantaneous power is

$$\begin{aligned} p_{2\phi}(t) &= |V^a||I^a|\,(\cos\phi + \cos(2\omega t + \theta_V + \theta_I)) + |V^a||I^a|\,(\cos\phi + \cos(2\omega t + (\theta_V + \pi) + (\theta_I + \pi))) \\ &= |V^a||I^a|\,(2\cos\phi + \cos\theta(t) + \cos(\theta(t) + 2\pi)) \\ &= P \;+\; 2|V^a||I^a|\cos\theta(t) \end{aligned}$$

It can be shown that for $K \geq 3$, a balanced K-phase system has $p_{K\phi}(t) = KP$ independent of t (Exercise 1.12). Even though a balanced four-phase system also has time-invariant instantaneous power, its design is more complex than a three-phase system.

Second a three-phase system typically saves materials and thermal loss ($r|I^2|$) compared with a single-phase system that serves the same load. For example, it is clear that the single-phase system that consists of three identical subsystems shown in Figure 1.7(a) needs twice as much transmission line and incurs twice as much thermal loss in transmission as the balanced three-phase system in Figure 1.7(b), since the balanced three-phase system has zero return current and hence does not need a neutral line.

The following example compares a balanced three-phase system with a single one-phase circuit with a higher ampacity, as opposed to three identical subcircuits in Figure 1.7(a), to supply the same load. The same conclusion holds that the three-phase system needs half as much conductor and incurs half as much transmission loss.

Example 1.8 (Single-phase vs three-phase systems)**.** Consider two systems that deliver a specified apparent power $|S|$ at a specified voltage magnitude $|V|$ to a constant power load, as shown in Figure 1.25. The distance between the generation and the load is d. The first system is single-phased and the second system is balanced three-phased. Compare the required amount of wire and thermal loss in the line in these systems.

The line has an impedance $z := r + jx$ per unit length where the resistance r per unit length is inversely proportional to the area of the line with proportionality constant ρ. The current density limit of the line is δ in ampere per unit area.

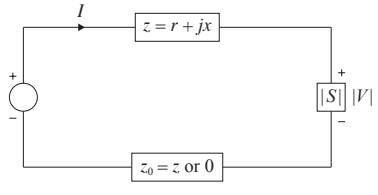


Figure 1.25 A system that delivers power $|S|$ to a load at voltage $|V|$. The distance between the generation and the load is d. The line has an impedance $z := r + jx$ per unit length.

Solution. A single-phase system requires two cables, one for return current, each carrying a current of magnitude $|I_{1\phi}| = |S|/|V|$. This is illustrated in Figure 1.25 with $z_0 = z$. A balanced three-phase system requires three cables, each carrying a per-phase

apparent power of $|S|/3$ and a per-phase current of magnitude $|I_{3\phi}| = |S|/(3|V|)$. The per-phase equivalent circuit is illustrated in Figure 1.25 with $z_0 = 0$.

For the single-phase system the required cross-sectional area of the cable is

$$A_{1\phi} := \frac{|I_{1\phi}|}{\delta} = \frac{|S|}{\delta|V|}$$

Hence the amount of material (volume of the cable) required is

$$m_{1\phi} := 2\,A_{1\phi}d = 2\frac{d|S|}{\delta|V|}$$

Moreover the resistance per-unit length of the cable is

$$r_{1\phi} := \frac{\rho}{A_{1\phi}} = \frac{\rho\delta|V|}{|S|}$$

and hence the active power loss in the cable is

$$l_{1\phi} := 2r_{1\phi}|I_{1\phi}|^2 d = \frac{2\rho\delta|V|}{|S|}\cdot\frac{d|S|^2}{|V|^2} = 2\frac{\rho\delta d|S|}{|V|}$$

For the balanced three-phase system the required cross-sectional area of the cable in each phase is

$$A_{3\phi} := \frac{|I_{3\phi}|}{\delta} = \frac{|S|}{3\delta|V|}$$

Hence the amount of material required is

$$m_{3\phi} := 3\,A_{3\phi}d = \frac{d|S|}{\delta|V|} = \frac{1}{2}m_{1\phi}$$

Moreover the resistance $r_{3\phi}$ per unit length of the cable is

$$r_{3\phi} := \frac{\rho}{A_{3\phi}} = \frac{3\rho\delta|V|}{|S|}$$

and hence the active power loss in the cable is

$$l_{3\phi} := 3r_{3\phi}|I_{3\phi}|^2 d = \frac{9\rho\delta|V|}{|S|}\cdot\frac{d|S|^2}{9|V|^2} = \frac{\rho\delta d|S|}{|V|} = \frac{1}{2}l_{1\phi}$$

i.e., the balanced three-phase system uses half as much material and incurs half as much loss as the single-phase system. □

Remark 1.4. 1 Example 1.8 also shows that thermal loss $r|I|^2$ is inversely proportional to $|V|$. Intuitively a higher load voltage $|V|$ requires a smaller load current $|I|$ to deliver the same amount of power $|S|$, resulting in a smaller thermal loss in the grid.

2 It is shown in Exercise 2.7 that, given a desired load power, the active line loss is inversely proportional to the square $|V|^2$ of the load voltage magnitude, rather than $|V|$ derived here. This is because, in Exercise 2.7, the line resistance is given and independent of load power and voltage $|V|$, whereas, here, the line resistance $r_{3\phi}$

is chosen to be proportional to $|V|$ (reducing the dependence of line loss $r_{3\phi}|I_{3\phi}|^2$ from $|V|^2$ to $|V|$).

3 Note that V is the voltage drop across the load, not the voltage drop across transmission line z which is $zdI = zdS^*/V^*$. In the case of balanced three-phase system (where $z_0 = 0$ in Figure 1.25), if the load power S and voltage V are specified then the required squared voltage magnitude at the source is

$$|zdI + V|^2 = \left|zd\frac{S^*}{V^*} + V\right|^2 = |V|^2 + d|z|^2\frac{|S|^2}{|V|^2} + 2d\mathrm{Re}\,(z^*S)$$

4 In practice most three-phase systems do include a grounded neutral line to carry unbalanced current during asymmetrical conditions, e.g., due to line faults, and reduce voltage transients during line switching or lightning events. Since the unbalanced current is much smaller than the phase currents, the neutral line is typically much smaller in size and ampacity and therefore much cheaper. □

1.4 Bibliographical notes

There are many excellent textbooks on basic power system concepts, e.g., [1, 2, 3, 4]. Many materials in this chapter follow [1]. The example comparing the savings of single-phase and three-phase systems is from [4]. Circuit theory is a well established field. For general circuit analysis using KCL and KVL, see, e.g., [8, Chapter 12]. The connection with algebraic graph theory is recently surveyed in [9].

1.5 Problems

Chapter 1.1.

Exercise 1.1 (ZIP load model)**.** A common load model, called ZIP, assumes that the real and reactive power (p, q) consumed by a load depends on the voltage magnitude $|V|$ across the load:

$$p := a_2|V|^2 + a_1|V| + a_0, \qquad q := a_2'|V|^2 + a_1'|V| + a_0'$$

for some real numbers (a_0, a_1, a_2) and (a'_0, a'_1, a'_2). This can be equivalently described in terms of the complex power $s := p + \mathbf{i}q$ consumed by the load, as [1]

$$s := b_2|V|^2 + b_1|V| + b_0 \tag{1.47a}$$

where $b_i = a_i + \mathbf{i}a'_i$. Instead of the complex power s, a ZIP model may describe how the apparent power $|s|$ consumed by the load depends on $|V|$:

$$|s| := c_2|V|^2 + c_1|V| + c_0 \tag{1.47b}$$

for some real numbers (c_0, c_1, c_2). Given a ZIP load, specified either by (1.47a) or (1.47b), show that its power consumption is equivalent to the sum of power consumed by a constant impedance z, a constant current device (source) J, and a constant power device (source) σ, and express the parameters (z, J, σ) of these devices in terms of the parameters of the ZIP load.

Exercise 1.2 (KVL)**.** Prove that Kirchhoff's voltage law (1.3b) is equivalent to (1.4b). (Hint: See Appendix A.11 and use Theorem A.35.1 and Theorem A.35.2.)

Exercise 1.3 (Circuit analysis)**.** Consider a 3-node 3-link circuit specified by:

incidence matrix $\hat{C} := \begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & 0 \\ 0 & -1 & -1 \end{bmatrix}$, impedances $z_{12} = z_{23} = 1$, voltage source v_{13}

Use (**??**) to determine the currents $J_1 := (J_{12}, J_{23}, J_{13})$, voltages $U_1 := (U_{12}, U_{23})$ and nodal voltages $V := (V_1, V_2)$, assuming without loss of generality that node 3 is the reference node with $V_3 := 0$.

Exercise 1.4 (Circuit analysis)**.** For the three-bus network in Figure 1.5, derive the current balance equation (1.10a) by analyzing the equivalent circuit using KCL, KVL, and Ohm's law, as explained in Chapter 1.1.4. Draw the equivalent circuit.

Exercise 1.5 (One-line diagram and Π circuit)**.** Derive (1.10) $I = YV$ from the one-line diagram of a general network by analyzing its equivalent circuit.

[1] The power consumption may depend also on the frequency. During transient, this dependence can be made explicit by the time-domain model

$$s(t) := \Big(a_2|v(t)|^2 + a_1|v(t)| + a_0\Big)\,(1 + a_3\Delta\omega(t))$$

where $s(t) := v(t)i(t)$ is the instantaneous power in the time-domain and $\Delta\omega(t)$ is the deviation from the nominal frequency during transient.

Chapter 1.2.

Exercise 1.6 ($\alpha := e^{-\mathbf{i}2\pi/3}$)**.** Prove the following properties of $\alpha := e^{-\mathbf{i}\angle 120^\circ}$:

1 $\alpha^2 = \overline{\alpha}$, $\alpha^3 = 1$, $\alpha^4 = \alpha$, $\alpha^k = \alpha^{k \mod 3}$ where $\overline{a}$ denotes the complex conjugate of a.
2 $1+\alpha+\alpha^2 = 0$.
3 $1-\alpha = \sqrt{3}\angle 30^\circ$, $1-\alpha^2 = \sqrt{3}\angle -30^\circ$.
4 $1+\alpha = -\alpha^2 = 1\angle -60^\circ$, $1+\alpha^2 = -\alpha = 1\angle 60^\circ$.
5 $\overline{\alpha}_+ = \alpha_-$, $\overline{\alpha}_- = \alpha_+$.

Exercise 1.7 (Proof of Theorem 1.2)**.** Let $\alpha := e^{-\mathbf{i}2\pi/3}$. Recall the matrices F defined in (1.18) and Γ in (1.12), reproduced here:

$$F := \frac{1}{\sqrt{3}}\begin{bmatrix}\mathbf{1} & \alpha_+ & \alpha_-\end{bmatrix} = \frac{1}{\sqrt{3}}\begin{bmatrix}1 & 1 & 1\\ 1 & \alpha & \alpha^2\\ 1 & \alpha^2 & \alpha\end{bmatrix}, \qquad \Gamma := \begin{bmatrix}1 & -1 & 0\\ 0 & 1 & -1\\ -1 & 0 & 1\end{bmatrix}$$

1 Suppose the entries x_j of $x := (x_1, x_2, x_3) \in \mathbb{C}^3$ have the same magnitude. Then x is balanced if and only if $x_1 + x_2 + x_3 = 0$.
2 The columns of F are orthonormal. Both F and $\overline{F}$ are complex symmetric, i.e., $F^\mathsf{T} = F$ and $\overline{F}^\mathsf{T} = \overline{F}$, where $\overline{F}$ is the complex conjugate of F componentwise. Hence

$$F^{-1} = F^\mathsf{H} = \overline{F} = \frac{1}{\sqrt{3}}\begin{bmatrix}\mathbf{1} & \alpha_- & \alpha_+\end{bmatrix}$$

3 Γ is a normal matrix, $\Gamma\Gamma^\mathsf{T} = \Gamma^\mathsf{T}\Gamma$.
4 *Spectral decomposition of* Γ:
   1 The eigenvalues and eigenvectors of Γ are

$$\Gamma\mathbf{1} = 0, \qquad \Gamma\alpha_+ = (1-\alpha)\alpha_+, \qquad \Gamma\alpha_- = (1-\alpha^2)\alpha_- \tag{1.48}$$

   where $1-\alpha = \sqrt{3}e^{\mathbf{i}\pi/6}$ and $1-\alpha^2 = \sqrt{3}e^{-\mathbf{i}\pi/6}$.
   2 Therefore the spectral decomposition of Γ is:

$$\Gamma = F\begin{bmatrix}0 & & \\ & 1-\alpha & \\ & & 1-\alpha^2\end{bmatrix}\overline{F}$$

5 *Spectral decomposition of* Γ^T:
   1 The eigenvalues and eigenvectors of Γ^T are

$$\Gamma\mathbf{1} = 0, \qquad \Gamma\alpha_- = (1-\alpha)\alpha_-, \qquad \Gamma\alpha_+ = (1-\alpha^2)\alpha_+ \tag{1.49}$$

   where $1-\alpha = \sqrt{3}e^{\mathbf{i}\pi/6}$ and $1-\alpha^2 = \sqrt{3}e^{-\mathbf{i}\pi/6}$.

2 Therefore the spectral decomposition of Γ^{T} is:

$$\Gamma^{\mathsf{T}} = \overline{F}\begin{bmatrix} 0 & & \\ & 1-\alpha & \\ & & 1-\alpha^2 \end{bmatrix} F \tag{1.50}$$

Exercise 1.8. Show that the voltage magnitude $|V^{cd}| = 208V$ in the split-phase Delta transformer in Figure 1.20(b), assuming the system is a balanced three-phase positive sequence.

Exercise 1.9. Consider the balanced three-phase system in Y configuration shown in Figure 1.26. Show that $V^{n_0 n_1} = 0$ provided $z \neq -(z_1 + l_1)/3$.[2]

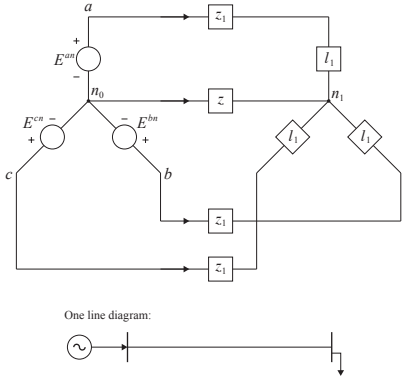


Figure 1.26 Balanced three-phase system in Y configuration where the impedances z, z_1, l_1 are given.

Exercise 1.10 (Balanced Y loads)**.** Consider the balanced three-phase system in Y configuration shown in Figure 1.27 where a three-phase voltage source in positive sequence supplies m three-phase loads in parallel. All transmission lines have a common admittance $T = 1$ and all loads have a common admittance L. Consider the following $10m$ variables:

[2] Suppose the impedances z, z_1, l_1 all have positive resistance, which is the case in practice. Then this condition is automatically satisfied. If $3z = -(z_1 + l_1)$ holds, however, then $V^{n_0 n_1}$ can take any value and Kirchhoff's laws will be satisfied because $I^{n_0 n_1} + I_a + I_b + I_c = 0$ will always be satisfied for any value of $V^{n_0 n_1}$.

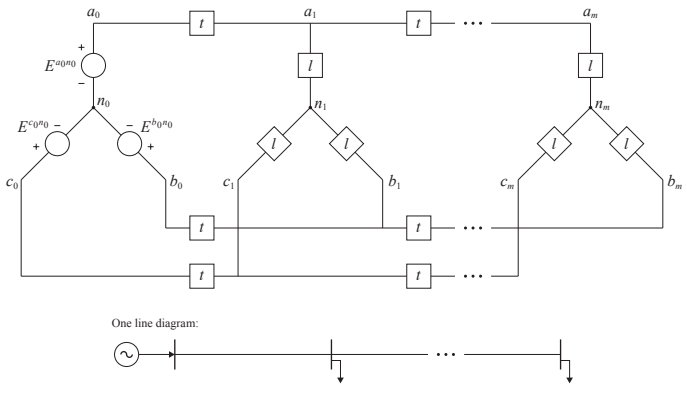


Figure 1.27 Balanced three-phase system in Y configuration where a three-phase voltage source in positive sequence supplies m three-phase loads in parallel.

- a voltage and a current for each phase at each stage $k = 1,\ldots,m$:

$$\tilde{V}_k \;:=\; \begin{bmatrix} V^{a_k n_k} \\ V^{b_k n_k} \\ V^{c_k n_k} \end{bmatrix} \text{ and } \tilde{I}_k \;:=\; \begin{bmatrix} I_{a_k n_k} \\ I_{b_k n_k} \\ I_{c_k n_k} \end{bmatrix}, \qquad k = 1,\ldots,m$$

  for a total of $6m$ variables.
- a current for each phase from stage $k-1$ to stage k:

$$\tilde{J}_{k-1,k} \;:=\; \begin{bmatrix} I_{a_{k-1}a_k} \\ I_{b_{k-1}b_k} \\ I_{c_{k-1}c_k} \end{bmatrix}, \qquad k = 1,\ldots,m$$

  for a total of $3m$ currents.
- a voltage between neutrals from stage $k-1$ to stage k: $V^{n_{k-1}n_k}$, $k = 1,\ldots,m$, for a total of m voltages.

1 Show that $V^{n_{k-1}n_k} = 0$ for $k = 1,\ldots,m$.
2 Show that

$$V^{a_k n_k} \;=\; \beta_k\, E^{a_0 n_0}, \qquad V^{b_k n_k} \;=\; \beta_k\, E^{b_0 n_0}, \qquad V^{c_k n_k} \;=\; \beta_k\, E^{c_0 n_0}, \qquad k = 1,\ldots,m$$

where β_k is:

$$\beta_k := \frac{r_1^k r_2^m (r_2 - 1) - r_2^k r_1^m (r_1 - 1)}{r_2^m (r_2 - 1) - r_1^m (r_1 - 1)}$$

and r_1, r_2 are given by:

$$r_{1,2} = \frac{1}{2}\left((L+2) \pm \sqrt{L(L+4)}\right) \tag{1.51}$$

(Hint: Derive a recursion on $\tilde{V}_k$ across stages k and solve the difference equation for each phase a, b, c separately.)

3 Show that $\tilde{V}_k, \tilde{I}_k, \tilde{J}_{k-1,k}$ are balanced positive-sequence sets for $k = 1, \ldots, m$.

Chapter 1.3.

Exercise 1.11. Show that the instantaneous power in the time domain can be expressed in terms of real and reactive powers in the phasor domain:

$$\begin{aligned} p(t) &= |V||I|\left(\cos\phi + \cos(2\omega t + \theta_V + \theta_I)\right) \\ &= P\ (1 + \cos 2(\omega t + \theta_I)) - Q\ \sin 2(\omega t + \theta_I) \end{aligned}$$

where $\phi := \theta_V - \theta_I$ is the power factor angle, $P := |V||I|\cos\phi$ is the real power and $Q := |V||I|\sin\phi$ is the reactive power.

Exercise 1.12 (Instantaneous power)**.** Consider a balanced K-phase system with $K \geq 3$ and for $k = 0, \cdots, K-1$,

$$v_k(t) = \sqrt{2}|V|\cos\left(\omega t + \left(\theta_V + k\frac{2\pi}{K}\right)\right), \quad i_k(t) = \sqrt{2}|I|\cos\left(\omega t + \left(\theta_I + k\frac{2\pi}{K}\right)\right)$$

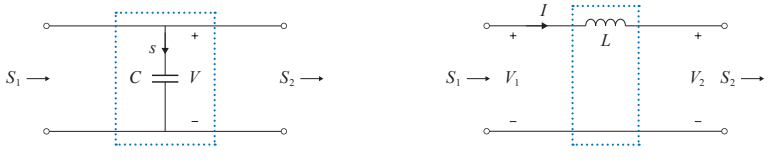
Show that $p_{K\phi}(t) := \sum_{k=0}^{K-1} v_k(t) i_k(t) = KP$ where $P := (1/T)\int_0^T v_0(t) i_0(t) dt = |V||I|\cos(\theta_V - \theta_I)$ and $T := 2\pi/\omega$.

Exercise 1.13. Suppose $z = 1/\mathbf{i}\omega c$ (capacitance). Prove directly in time domain that the average delivered power is 0 and the magnitude of the instantaneous power is Q.

Exercise 1.14 (Power meter)**.** A power meter measures voltage and current magnitudes (rms values) $(|V|, |I|)$ and instantaneous power $p(t)$ over 1 or more period T. In addition to reporting $(|V|, |I|)$, it usually reports real and reactive power (P, Q), apparent power $|S|$, and power factor as well. Explain how to calculate these quantities.

Exercise 1.15. Consider Figure 1.28.

1 *Shunt capacitor is VAR source:* Prove that in Figure 1.28(a), $S_2 = S_1 + \mathbf{i}\omega C|V|^2$.
2 *Short transmission line is inductive:* Prove that in Figure 1.28(b), if $|V_2| = |V_1|$ then $S_2 = S_1^{\mathsf{H}}$.



(a) Shunt capacitor is VAR source

(b) Short transmission line is inductive

Figure 1.28 Conservation of power

2 Transmission line models

An electric network consists of transmission lines that transfer power from generators to loads. In this chapter we develop models for the terminal behavior of a three-phase transmission line that map the voltage and current at one end of the line to those at the other end, in two steps. In Chapter 2.1 we derive inductance and capacitance parameters of a transmission line as functions of line geometry. In Chapter 2.2 we use these parameters to develop circuit models for short, medium, and long-distance transmission lines. These line models are building blocks for network models developed in later chapters.

2.1 Line characteristics

The alternating currents in the conductors of a three-phase transmission line create electromagnetic interactions among them that couple the voltages on, and currents and charges in these conductors. In a balanced operation however the interactions are as if the phases are decoupled. This allows per-phase analysis where, in each phase, the line can be characterized as a combination of a series impedance and a shunt admittance parameterized by:

$$\begin{aligned} \text{series impedance per meter} \quad & z := r + \mathbf{i}\omega l \qquad \Omega/\text{m} \\ \text{shunt admittance per meter to neutral} \quad & y := g + \mathbf{i}\omega c \qquad \Omega^{-1}/\text{m} \end{aligned}$$

In this section we present models for these per-meter line parameters (r, l) and (g, c). In the next section we will use these parameters to derive lumped-circuit models of the line. A three-phase line consists of multiple wires and therefore we need to derive the series inductance l and shunt capacitance c due to currents and charges in multiple wires. The key property that will be important in our derivation is that the set of wires carry currents in both directions so that the currents and charges in all the wires sum to zero at all times, as expressed in (2.2) and (2.5) below.

2.1.1 Series resistance r and shunt conductance g

The direct current (dc) resistance of a conductor is

$$r_{\text{dc}} := \frac{\rho_T}{A} \qquad \Omega/\text{m}$$

where ρ_T is called the conductor resistivity at temperature T and A is the cross-sectional area of the conductor. Hence the per-meter resistance is inversely proportional to the size of the line. The alternating current (ac) resistance (or effective resistance) of a conductor is defined to be

$$r_{\text{ac}} := \frac{P_{\text{loss}}}{|I|^2} \qquad \Omega/\text{m}$$

where P_{loss} is the real power loss in W and $|I|$ is the root-mean-square of the current in A in the conductor. The current distributes uniformly throughout the conductor's cross-sectional area for dc. For ac, the current density is lower at the conductor center and higher near the conductor surface. This is called the skin effect and is more pronounced at higher ac frequencies. As frequency increases, the real power loss, and hence the ac resistance, also increase. At 60 Hz the ac resistance is at most a few percent higher than dc resistance. These effects are modeled by the series resistance r in Ω/m in transmission line models.

Shunt conductance g in Ω^{-1}/m accounts for real power loss between conductors or between conductors and ground, typically due to either leakage currents at insulators or to corona. Insulator loss depends on the environment such as moisture level. Corona occurs when a strong electric field at a conductor surface ionizes the air, causing it to conduct. It depends on meteorological conditions such as rain. Losses due to insulator leakage and corona are typically negligible compared to resistance loss $r|I|^2$. It is therefore common to assume zero shunt conductance g in transmission line models.

2.1.2 Series inductance l

Roughly, the per-meter series inductance l in henrys/m of a wire is the proportionality constant between the current i in a meter of the wire and the total magnetic flux linkages λ, i.e., $\lambda(t) = li(t)$, where $i(t)$ is in ampere and λ is in webers. We now study how the per-meter series inductance l of a wire depends on the geometry of the transmission lines.

Single conductor. Consider a straight infinitely long wire of radius r with uniform current density in the wire with a total current i (dropping t from the notation for simplicity). The total flux linkages λ_R per meter of the wire within a radius R of the wire is related to the current i and the geometry by:

$$\lambda_R = \frac{\mu_0}{2\pi}\left(\frac{\mu_r}{4} + \ln\frac{R}{r}\right) i$$

where $\mu_0 := 4\pi \times 10^{-7}$ weber/ampere-meter is the permeability of free space, and μ_r is the relative permeability of the wire. If the conductor is nonmagnetic (e.g. copper or aluminum), then $\mu_r \approx 1$. The first term is due to flux linkages inside the wire and the second term is due to flux linkages outside the wire up to radius R. The details are explained in [1, pp.54–59].

Multiple conductors. We will calculate approximately the per-meter total flux linkages λ_1 of conductor 1 that carries a current i_1. The total flux linkages λ_1 is determined not only by current i_1, but also by currents i_k from other conductors $k = 2, \ldots, n$, that carry currents i_k and are at distances d_{1k} from the center of conductor 1. See Figure 2.1.

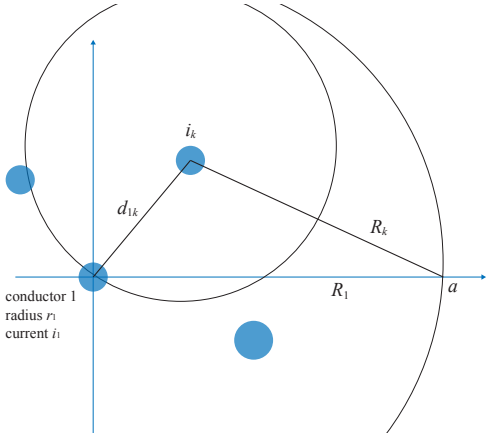


Figure 2.1 Per-meter total flux linkages in a volume within a radius R_1 from the center of conductor 1 due to all conductors. Conductors k carry currents i_k and their centers are distances d_{1k} from the center of conductor 1 and R_k from point a.

Denote by R_1 the distance of point a from the origin (center of conductor 1) and by R_k the distance of the center of conductor k from point a. Then the total flux linkages of conductor 1 is

$$\lambda_1 = \lim_{R_1 \to \infty} \frac{\mu_0}{2\pi} \left(i_1 \left(\frac{\mu_r}{4} + \ln \frac{R_1}{r_1} \right) + \sum_{k=2}^{n} i_k \ln \frac{R_k}{d_{1k}} \right) \tag{2.1}$$

where ln denotes the natural log. We make the key assumption

$$\sum_{k=1}^{n} i_k(t) = 0 \qquad \text{at all times } t \tag{2.2}$$

This is a reasonable assumption as in practice the lines carrying power from generation to load and the lines carrying the return currents follow the same physical path by design. The implication is that the magnetic inductances due to all the lines cancel each other at infinity. Formally, we add $-\ln R_1 \sum_{k=1}^{n} i_k$ into the bracket on the right-hand

side of (2.1) to get

$$\lambda_1 = \lim_{R_1\to\infty} \frac{\mu_0}{2\pi}\left(i_1\left(\frac{\mu_r}{4}+\ln\frac{1}{r_1}\right)+\sum_{k=2}^{n} i_k \ln\frac{1}{d_{1k}}\right) + \frac{\mu_0}{2\pi}\sum_{k=1}^{n} i_k \ln\frac{R_k}{R_1}$$

As $R_1 \to \infty$, $\ln(R_k/R_1) \to 0$. Hence

$$\lambda_1 = \frac{\mu_0}{2\pi}\left(i_1 \ln\frac{1}{r_1'} + \sum_{k=2}^{n} i_k \ln\frac{1}{d_{1k}}\right)$$

where $r_1' := r_1 e^{-\mu_r/4}$ is the radius of an equivalent hollow conductor with the same flux linkages as the solid conductor of radius r. For a nonmagnetic wire, $\mu_r \approx 1$ and $r_1' \approx 0.78 r_1$.

In general the total flux linkages λ_k of conductor k depends not only on current i_k but currents $i_{k'}$ in other conductors as well, and is given by

$$\lambda_k = \left(\frac{\mu_0}{2\pi}\ln\frac{1}{r_k'}\right) i_k + \sum_{k'\neq k}\left(\frac{\mu_0}{2\pi}\ln\frac{1}{d_{kk'}}\right) i_{k'} \tag{2.3}$$

where $r_k' := r_k e^{-\mu_r/4}$. In vector form this is

$$\lambda = Li$$

where $\lambda := (\lambda_k, k = 1,\ldots,n)$, $i := (i_k, i = 1,\ldots,n)$, and the (k,k')-th entry of the $n\times n$ matrix L is

$$l_{kk'} = \begin{cases} \frac{\mu_0}{2\pi}\ln\frac{1}{r_k'} & \text{if} \quad k = k' \\ \frac{\mu_0}{2\pi}\ln\frac{1}{d_{kk'}} & \text{if} \quad k \neq k' \end{cases}$$

The voltage drop $v_k(t)$ between two points on conductor k that are separated by an infinitesimal distance is related to the rate of change of the total flux linkages $\lambda_k(t)$ (Faraday's law), i.e.,

$$v_k(t) = \frac{d}{dt}\lambda_k(t) = \sum_{k'} l_{kk'}\frac{d}{dt} i_{k'}(t)$$

This relation, in the phasor domain, is used in Chapter 2.2.1 to derive a circuit model of a transmission line. In a circuit model, the term

$$l_{kk} := \frac{\mu_0}{2\pi}\ln\frac{1}{r_k'} \qquad \text{henrys/m}$$

is called the *self-inductance* per meter of conductor k and the term

$$l_{kk'} := \frac{\mu_0}{2\pi}\ln\frac{1}{d_{kk'}} \qquad \text{henrys/m}$$

is called the *mutual inductances* per meter between conductors k and k'. The larger the conductor r_k the smaller the self-inductance l_k.

2.1.3 Shunt capacitance c

Roughly, the per-meter shunt capacitance c, in farads/m, of a wire is the proportionality constant between the charge q, in coulombs/m, in a meter of the wire and the voltage v on the surface of the wire, i.e., $q(t) = cv(t)$. We now study how the per-meter shunt capacitance c of a wire depends on the geometry of the transmission lines.

Consider the situation in Figure 2.1 with multiple conductors. A similar analysis to that in Chapter 2.1.2 shows that the voltage, with respect to a reference at infinity, at a point on the surface of conductor k is

$$v_k = \left(\frac{1}{2\pi\epsilon}\ln\frac{1}{r_k}\right)q_k \;+\; \sum_{k'\neq k}\left(\frac{1}{2\pi\epsilon}\ln\frac{1}{d_{kk'}}\right)q_{k'} \tag{2.4}$$

where ϵ is the permittivity of the medium ($\epsilon = 8.854\times10^{-12}$ farads/meter in free space and $\epsilon \approx 1$ farad/meter in dry air). As before, r_k is the radius of conductor k and $d_{kk'}$ is the distance between the centers of conductors k and k'. Here q_k is the total charge per unit length of wire k in coulombs/m. In vector form this is

$$v = Fq$$

where $v := (v_k, k = 1,\ldots,n)$, $q := (q_k, k = 1,\ldots,n)$, and the (k,k')-th entry of the $n\times n$ matrix F is

$$f_{kk'} = \begin{cases} \frac{1}{2\pi\epsilon}\ln\frac{1}{r_k} & \text{if} \quad k = k' \\ \frac{1}{2\pi\epsilon}\ln\frac{1}{d_{kk'}} & \text{if} \quad k \neq k' \end{cases}$$

Taking time derivatives relates the currents in the conductors to the rate of change in a voltage on the surface of the conductor relative to the reference, $\dot{v} = Fi(t)$. Let $C := F^{-1}$. The diagonal entries c_{kk} of C are called self-capacitances per meter of conductor k and the off-diagonal entries $c_{kk'}$ of C are called mutual capacitances per meter between conductors k and k', in farads/m. The larger the conductor r_k the larger the self-capacitance c_{kk}.

The key assumption (among others) in deriving (2.4) is

$$\sum_{k=1}^{n} q_k(t) = 0 \qquad \text{at all times } t \tag{2.5}$$

Compare this assumption with the assumption (2.2), and the expressions (2.3) and (2.4).

Example 2.1. The voltage v_k in (2.4) is the potential, or voltage with respect to the reference at infinity, at a point on the surface of conductor k. The voltage difference v_{jk} between two points on the surfaces of two parallel conductors j and k that are on a plane perpendicular to conductor j is:

$$v_{jk} \;:=\; v_j - v_k = \frac{1}{2\pi\epsilon}\left(q_j\ln\frac{d_{kj}}{r_j} \;-\; q_k\ln\frac{d_{jk}}{r_k} \;+\; \sum_{k'\neq j,k} q_{k'}\ln\frac{d_{kk'}}{d_{jk'}}\right)$$

2.1.4 Balanced three-phase line

Consider the simplest model of a symmetric three-phase transmission line in balanced operation, as shown in Figure 2.2, with the assumptions:

1 the conductors are equally spaced at D and have equal radii r;[1]
2 $i_a(t)+i_b(t)+i_c(t)=0$ at all times t;
3 $q_a(t)+q_b(t)+q_c(t)=0$ at all times t.

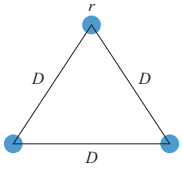


Figure 2.2 Per-meter inductance and capacitance of a symmetric three-phase transmission line in balanced operation.

It can be shown (see Exercise 2.1) that in this symmetric arrangement the effect of mutual inductances and capacitances among the transmission lines is particularly simple, resulting in the following equal per-phase inductance for each line:

$$l = \frac{\mu_0}{2\pi} \ln \frac{D}{r'} \qquad \text{H/m}$$

where $r' := re^{-\mu_r/4}$, and equal per-phase capacitance for each line:

$$c = \frac{2\pi\epsilon}{\ln(D/r)} \qquad \text{F/m}$$

Note that l and c include not only the self-inductance and self-capacitance of the line, but also mutual inductances and capacitances. Two implications are as follows.

1 Although there is magnetic coupling between phases, the conditions $i_a(t)+i_b(t)+i_c(t)=0$, $q_a(t)+q_b(t)+q_c(t)=0$ and the symmetry (equal radii r and distances D) reduce the effect of the magnetic coupling to the term $\ln D$. This allows us to model the magnetic effect *as if* it consists of only self-inductance and electric effect *as if* it consists of only self-capacitance. Moreover, the inductances and capacitances are equal for each phase, permitting per-phase analysis.
2 To reduce the impedance per meter due to inductance or capacitance, we can reduce the spacing D or increase the wire radius r. Both have limitations. Other techniques are used in practice to approximate condition 1 above on the symmetry of line geometry, e.g., conductor bundling and transposition of the transmission lines.

[1] We use r to denote both the per-meter series resistance and the radius of the conductor; the meaning should be clear from the context.

Consider any point p that is equidistant from the centers of the conductors a, b, c, e.g., the point at the center of the triangle in Figure 2.2. The potential, or the voltage relative to the reference point at infinity, at this point p can be shown to be

$$v_p = \frac{1}{2\pi\epsilon}\left(q_a \ln\frac{1}{d_{pa}} + q_b \ln\frac{1}{d_{pb}} + q_c \ln\frac{1}{d_{pc}}\right) \tag{2.6}$$

where $d_{pa} = d_{pb} = d_{pc}$ are the distances between p and the centers of the conductors. Since $q_a + q_b + q_c = 0$ we have $v_p = 0$, and hence p has the same potential as the reference point at infinity and can therefore be taken as the reference point. We will construct an imaginary geometric line parallel to the conductors pass through the equidistance point from these conductors. Every point on this line is the reference potential. By default we will pick this as the neutral potential that defines the phase-to-neutral voltages. The current supplied to the transmission line capacitance is called the *charging current* and the corresponding capacitance is also called the *line charging*. Figure 2.3 shows the corresponding circuit model of a transmission line. When the

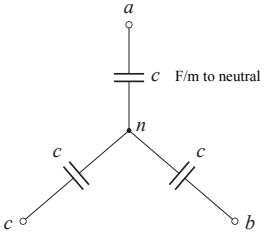


Figure 2.3 Circuit model of the cross section of a balanced three-phase transmission line.

phase a line-to-neutral voltage is V_{an} the phase a charging current is

$$I_{a,\text{charging}} = \mathbf{i}\omega c V_{an} \qquad \text{A/m}$$

from phase a conductor to neutral.

2.2 Line models

Consider a three-phase transmission line in balanced operation in sinusoidal steady state, modeled as in Figure 2.3. A key conclusion of Chapter 2.1.4 is that for balanced three-phase lines, we can analyze each phase separately. Consider now a transmission line on one of the phases. Let

$$\text{series impedance per meter } z := r + \mathbf{i}\omega l \qquad \Omega/\text{m}$$
$$\text{shunt admittance per meter to neutral } y := g + \mathbf{i}\omega c \qquad \Omega^{-1}/\text{m}$$

where the per-meter resistance $r > 0$ and conductance $g > 0$ depend on the material and size of the line, and the per-meter inductance $l > 0$ and parameter $c > 0$ of the line can be

calculated as in Chapters 2.1.2–2.1.4. In this section we derive two equivalent models of a balanced three-phase transmission line. The first model represents the terminal behavior, i.e., the mapping of the voltage and current between one end of the line and those at the other end, by a transmission matrix in (2.9) below. The second model represents the terminal behavior of the line by a linear circuit with series impedance and shunt admittances given in (2.14) below.

2.2.1 Transmission matrix

Distributed-element model. We start by deriving the V-I relations between two ends of a transmission line. Figure 2.4 shows a per-phase model of a balanced three-phase line of length ℓ. The voltages are phase (line-to-neutral) voltages as illustrated in Figure 2.3. We will call the left end the sending end and the right end the receiving end. When we apply a voltage V_1, with respect to neutral, at the sending end driving a current I_1 towards the receiving end, the voltage drops and the current leaks from the sending end to the receiving end so that the voltage $V(x)$ and current $I(x)$ at each point x of the line vary. We will derive a relation between the sending end (V_1, I_1) and the receiving end (V_2, I_2) by solving for $(V(x), I(x))$ in terms of (V_2, I_2) for all $0 \le x \le \ell$.

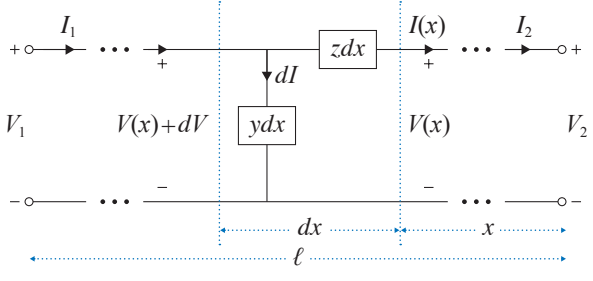


Figure 2.4 Per-phase model of a balanced three-phase line of length ℓ with impedance parameters z, y.

To this end consider the infinitesimal segment of length dx at a distance x from the receiving end. This segment is modeled by the circuit with series impedance zdx and shunt admittance ydx to neutral as shown in Figure 2.4. Let the voltage and current at point x be $V := V(x)$ and $I := I(x)$ respectively. Let the corresponding quantities at point $x + dx$ be $V(x) + dV$ and $I(x) + dI$. Applying Kirchhoff's laws to the segment, we have

$$\begin{aligned} dV &= zI(x)\, dx \\ dI &= (V(x) + dV)y\, dx \;\approx\; yV(x)\, dx \end{aligned}$$

where the approximation results from ignoring the second-order term $dV dx$. Hence

we have

$$\begin{bmatrix} \frac{dV}{dx} \\ \frac{dI}{dx} \end{bmatrix} = \begin{bmatrix} 0 & z \\ y & 0 \end{bmatrix} \begin{bmatrix} V \\ I \end{bmatrix} \tag{2.7}$$

Transmission matrix.

The ordinary differential equation (2.7) can be easily solved using standard method (see below for details), and the general solution is:

$$\begin{bmatrix} V(x) \\ I(x) \end{bmatrix} = U \begin{bmatrix} e^{\gamma x} & 0 \\ 0 & e^{-\gamma x} \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \end{bmatrix} \tag{2.8a}$$

for some constants k_1, k_2, where

$$U := \begin{bmatrix} Z_c & -Z_c \\ 1 & 1 \end{bmatrix} \text{ and } U^{-1} := \frac{1}{2Z_c} \begin{bmatrix} 1 & Z_c \\ -1 & Z_c \end{bmatrix} \tag{2.8b}$$

Here

$$Z_c := \sqrt{\frac{z}{y}} \quad \Omega m^{-1} \qquad \text{and} \qquad \gamma := \sqrt{zy} \quad m^{-1} \tag{2.8c}$$

are called the *characteristic impedance* and *propagation constant* of the line respectively. At $x = 0$, $V(0) = V_2$ and $I(0) = I_2$. From (2.8) we have

$$\begin{bmatrix} V_2 \\ I_2 \end{bmatrix} = U \begin{bmatrix} k_1 \\ k_2 \end{bmatrix}$$

and hence

$$\begin{bmatrix} V(x) \\ I(x) \end{bmatrix} = U \begin{bmatrix} e^{\gamma x} & 0 \\ 0 & e^{-\gamma x} \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \end{bmatrix} = U \begin{bmatrix} e^{\gamma x} & 0 \\ 0 & e^{-\gamma x} \end{bmatrix} U^{-1} \begin{bmatrix} V_2 \\ I_2 \end{bmatrix}$$

The sending-end voltage and current are therefore related to the receiving-end (V_2, I_2) as

$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = U \begin{bmatrix} e^{\gamma \ell} & 0 \\ 0 & e^{-\gamma \ell} \end{bmatrix} U^{-1} \begin{bmatrix} V_2 \\ I_2 \end{bmatrix}$$

Expanding, we have

$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = \begin{bmatrix} \cosh(\gamma \ell) & Z_c \sinh(\gamma \ell) \\ Z_c^{-1} \sinh(\gamma \ell) & \cosh(\gamma \ell) \end{bmatrix} \begin{bmatrix} V_2 \\ I_2 \end{bmatrix} \tag{2.9}$$

where $\cosh x := (e^x + e^{-x})/2$ and $\sinh x := (e^x - e^{-x})/2$. This defines a linear mapping that maps the voltage and current (V_2, I_2) at the receiving end to the voltage and current (V_1, I_1) at the sending end. The matrix in (2.9) is called a *transmission matrix*.

The ratio V_1/I_1 at the sending end is called the *driving-point impedance*. It is the equivalent impedance across the two sending-end terminals.

Example 2.2 (Driving-point impedance)**.** Consider the terminal model (2.9) of a transmission line. Suppose the receiving end is connected to an impedance load Z_l. Show that the driving-point impedance V_1/I_1 is equal to the characteristic impedance Z_c of the line under one of the following conditions:

- if the load is matched to the line, i.e., $Z_l = Z_c$; or
- if the line length ℓ grows to infinity, since the line parameters satisfy $r, x, g, c > 0$.

The second condition implies that as the line grows in length its impedance comes to dominate the load impedance Z_l.

Solution. Since $V_2 = Z_l I_2$, we have from (2.9) that when $Z_l = Z_c$

$$\frac{V_1}{I_1} = Z_c \frac{\cosh(\gamma\ell) + \sinh(\gamma\ell)}{\sinh(\gamma\ell) + \cosh(\gamma\ell)} = Z_c$$

For the second case, we have from (2.9)

$$\frac{V_1}{I_1} = Z_c \frac{Z_l \cosh(\gamma\ell) + Z_c \sinh(\gamma\ell)}{Z_l \sinh(\gamma\ell) + Z_c \cosh(\gamma\ell)} = Z_c \frac{Z_l + Z_c \tanh(\gamma\ell)}{Z_l \tanh(\gamma\ell) + Z_c}$$

Now $\gamma = \sqrt{zy} =: \sqrt{\hat{\gamma}}$ where $\hat{\gamma} := (rg - \omega^2 lc) + \mathbf{i}\omega(rc + gl)$. Note that $\mathrm{Im}\hat{\gamma} > 0$ and hence $\angle\hat{\gamma} \in (0, \pi)$ and $\gamma \in (0, \pi/2)$. If we write $\gamma =: \alpha + \mathbf{i}\beta$ then $\alpha > 0$. Hence

$$\cosh(\gamma\ell) = \frac{1}{2}\left(e^{\gamma\ell} + e^{-\gamma\ell}\right) = \frac{1}{2}\left(e^{(\alpha+\mathbf{i}\beta)\ell} + e^{-(\alpha+\mathbf{i}\beta)\ell}\right)$$
$$\sinh(\gamma\ell) = \frac{1}{2}\left(e^{\gamma\ell} - e^{-\gamma\ell}\right) = \frac{1}{2}\left(e^{(\alpha+\mathbf{i}\beta)\ell} - e^{-(\alpha+\mathbf{i}\beta)\ell}\right)$$

and

$$\tanh(\gamma\ell) = \frac{e^{(\alpha+\mathbf{i}\beta)\ell} - e^{-(\alpha+\mathbf{i}\beta)\ell}}{e^{(\alpha+\mathbf{i}\beta)\ell} + e^{-(\alpha+\mathbf{i}\beta)\ell}} = \frac{1 - e^{-2(\alpha+\mathbf{i}\beta)\ell}}{1 + e^{-2(\alpha+\mathbf{i}\beta)\ell}} \to 1 \quad \text{as} \quad \ell \to \infty$$

Hence $V_1/I_1 \to Z_c$ as $\ell \to \infty$. □

Example 2.3 (Matched load)**.** Suppose the line is terminated in its characteristic impedance Z_c, i.e., $V_2 = Z_c I_2$. Then (2.9) yields

$$V_1 = (\cosh(\gamma\ell) + \sinh(\gamma\ell))\, V_2 = V_2 e^{\gamma\ell}$$
$$I_1 = (\cosh(\gamma\ell) + \sinh(\gamma\ell))\, I_2 = I_2 e^{\gamma\ell}$$

Therefore the driving-point impedance V_1/I_1 is also the characteristic impedance Z_c of the line. Moreover the ratio of the receiving to sending end voltages and currents are

$$\frac{V_2}{V_1} = \frac{I_2}{I_1} = e^{-\gamma\ell}$$

The ratio of the receiving power to the sending power is:

$$\frac{-S_{21}}{S_{12}} = \frac{V_2 I_2^*}{V_1 I_1^*} = e^{-\gamma\ell}\left(e^{-\gamma\ell}\right)^*$$

Writing $\gamma = \sqrt{zy} = \sqrt{(rg - \omega^2 lc) + \mathbf{i}\omega(rc + gl)} =: \alpha + \mathbf{i}\beta$, we have

$$\frac{-S_{21}}{S_{12}} = e^{-2\alpha\ell}$$

Since $e^{-2\alpha\ell}$ is real, the powers have the same phase angle $\angle(-S_{21}) = \angle S_{12} =: \theta$. This implies that the transmission efficiency has the same ratio in terms of real power $-P_{21}$ received and real power P_{12} sent:

$$\frac{-P_{21}}{P_{12}} = \frac{-S_{21}\cos\theta}{S_{12}\cos\theta} = e^{-2\alpha\ell}$$

Hence for an impedance load that is matched to the line impedance Z_c, the transmission efficiency η decreases exponential in the line length ℓ. For high-voltage transmission lines, $\alpha \approx 0$ so the loss is small and $\eta \approx 1$.

Indeed, for a lossless line, $r = g = 0$. Then $z = \mathbf{i}\omega l$ and $y = \mathbf{i}\omega c$. Hence

$$Z_c = \sqrt{\frac{z}{y}} = \sqrt{\frac{l\ell}{c\ell}} = \sqrt{\frac{L}{C}}$$

is real, where L is the total inductance of the line and C the total capacitance of the line, and

$$\gamma = \sqrt{zy} = \mathbf{i}\omega\sqrt{lc}$$

is purely imaginary ($\alpha = 0$). The transmission efficiency is $\eta = -P_{21}/P_{12} = 1$. We will study lossless lines in more detail in Chapter 2.2.4. □

Solution of (2.7).

First we note that even though (V, I) and the parameters (y, z) are complex variables, the variable x (distance from terminal 2) is a real variable. Hence the ordinary differential equation (ode) (2.7) can be solved in the same way as an ode in the real domain. To see this consider a general ode:

$$\dot{z} := \frac{dz}{dt} = Mz \tag{2.10}$$

where $z := x + jy \in \mathbb{C}^n$ with x, y in $\mathbb{R}^n$ and $M := A + jB \in \mathbb{C}^{n\times n}$ with A, B in $\mathbb{R}^{n\times n}$, with the interpretation $\dot{x} + j\dot{y} = (A + jB)(x + jy)$. Rewrite this in the real domain:

$$\begin{bmatrix}\dot{x}\\ \dot{y}\end{bmatrix} = \underbrace{\begin{bmatrix}A & -B\\ B & A\end{bmatrix}}_{\tilde{M}}\begin{bmatrix}x\\ y\end{bmatrix} \tag{2.11}$$

Two matrices

$$M = A + jB \quad \text{and} \quad \tilde{M} = \begin{bmatrix}A & -B\\ B & A\end{bmatrix}$$

are *equivalent*, written $M \leftrightarrow \tilde{M}$, in the sense that for any $z = x + \mathbf{i}y$ with $x, y \in \mathbb{R}^n$,

$$\begin{bmatrix} \mathrm{Re}(Mz) \\ \mathrm{Im}(Mz) \end{bmatrix} = \tilde{M} \begin{bmatrix} x \\ y \end{bmatrix}$$

Since

$$M^2 = \left(A^2 - B^2\right) + j\,(AB + BA) \quad \text{and} \quad \tilde{M}^2 = \begin{bmatrix} A^2 - B^2 & -(AB + BA) \\ AB + BA & A^2 - B^2 \end{bmatrix}$$

we have $\tilde{M}^2 \leftrightarrow M^2$, and by induction $\tilde{M}^k \leftrightarrow M^k$ for all k. Hence $e^{\tilde{M}} \leftrightarrow e^M$. This implies that a trajectory $z(t) \in \mathbb{C}^n$ is a solution of (2.10) if and only if $(x(t), y(t)) \in \mathbb{R}^{2n}$ with $z(t) =: x(t) + \mathbf{i}y(t)$ is a solution of (2.11). Hence solving (2.11) using $\tilde{M}$ in the real domain is equivalent to solving (2.10) using M directly in the complex domain.

We now solve the ode (2.7). Let

$$A := \begin{bmatrix} 0 & z \\ y & 0 \end{bmatrix}$$

Then the eigenvalues of A are $\pm\gamma$ where $\gamma := \sqrt{yz}$ is the propagation constant defined in (2.8c). Recall the characteristic impedance of the line $Z_c := \sqrt{\frac{z}{y}}$ also defined in (2.8c). The corresponding eigenvectors are (any vectors proportional to) the columns of the matrix U defined in (2.8b). Let U^{-1} be its inverse. Since $AU = U\mathrm{diag}(\gamma, -\gamma)$, if we define

$$\begin{bmatrix} \tilde{V}(x) \\ \tilde{I}(x) \end{bmatrix} := U^{-1} \begin{bmatrix} V(x) \\ I(x) \end{bmatrix} \tag{2.12}$$

then

$$\frac{d}{dx}\begin{bmatrix} \tilde{V} \\ \tilde{I} \end{bmatrix} = U^{-1}\frac{d}{dx}\begin{bmatrix} V \\ I \end{bmatrix} = U^{-1}A\begin{bmatrix} V(x) \\ I(x) \end{bmatrix} = U^{-1}AU\left(U^{-1}\begin{bmatrix} V(x) \\ I(x) \end{bmatrix}\right) = \mathrm{diag}(\gamma, -\gamma)\begin{bmatrix} \tilde{V}(x) \\ \tilde{I}(x) \end{bmatrix}$$

i.e., $\tilde{V}$ and $\tilde{I}$ are decoupled. Hence

$$\tilde{V}(x) = k_1 e^{\gamma x} \quad \text{and} \quad \tilde{I}(x) = k_2 e^{-\gamma x}$$

for some constants k_1, k_2. Then (2.12) implies that the general solution of (2.7) is (2.8). □

2.2.2 Lumped-element Π-circuit model

If we are only interested in the terminal voltages and currents of a line, then we can represent the line by a lumped-circuit model as shown in Figure 2.5 that consists of a series impedance Z' and a shunt admittance $Y'/2$ at each end of the line. This is called the Π *model* or Π-*circuit model* of a transmission line. We now derive the parameters (Z', Y') in the Π model in terms of line characteristics (Z_c, γ).

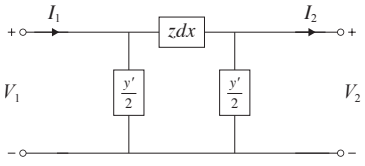


Figure 2.5 Lumped-circuit Π model of a transmission line.

Applying Kirchhoff's laws we have

$$I_1 = \frac{Y'}{2}V_1 + \frac{Y'}{2}V_2 + I_2$$
$$V_1 - V_2 = Z'\left(\frac{Y'}{2}V_2 + I_2\right)$$

Hence

$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = \begin{bmatrix} 1 + Z'Y'/2 & Z' \\ Y'(1 + Z'Y'/4) & 1 + Z'Y'/2 \end{bmatrix} \begin{bmatrix} V_2 \\ I_2 \end{bmatrix} \tag{2.13}$$

Comparing (2.13) and (2.9) we find that the Π model in Figure 2.5 is given by:

$$Z' = Z_c \sinh(\gamma\ell) = \sqrt{\frac{z}{y}} \sinh(\gamma\ell) = Z \frac{\sinh(\gamma\ell)}{\gamma\ell} \tag{2.14a}$$

$$\frac{Y'}{2} = \frac{1}{Z_c} \frac{\cosh(\gamma\ell) - 1}{\sinh(\gamma\ell)} = \frac{1}{Z_c} \frac{\sinh(\gamma\ell/2)}{\cosh(\gamma\ell/2)} = \frac{Y}{2} \frac{\tanh(\gamma\ell/2)}{\gamma\ell/2} \tag{2.14b}$$

where $Z := z\ell$ is the total series impedance of the line and $Y := y\ell$ is the total shunt admittance to neutral of the line.

When $|\gamma\ell| \ll 1$ then $\sinh(\gamma\ell)/(\gamma\ell) \approx 1$ and $\tanh(\gamma\ell/2)/(\gamma\ell/2) \approx 1$, in which case the Π model in Figure 2.5 can be approximated by the total series impedance Z and total shunt admittance Y to neutral of the line.

In summary each phase of a balanced three-phase transmission line can be modeled as follows:

- *Long line ($\ell > 150$ miles approximately):* Use either (2.9) or the Π circuit model with Z' and Y' given by (2.14).
- *Medium line ($50 < \ell < 150$ miles approximately):* Use the Π circuit model with $Z := z\ell$ and $Y := y\ell$ instead of Z' and Y'. Here $Z = R + \mathbf{i}\omega L$ is the total series impedance of the line and $Y = \mathbf{i}\omega C$ is the total shunt admittance to neutral of the line. In particular, for medium lines, the shunt resistance is negligible.
- *short line ($\ell < 50$ miles approximately):* Use the Π circuit model with Z only and neglect Y.

2.2.3 Real and reactive line losses

The power injected at terminal 1 towards terminal 2 and that at terminals 2 towards 1 are (from Kirchhoff's laws):

$$\begin{aligned} S_{12} := V_1 I_1^H &= \left(\frac{1}{Z'}\right)^H \left(|V_1|^2 - V_1 V_2^H\right) + \left(\frac{Y'}{2}\right)^H |V_1|^2 \\ S_{21} := V_2(-I_2)^H &= \left(\frac{1}{Z'}\right)^H \left(|V_2|^2 - V_2 V_1^H\right) + \left(\frac{Y'}{2}\right)^H |V_2|^2 \end{aligned}$$

They are not negatives of each other because of power loss along the line. Indeed the total complex power loss is their sum:

$$S_{12} + S_{21} = \left(\frac{1}{Z'}\right)^H |V_1 - V_2|^2 + \left(\frac{Y'}{2}\right)^H \left(|V_1|^2 + |V_2|^2\right) = Z' |I_{12}^s|^2 + \left(\frac{Y'}{2}\right)^H \left(|V_1|^2 + |V_2|^2\right)$$

where I_{12}^s denotes the current through the series impedance Z'. The first term on the right-hand side is loss due to series impedance and the last term are losses due to shunt admittances of the line. Suppose $Z' = R^s + \mathbf{i}X^s$ and the shunt admittance is purely capacitive, i.e., $Y' = \mathbf{i}B^m$ with $R^s, X^s, B^m > 0$. Then, over the transmission line,

$$\begin{aligned} \text{real power loss} \quad & \mathrm{Re}\,(S_{12} + S_{21}) = R^s |I_{12}^s|^2 \\ \text{reactive power loss} \quad & \mathrm{Im}\,(S_{12} + S_{21}) = X^s |I_{12}^s|^2 - \frac{B^m}{2}\left(|V_1|^2 + |V_2|^2\right) \end{aligned}$$

Remark 2.1 (High voltage reduces line loss)**.** Consider a load supplied by a source through a transmission line modeled by a series impedance $R + \mathbf{i}X$ and zero shunt admittances. Suppose the load draws an active power P_{load} with power factor $\cos\phi$ at a specified voltage magnitude $|V_{\text{load}}|$. It can be shown that, given a desired active load power P_{load}, the active line loss P_{line} is inversely proportional to the square of the load voltage magnitude $|V_2|$ and its power factor $\cos\phi$ (Exercise 2.7):

$$P_{\text{line}} = R|I_{\text{load}}|^2 = R\,\frac{P_{\text{load}}^2}{|V_2|^2 \cos^2\phi}$$

Therefore a higher voltage (magnitude) reduces line loss.

Note that the higher voltage refers to the voltage $|V_2|$ across the load (and eventually the source voltage $|V_1|$), not the voltage across the transmission line which is $|V_1 - V_2|$; see Figure 2.5. It is derived in Example 1.8 that, given a desired load power, the active line loss is inversely proportional to the load voltage magnitude $|V_2|$, rather than $|V_2|^2$. This is because, in Exercise 2.7, the line resistance R is given and independent of load power and voltage $|V_2|$, whereas, in Example 1.8, the line resistance R is chosen to be proportional to $|V_2|$ (reducing the dependence of line loss $R|I_{\text{load}}|^2$ from $|V_2|^2$ to $|V_2|$). □

2.2.4 Lossless line

In this subsection we look at some properties of a lossless line, i.e., when $r = g = 0$. A lossless line is an important model because a high-voltage transmission line typically has very small power loss compared with the power flow on the line, and can be modeled as a lossless line. As noted above we have

$$\begin{aligned} Z_c &= \sqrt{\frac{z}{y}} = \sqrt{\frac{\mathbf{i}\omega l}{\mathbf{i}\omega c}} = \sqrt{\frac{l}{c}} \quad \Omega \\ \gamma &= \sqrt{zy} = \sqrt{(\mathbf{i}\omega l)(\mathbf{i}\omega c)} = \mathbf{i}\omega\sqrt{lc} =: \mathbf{i}\beta \quad m^{-1} \end{aligned}$$

with $\beta := \omega\sqrt{lc}$. Therefore the characteristic impedance Z_c is purely resistive while the propagation constant γ is purely reactive. The characteristic impedance Z_c is called a *surge impedance* for a lossless line. This implies

$$\cosh(\gamma x) = \cos(\beta x) \text{ and } \sinh(\gamma x) = \mathbf{i}\sin(\beta x)$$

Π-circuit model.

Substituting Z_c and γ into (2.9) the transmission matrix reduces to

$$\begin{bmatrix} V(x) \\ I(x) \end{bmatrix} = \begin{bmatrix} \cosh(\gamma x) & Z_c \sinh(\gamma x) \\ Z_c^{-1}\sinh(\gamma x) & \cosh(\gamma x) \end{bmatrix} \begin{bmatrix} V_2 \\ I_2 \end{bmatrix} = \begin{bmatrix} \cos(\beta x) & \mathbf{i}Z_c \sin(\beta x) \\ \mathbf{i}Z_c^{-1}\sin(\beta x) & \cos(\beta x) \end{bmatrix} \begin{bmatrix} V_2 \\ I_2 \end{bmatrix} \tag{2.15}$$

for $x \in [0, \ell]$. The circuit elements Z' and Y' in the Π circuit model of a transmission line reduces to (from (2.14)):

$$Z' = Z_c \sinh(\gamma\ell) = \mathbf{i}Z_c \sin(\beta\ell) =: \mathbf{i}X \quad \Omega \tag{2.16a}$$

$$\frac{Y'}{2} = \frac{Y}{2}\frac{\tanh(\gamma\ell/2)}{\gamma\ell/2} = \frac{Y}{2}\frac{\tan(\beta\ell/2)}{\beta\ell/2} =: \mathbf{i}\frac{\omega C'}{2} \quad \Omega^{-1} \tag{2.16b}$$

where $Y := \mathbf{i}\omega c\ell$ and $C' := c\ell\,(\tan(\beta\ell/2)/(\beta\ell/2))$. If ℓ is small then $C' \approx c\ell$. When $\beta\ell < \pi$ radian, both $Z' > 0$ and $Y' > 0$, i.e., the series impedance is purely inductive and the shunt admittances are purely capacitive. In practice, for overhead lines, $1/\sqrt{lc} \approx 3 \times 10^8$ ms^{-1}. At 60 Hz (using $\beta := \omega\sqrt{lc}$)

$$\frac{\pi}{\beta} = \frac{\pi}{2\pi(60)\sqrt{lc}} \approx 2{,}500 \text{ km}$$

Hence a lossless overhead transmission line less than 2,500 km can be modeled by the simple circuit in Figure 2.6 where X and C' are given in (2.16). It is a model for either

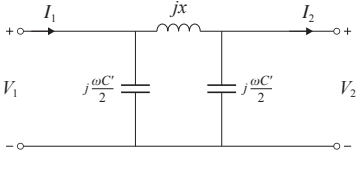


Figure 2.6 Π circuit model for a lossless line with length $\ell < \pi/\beta$.

a single-phase line or the phase-to-neutral of a balanced three-phase line.

Voltage profile.

Usually power must be delivered to a load at a specified nominal voltage magnitude $|V_2|$ at the load. To see how the voltage magnitude changes along a line from the source $x = \ell$ to the load $x = 0$, we determine the voltage $V(x)$ for $x \in [0, \ell]$ using (2.15):

$$V(x) = V_2 \cos(\beta x) + \mathbf{i} Z_c I_2 \sin(\beta x) \tag{2.17}$$

Suppose the line terminates at an impedance load $Z_{\text{load}} := R_{\text{load}} + \mathbf{i} X_{\text{load}}$. Then the voltage $V(x)$ at each point x depends on the load impedance because $V_2 = Z_{\text{load}} I_2$. There are four cases of load impedance:

1 *No load* $I_2 = 0$: $V(x) = V_2 \cos(\beta x)$ is real. Hence the voltage magnitude $V(x)$ increases from the source at $x = \ell$ to the end of the line at $x = 0$ as long as $\beta\ell < \pi/2$ radian.
2 *Surge impedance load* $Z_{\text{load}} = Z_c$: The voltage magnitude $|V(x)|$ is constant. Moreover the power delivered $S(x)$ at every point $x \in [0, \ell]$ is real and constant $|V_2|^2/Z_c$, so only real power is delivered. See Exercise 2.4.
3 *Full load*: Since $I_2 = V_2/Z_{\text{load}}$ we have

$$\begin{aligned} V(x) &= \left(\cos(\beta x) + \mathbf{i} \frac{Z_c}{Z_{\text{load}}} \sin(\beta x)\right) V_2 \\ &= \left(\cos(\beta x) + \frac{Z_c X_{\text{load}}}{|Z_{\text{load}}|^2} \sin(\beta x) + \mathbf{i} \frac{Z_c R_{\text{load}}}{|Z_{\text{load}}|^2} \sin(\beta x)\right) V_2 \end{aligned} \tag{2.18}$$

   In Exercise 2.5 we derive for special cases sufficient conditions under which the voltage magnitude $|V(x)|$ decreases from the source at $x = \ell$ to the load Z_{load} at $x = 0$.
4 *Short circuit* $V_2 = 0$: $V(x) = \mathbf{i} Z_c I_2 \sin(\beta x)$. Hence the voltage magnitude $|V(x)|$ decreases from the source at $x = \ell$ to the load at $x = 0$ as long as $\beta\ell < \pi/2$ radian.

This is illustrated in Figure 2.7. The general trend of decreasing voltage magnitude

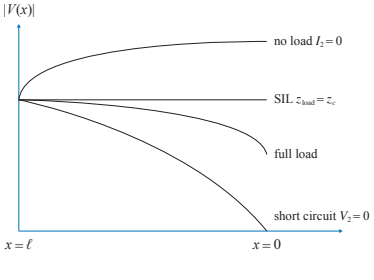


Figure 2.7 Voltage magnitude $|V(x)|$ on a lossless line.

towards the load (case 3 above) can be problematic because loads are generally designed to work with specific voltages. As mentioned above low load voltage also increases line loss in the network. Voltages are regulated tightly around their nominal values through various voltage compensation devices in generating units and inside the network.

Example 2.4 (Steady-state stability limit)**.** To derive the power delivered to a generic load we have from (2.16) that

$$I_2 = \frac{V_1 - V_2}{\mathbf{i}X} - \mathbf{i}\frac{\omega C'}{2} V_2$$

Hence the complex power delivered is

$$-S_{21} = V_2(I_2^*) = -\left(\frac{|V_2|^2 - V_2 V_1^*}{-\mathbf{i}X} - \mathbf{i}\frac{\omega C'}{2}|V_2|^2\right)$$

and the real power delivered is

$$-P_2 = \frac{|V_1||V_2|}{X}\sin\delta$$

where $\delta := \angle V_1 - \angle V_2$ is the angle difference between V_1 and V_2. Hence the maximum power is delivered on a lossless line if $\delta = \pi/2$ and the maximum power would have been $|V_1||V_2|/X$. This $\delta = \pi/2$ is called the steady-state stability limit. If the load exceeds this limit, there is no solution for δ for this equation. In practice a transmission network operates with $\delta \ll \pi/2$ because a line is typically limited by three other factors. First the voltage drop from the source to the load must be small, e.g., $|V_2|/|V_1| \geq 95\%$. Second δ is usually limited to 30° or 35° by transient stability. Third δ can be limited by the thermal rating of the conductor insulation materials. □

2.2.5 Short line

Consider a three-phase transmission line connecting two buses in balanced operation so we can analyze each phase separately. Assume the line is short and can be modeled by a Π equivalent circuit with only a series impedance $Z = R + \mathbf{i}X$ and no shunt admittances. We explain some properties of complex power transfer over this line.

Let V_i and I_i be the voltages and currents at buses $i = 1, 2$. Let S_{ij}, $i, j = 1, 2$, be the sending-end complex power from bus i to bus j, $i \neq j$, and I_{ij} be the complex current from bus i to bus j. Then

$$S_{ij} = V_i I_{ij}^* = V_i \frac{V_i^* - V_j^*}{Z^*} = \frac{1}{Z^*}\left(|V_i|^2 - V_i V_j^*\right) \tag{2.19}$$

If the voltage magnitudes $|V_i|$, $i = 1, 2$, are fixed, the branch powers depend only on the *power angle* $\theta_{ij} := \theta_i - \theta_j$:

$$S_{ij} = \frac{1}{Z^*}\left(|V_i|^2 - |V_i||V_j|e^{j\theta_{ij}}\right)$$

Taking the sum of the branch powers in (2.19), the complex loss over the line is

$$S_{12}+S_{21}=\frac{|V_1-V_2|^2}{Z^*} \;=\; Z\,|I_{12}|^2$$

where I_{12} is the current from buses 1 to 2. In particular the real power loss is $P_{12}+P_{21}=R|I_{12}|^2$.

Nose curve and voltage collapse.

Suppose bus 1 has a generator with a fixed $V_1 := |V_1|\angle 0^\circ$ supplying a load at bus 2 through a line with impedance Z. Let the power supplied to the load be $-S_{21} = |S_{21}|(\cos\phi+\mathbf{i}\sin\phi) =: P(1+\mathbf{i}\tan\phi)$ where $P>0$ is the active load power and ϕ is the power factor angle. The power flow equation (2.19) hence becomes

$$P(1+\mathbf{i}\tan\phi)=-\frac{1}{Z^*}\left(|V_2|^2-|V_2||V_1|\,e^{\mathbf{i}\theta_{21}}\right) \tag{2.20}$$

where $\theta_{21} := \angle V_2-\angle V_1=\angle V_2$. Voltage support is typically available on the generator side, so we assume $|V_1|$ is fixed even when the load power varies.[2] Voltage support may not be available on the load side and we are interested in the behavior of the load voltage $|V_2|$ as the active load power P increases while keeping the power factor angle ϕ constant.

Fix V_1 and ϕ. For each P, (2.20) defines two real equations in two variables $|V_2|$ and θ_{21}. For this simple system we can analytically solve for $|V_2|$ for each P. Depending on the value of P, there may be zero, one, or two solutions for $|V_2|$. As P varies, the solutions $|V_2|$ trace out a curve called a *nose curve*. As P increases from zero with fixed power factor angle ϕ, there are exactly two solutions for $|V_2|$, one with a high voltage and the other with a low voltage. The difference between the high-voltage solution and the low-voltage solution of $|V_2|$ decreases until they coincide. This is the point where the active load power $P=P_{\max}$ is maximum and represents the limit of power transfer from the voltage source V_1 through the transmission line Z to the load. If P increases further, real solutions for $|V_2|$ cease to exist. This phenomenon is called *voltage collapse*. This is studied in Exercise 2.9. See Chapter **??** for discussions on voltage collapse beyond the infinite bus model.

Short and lossless line $R=0$.

Suppose the series resistance is negligible (which is a reasonable approximation for high voltage transmission lines), $Z=\mathbf{i}X$. Then (2.19) reduces to

$$S_{ij}=\mathbf{i}\frac{1}{X}\left(|V_i|^2-V_iV_j^*\right)$$

[2] An ideal voltage source whose complex bus voltage is fixed regardless of its power generation is called an *infinite bus*.

Hence

$$P_{12} = \frac{|V_1||V_2|}{X} \sin\theta_{12} = -P_{21} \tag{2.21}$$
$$Q_{12} = \frac{1}{X}\left(|V_1|^2 - |V_1||V_2|\cos\theta_{12}\right)$$
$$Q_{21} = \frac{1}{X}\left(|V_2|^2 - |V_1||V_2|\cos\theta_{12}\right)$$

where $\theta_{12} := \angle V_1 - \angle V_2$. This has the following implications.

1 *Transmission efficiency.* The transmission efficiency $\eta := -P_{21}/P_{12} = 1$ since there is zero real power loss. The maximum power transfer $|V_1||V_2|/X$ is proportional to voltage magnitude product. This is another reason why transmission networks tend to operate at very high voltage levels. Indeed doubling the voltage increases the maximum power transfer capability by fourfold.

2 *DC power flow model.* When voltage magnitudes are fixed, the real power depends only on the power angle θ_{12}. When the power angle is small $|\theta_{12}| \approx 0$, $\sin\theta_{12} \approx \theta_{12}$ and the real powers P_{ij} are roughly *linear* in the phase angles (θ_1, θ_2). These assumptions are called the *DC power flow* approximation ($R = 0$, fixed $|V_i|$, small $|\theta_{ij}|$, ignore Q_{ij}); see Chapter 4.6.2 for more details.

3 *Decoupling.* When $|\theta_{12}| \approx 0$, there is a decoupling between real and reactive powers:

$$\frac{\partial P_{12}}{\partial \theta_{12}} = -\frac{\partial P_{21}}{\partial \theta_{12}} = \frac{|V_1||V_2|}{X}\cos\theta_{12} \approx \frac{|V_1||V_2|}{X}$$
$$\frac{\partial P_{12}}{\partial |V_i|} = -\frac{\partial P_{21}}{\partial |V_i|} = \frac{|V_j|}{X}\sin\theta_{12} \approx 0$$

Hence the real powers P_{ij} depend strongly on θ_{12} but not on the voltage magnitudes $|V_k|$.

On the other hand

$$\frac{\partial Q_{ij}}{\partial \theta_{12}} = \frac{|V_1||V_2|}{X}\sin\theta_{12} \approx 0$$

i.e., the reactive powers Q_{ij} depend weakly on the power angle θ_{12}. Moreover

$$\frac{\partial Q_{12}}{\partial |V_2|} = -\frac{|V_1|}{X}\cos\theta_{12} < 0, \qquad \frac{\partial Q_{21}}{\partial |V_2|} = \frac{1}{X}(2|V_2| - |V_1|\cos\theta_{12})$$

Typically $|V_1| \approx |V_2|$ and hence the second expression above is positive. Hence to maintain a high load voltage $|V_2|$, we should increase Q_{21} and/or decrease Q_{12}, i.e., the load should supply reactive power and the generation should absorb reactive power. This motivates the use of reactive power to regulate voltage magnitudes. The decoupling property holds in a network setting as well and leads to a fast algorithm to solve power flow problems; see Chapter 4.4.3.

4 *Out-of-step generators.* When generators are not synchronized, i.e., they operate with slightly different frequencies, the long-run average active power transmitted

across a lossless line is zero. To see this, consider voltages at buses 1 and 2 given by

$$v_1(t) = \sqrt{2}|V_1|\cos(\omega' t + \theta_1)$$
$$v_2(t) = \sqrt{2}|V_2|\cos(\omega t + \theta_2)$$

where the frequency ω' at bus 1 is slightly out of step, with $\omega' \approx \omega$. Write

$$v_1(t) = \sqrt{2}|V_1|\cos(\omega t + \theta_1'(t))$$

with a slowly-varying phase $\theta_1'(t) := \theta_1 + (\omega' - \omega)t$. If the phase $\theta_1'(t)$ varies slowly enough, we can still use the steady-state expressions above as reasonable approximations of powers. Then the short-term active power is given by (from (2.21)):

$$P_{12} = \frac{|V_1||V_2|}{X}\ \sin\left((\omega' - \omega)t + \theta_{12}\right)$$

Hence the long-term average of active power transfer is zero. This is not only ineffective, but highly undesirable because the line current can be very large. In practice protective devices would remove the out-of-step generator.

2.3 Bibliographical notes

There are many excellent textbooks on basic power system concepts and many materials in this chapter follow [1]; see also [2, Chapter 4]. We develop line characteristics in Chapter 2.1 based on basic results in physics that we do not elaborate. For example, the derivation of shunt capacitance c of a transmission line in Chapter 2.1.3 is explained in [1, Chapters 3.7–3.8] or [2, Chapters 4.8–4.12]). The expression (2.6) for the potential v_p at the center of a balanced three-phase transmission line is from [1, Example 3.8, p. 79]. Some of the materials on lossless lines follow [2].

2.4 Problems

Chapter 2.1.

Exercise 2.1. Consider the simplest model of a symmetric three-phase transmission line in balanced operation, as shown in Figure 2.2, with the assumptions

- the conductors are equally spaced at D and have equal radii r;
- $i_a(t) + i_b(t) + i_c(t) = 0$ at all times t;
- $q_a(t) + q_b(t) + q_c(t) = 0$ at all times t.

where $i_k(t)$ are currents and q_k are the total charge per unit length of wire k in coulombs/meter. Show that the per-phase inductance per meter of the three-phase transmission line is

$$l = \frac{\mu_0}{2\pi} \ln \frac{D}{r'} \qquad \text{(in H/m)}$$

where $r' := re^{-\mu_r/4}$, and the per-phase capacitance per meter is

$$c = \frac{2\pi\epsilon}{\ln(D/r)} \qquad \text{(in F/m)}$$

Chapter 2.2.

Exercise 2.2. Consider the per-phase transmission line model described by (2.9). We are to determine the line characteristic impedance Z_c and propagation constant $\gamma\ell$ from two measurements:

1 **Open-circuit test.** The load side is open-circuited so that $I_2 = 0$ and the driving-point impedance is measured as

$$Z_{oc} := \frac{V_1}{I_1}$$

2 **Short-circuit test.** The load side is short-circuited so that $V_2 = 0$ and the driving-point impedance is measured as

$$Z_{cc} := \frac{V_1}{I_1}$$

Derive Z_c and $\gamma\ell$ in terms of Z_{oc} and Z_{sc} (sign ambiguity is fine).

Exercise 2.3 (Lumped-circuit Π model)**.** Consider a general transmission matrix T that maps the receiving-end voltage and current (V_2, I_2) to those (V_1, I_1) at the sending-end:

$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = \underbrace{\begin{bmatrix} a & b \\ c & d \end{bmatrix}}_{T} \begin{bmatrix} V_2 \\ I_2 \end{bmatrix}$$

1 Show that the transmission matrix T in (2.9) has the property $ad - bc = 1$.
2 Suppose $b \neq 0$ in T. Show that the condition $ad - bc = 1$ is necessary and sufficient for interpreting the transmission matrix T as a Π equivalent circuit consisting of a series impedance $Z \neq 0$ and shunt admittances (line charging) Y_1 and Y_2 at the sending and receiving ends respectively (note that Y_1 may not necessarily equal Y_2).

Exercise 2.4 (Surge impedance load (SIL) on lossless line.)**.** Consider a lossless line with $r = g = 0$ that terminates in an impedance load that is equal to the characteristic (surge) impedance $Z_{\text{load}} = Z_c = \sqrt{l/c}\ \Omega$ of the line. The power delivered by a lossless line to the resistive load Z_c is called the *surge impedance loading* (SIL).

1 Show that the voltage magnitude $|V(x)|$ is constant over $x \in [0, \ell]$.
2 Calculate SIL.

Exercise 2.5 (Voltage drop along lossless line)**.** We have derived in Chapter 2.2.4 the voltage $V(x)$ at each point $x \in [0, \ell]$ along a lossless line terminating at an impedance load $Z_{\text{load}} = R_{\text{load}} + \mathbf{i} X_{\text{load}}$ to be (from (2.18)):

$$V(x) = \left(\cos(\beta x) + \frac{Z_c X_{\text{load}}}{|Z_{\text{load}}|^2} \sin(\beta x) \ + \ \mathbf{i} \frac{Z_c R_{\text{load}}}{|Z_{\text{load}}|^2} \sin(\beta x) \right) V_2$$

Assume $\beta\ell < \pi/4$. Prove the following:

1 If the load is purely resistive $Z_{\text{load}} = R_{\text{load}}$ then $|V(x)|$ is an increasing function for all $x \in [0, \ell]$ (i.e., the voltage magnitude $|V(x)|$ drops from the source at $x = \ell$ to the load Z_{load} at $x = 0$) if and only if $R_{\text{load}} \le Z_c$.
2 If the load is purely inductive $Z_{\text{load}} = \mathbf{i} X_{\text{load}}$ with $X_{\text{load}} > 0$ then $|V(x)|$ is an increasing function for all $x \in [0, \ell]$ if and only if

$$X_{\text{load}} \le \frac{\sin(2\beta\ell)}{1 - \cos(2\beta\ell)} Z_c$$

3 If $Z_{\text{load}} = R_{\text{load}}(1 + \mathbf{i})$ then $|V(x)|$ is an increasing function for all $x \in [0, \ell]$ if and only if

$$R_{\text{load}} \ \le \ \left(\sqrt{1 + \frac{1}{\sin^2(2\beta\ell)}} \ - \ \cot(2\beta\ell) \right)^{-1} Z_c$$

Exercise 2.6 (Voltage, reactive power compensation)**.** Consider a generator with voltage and power injection (V_j, s_j) supplying a load with voltage and power *injection* (V_k, s_k) through a transmission line parametrized by series and shunt admittances $\left(y^s_{jk}, y^m_{jk}, y^m_{kj}\right)$. Power balance at the load bus k is (with $y^s_{kj} = y^s_{jk}$)

$$s_k \ = \ \left(y^s_{kj}\right)^{\mathsf{H}} \left(|V_k|^2 - V_k V_j^{\mathsf{H}} \right) + \left(y^m_{kj}\right)^{\mathsf{H}} |V_k|^2 \tag{2.22}$$

Let $y^s_{kj} =: g^s_{kj} + \mathbf{i} b^s_{kj}$ and $y^m_{kj} =: g^m_{kj} + \mathbf{i} b^m_{kj}$ and suppose $g^s_{kj} \ge 0$, $b^s_{kj} < 0$ (inductive) and $g^m_{kj} \ge 0$, $b^m_{kj} \ge 0$ (capacitive). Let $s_k =: p_k + \mathbf{i} q_k$, and $V_i =: |V_i| e^{\mathbf{i}\theta_i}$, $i = j, k$. Use (2.22) to express the receiving real power $-p_k$ and receiving reactive power $-q_k$ in terms of the voltage magnitudes $|V_j|, |V_k|$, and the angle difference $\theta_{kj} := \theta_k - \theta_j$.

Suppose $y^m_{kj} = 0$ (zero shunt), $g^s_{jk} = 0$ (loss line), and $0 < |\theta_{kj}| \le \pi/2$ (power flow solution stability).

1 Show that real power is delivered to the load (i.e., $-p_k > 0$) if and only if $-\pi/2 \le \theta_{kj} < 0$.
2 The next few questions study the relation between load voltage magnitude $|V_k|$ and reactive power injection q_k. Show that:
   1 For DC load (i.e., $q_k = 0$), we must have $|V_k| < |V_j|$, i.e., the load voltage magnitude must be smaller than the generator voltage magnitude.
   2 On the other hand, $|V_k| = |V_j|$ implies that $q_k > 0$, i.e., the load must inject reactive power to maintain a high load voltage magnitude.
   3 If $-q_k > 0$ (i.e., the load withdraws reactive power), then $|V_k| < |V_j| \cos\theta_{kj}$ (i.e., load voltage magnitude will be further suppressed).
3 The power factor angle is $\phi_k := \tan^{-1}(q_k/p_k)$ and the power factor PF is $\cos\phi_k$. Show that

$$1 + \tan\phi_k \tan\theta_{kj} = \frac{|V_k|}{|V_j|\cos\theta_{kj}}$$

   When $|V_k| = |V_j| \cos\theta_{kj}$, what is the PF and is the load withdrawing or injecting real power?
4 Suppose further that $V_j := 1\angle 0^\circ$ and $b^s_{jk} = -1$. Suppose that the load voltage magnitude $|V_k|$ must lie between $[1-\epsilon, 1+\epsilon]$.
   1 At unity power ($q_k = 0$), find the maximum received power $-p_k$ and the corresponding load voltage phasor $V_k = |V_k| e^{\mathbf{i}\theta_k}$. Conclude that the maximum received real power satisfies $-p_k \le \frac{1}{2}$.
   2 Show that the maximum received real power is $-p_k = (1+\epsilon)$ when the load must inject the reactive power $q_k = (1-\epsilon)^2$.

Exercise 2.7 (Voltage, line loss and voltage drop)**.** Consider two buses 1 and 2 connected by a transmission line modeled by a per-phase Π circuit model with series impedance Z and shunt admittance (line charging) $Y/2$ at each end of the line, as shown in Figure 2.8. Let S_{12} be the sending-end complex power from buses 1 to 2 and S_{21} be the sending-end complex power from buses 2 to 1 (or, equivalently, $-S_{21}$ is the receiving-end complex power at bus 2). Note that the direction of load current I_2 is opposite to the convention we used in Chapter 2.2.2.

1 Calculate the complex line loss as a function of voltages (V_1, V_2). Can you express the complex line loss in terms of the load voltage and current (V_2, I_2) instead?
2 Suppose bus 2 is connected to a load that draws a fixed active power P_{load} with a fixed power factor $\cos\phi$ at a fixed voltage magnitude $|V_2|$. Suppose $Z = R + \mathbf{i}X$ and the shunt admittance $Y/2 = \mathbf{i}B/2$ is purely reactive (i.e., zero conductance). Calculate the active power loss P_{line} over the line in terms of the active load power P_{load}, the power factor angle ϕ, and the load voltage $|V_2|$.

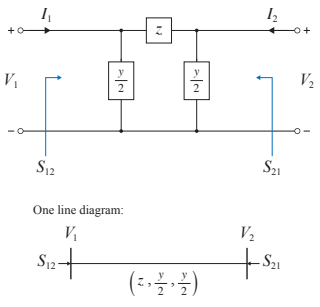


Figure 2.8 Two buses connected by a transmission line.

For the following subproblems, assume $Y = \mathbf{i}B = 0$ (short transmission line).

3. Given the fixed active load power P_{load}, show that the active line loss P_{line} derived in part 2 of the problem is inversely proportional to the squared load voltage $|V_2|^2$ and to the squared power factor $\cos^2\phi$.
4. Suppose now the load at bus 2 is an electric vehicle that draws an active power of $P_{\text{load}} = 20$ kW with unity power factor at a voltage magnitude of $|V_2| = 200$V. Calculate the ratio of the active power loss to the active load power if $R = 0.04\Omega$ (wires with gauge number 6 at 100ft).
5. What is the magnitude of the voltage drop $|V_1 - V_2|$ across the transmission line (the series impedance Z), relative to the load voltage $|V_2|$, in terms of $Z, P_{\text{load}}, |V_2|, \cos\phi$?

Exercise 2.8. Consider the short-line model $S_{12} = (Z^*)^{-1}\left(|V_1|^2 - V_1 V_2^*\right)$ of a transmission line with $Z := y^{-1}e^{\mathbf{i}\phi}$ that connects bus 1 and bus 2. Let V_1, V_2 be the complex voltages at buses 1 and 2 respectively and assume $|V_1| = |V_2| = 1$. Let $\theta_{12} := \angle V_1 - \angle V_2$.

1 For what value of θ_{12} is S_{12} real and nonzero?
2 What is the maximum real power $-P_{21}$ that can be received at bus 2 and what is θ_{12} that delivers it?

Exercise 2.9 (Nose curve and voltage collapse)**.** Consider a voltage source with a fixed magnitude $|V_1|$ supplying a load through a line modeled by a series impedance $z := |z|e^{\mathbf{i}\theta_z}$ with $|\theta_z| < \pi/2$. Let the power supplied to the load be $S_2 = |S_2|(\cos\phi + \mathbf{i}\sin\phi) =: P(1 + \mathbf{i}\tan\phi)$ where $P > 0$ is the active load power and ϕ is the power factor angle. The power flow equation is:

$$P(1 + \mathbf{i}\tan\phi) = -\frac{1}{z^*}\left(|V_2|^2 - |V_2||V_1|e^{\mathbf{i}\theta_{21}}\right) \tag{2.23}$$

where $\theta_{21} := \angle V_2 - \angle V_1$.

1 For each P, solve (2.23) for $|V_2|$ with $|V_1|$ and ϕ fixed.
2 Show that $|V_2|$ behaves as follows as P increases from $P = 0$ with the power factor angle ϕ kept constant: $|V_2|$ is a nonunique rool of a polynomial equation in P. As P increases, the resulting nonunique roots $|V_2|$ trace out a curve called the *nose curve*. As P keeps increasing, eventually, the polynomial equation has no real root, which is the phenomenon of *voltage collapse*.
3 Find the maximum power transfer $P = P_{\max}$ at which solutions for $|V_2|$ exist.

3 Transformer models

A large electric network is composed of multiple areas that have different nominal voltage magnitudes. These areas are connected by transformers that convert between different voltage levels. The ease of converting between voltage levels is an important advantage of AC over DC transmission systems. It allows, for example, the transmission network to operate at $765kV$ to reduce power loss and household appliances to operate at $120V$ for safety. In this chapter we develop transformer models and explain how to analyze a balanced three-phase system that contains transformers.

We start in Chapter 3.1 with models of a single-phase transformer and use them in Chapter 3.2 to develop models of three-phase transformers in balanced operation. We describe in Chapter 3.3 how to refer impedances from one side of a transformer to the other side. We apply this method in Chapter 3.4 to simplify per-phase analysis of circuits that contain transformers. We explain in Chapter 3.5 per-unit normalization that further simplifies the analysis of balanced three-phase systems.

3.1 Single-phase transformer

We first model an ideal single-phase transformer by a transmission matrix and then describe circuit models of a nonideal single-phase transformer.

3.1.1 Ideal transformer

An *ideal transformer* has no loss (zero resistance), no leakage flux, and the magnetic core has infinite permeability. Let N_1 be the number of turns in the primary winding, N_2 that in the secondary winding, and

$$n := \frac{N_2}{N_1}, \qquad a := \frac{1}{n} = \frac{N_1}{N_2}$$

An ideal transformer is represented schematically in Figure 3.1. We will call n the *voltage gain* and its reciprocal a the *turns ratio*. The voltage gain n relates the voltages

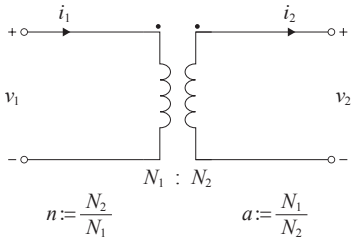


Figure 3.1 Single-phase ideal transformer.

and currents in the primary and secondary circuits, both at all times in the time domain:

$$\frac{v_2(t)}{v_1(t)} = n, \qquad \frac{i_2(t)}{i_1(t)} = a$$

and in the phasor domain:

$$\frac{V_2}{V_1} = n, \qquad \frac{I_2}{I_1} = a$$

This relation can also be written as

$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = \begin{bmatrix} a & 0 \\ 0 & n \end{bmatrix} \begin{bmatrix} V_2 \\ I_2 \end{bmatrix} \tag{3.1}$$

The matrix on the right-hand side is called a *transmission matrix* of an ideal transformer. It maps (V_2, I_2) to (V_1, I_1). The dot notation indicates that the currents I_1, I_2 are defined to be positive when one flows into and the other out of the dotted terminals, as indicated in Figure 3.1. This notation is convenient when we use single-phase transformers to construct three-phase transformers.

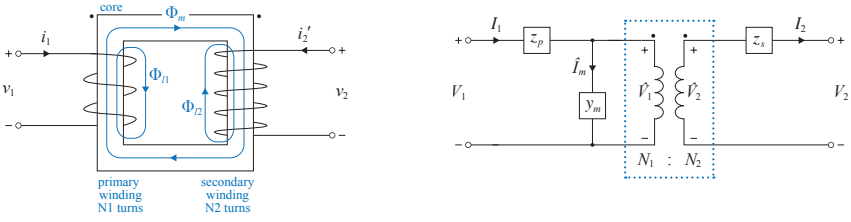
The ratio of the complex receiving-end to sending-end power is

$$\frac{-S_{21}}{S_{12}} := \frac{V_2 I_2^*}{V_1 I_1^*} = n \cdot a = 1$$

i.e., an ideal transformer has no power loss.

3.1.2 Nonideal transformer

A real transformer has power losses due to resistance in the windings $(r|I|^2)$, eddy currents and hysteresis losses. It also has nonzero leakage fluxes and finite permeability of the magnetic core. Figure 3.2(a) shows elements of a (nonideal) transformer. The primary winding has N_1 turns around the magnetic core and the secondary winding has N_2 turns. The mutual flux Φ_m due to the currents i_1 and i_2' links all the turns of the primary and secondary coils. The two dots indicate that the mutual flux components due to i_1 and i_2' add when these currents both enter (or exit) the dotted terminals according to the right-hand rule. The leakage fluxes Φ_{l1} and Φ_{l2} links the individual



(a) Nonideal transformer (b) Circuit model

Figure 3.2 Single-phase nonideal transformer. The dotted box represents an ideal transformer with $a := N_1/N_2$.

coils. The flux linkages $\lambda_{l1} =: L_{l1} i_1$ and $\lambda_{l2} =: L_{l2} i_2'$ due to Φ_{l1} and Φ_{l2} are proportional to the currents i_1 and i_2' respectively. The proportionality constants L_{l1}, L_{l2} are called inductances. Then the total flux linkages of the primary and secondary circuits are the sums of the leakage flux linkages and the mutual flux linkage:

$$\lambda_1 \;=\; \lambda_{l1} + N_1 \Phi_m, \qquad\qquad \lambda_2 \;=\; \lambda_{21} + N_2 \Phi_m$$

The voltages are

$$v_1 \;=\; r_1 i_1 + \frac{d\lambda_1}{dt} \;=\; r_1 i_1 + L_{l1} \frac{di_1}{dt} + N_1 \frac{d\Phi_m}{dt} \tag{3.2a}$$

$$v_2 \;=\; r_2 i_2' + \frac{d\lambda_2}{dt} \;=\; r_2 i_2' + L_{l2} \frac{di_2'}{dt} + N_2 \frac{d\Phi_m}{dt} \tag{3.2b}$$

where $r_1 i_1$ and $r_2 i_2'$ represent power losses in the core. The model for an ideal transformer neglects losses ($r_1 = r_2 = 0$) and leakage fluxes ($\lambda_{l1} = \lambda_{l2} = 0$) in (3.2) and hence $v_1 = N_1 \frac{d\Phi_m}{dt}$ and $v_2 = N_2 \frac{d\Phi_m}{dt}$, yielding $v_1/v_2 = N_1/N_2$.

The total magnetomotive force F due to the currents i_1 and i_2' is proportional to the mutual flux Φ_m:

$$F \;=\; N_1 i_1 + N_2 i_2' \;=\; R\Phi_m \tag{3.3}$$

where R is called the reluctance of the core. The model for an ideal transformer assumes infinite permeability of the magnetic core and hence $R = 0$, yielding $i_1/(-i_2') = N_2/N_1$. In practice the magnetic core has finite permeability, i.e., $R > 0$ and the the magnetomotive force F is nonzero. When the secondary circuit is open, $i_2' = 0$. The resulting primary current, denoted $\hat{i}_m$, is called the primary magnetizing current and satisfies $N_1 \hat{i}_m = R\Phi_m$ from (3.3).[1] Define

$$\hat{v}_1 \;:=\; N_1 \frac{d\Phi_m}{dt} \;=\; L_m \frac{d\hat{i}_m}{dt}, \qquad\qquad \hat{v}_2 \;:=\; N_2 \frac{d\Phi_m}{dt} \;=\; \frac{N_2}{N_1} \hat{v}_1$$

[1] Instead of $i_m := (R/N_1)\Phi_m$, we can define $i_m' := (R/N_2)\Phi_m$ as the secondary magnetizing current when the primary circuit is open $i_1 = 0$. In this case the shunt admittance y_m in Figure 3.4(a) will be in the secondary circuit.

where $L_m := N_1^2/R$. Substituting into (3.2) yields, denoting $i_2 := -i_2'$, we have

$$\text{Nonideal elements:} \quad v_1 = r_1 i_1 + L_{l1}\frac{di_1}{dt} + \hat{v}_1, \quad \hat{v}_1 = L_m \frac{d\hat{i}_m}{dt}, \quad v_2 = -r_2 i_2 - L_{l2}\frac{di_2}{dt} + \hat{v}_2$$

$$\text{Ideal transformer:} \quad \hat{v}_2 = \frac{N_2}{N_1}\hat{v}_1, \quad i_2 = \frac{N_1}{N_2}(i_1 - \hat{i}_m)$$

where the last equality follows from substituting $R\Phi_m = N_1 \hat{i}_m$ into (3.3). This set of equations in the phasor domain is

$$\text{Nonideal elements:} \quad V_1 = z_p I_1 + \hat{V}_1, \quad \hat{I}_m = y_m \hat{V}_1, \quad \hat{V}_2 = z_s I_2 + V_2 \tag{3.4a}$$

$$\text{Ideal transformer:} \quad \hat{V}_2 = \frac{N_2}{N_1}\hat{V}_1, \quad I_2 = \frac{N_1}{N_2}(I_1 - \hat{I}_m) \tag{3.4b}$$

where the series impedances $z_p := r_1 + \omega L_{l1}$ and $z_s := r_2 + \omega L_{l2}$ model the core losses and leakage fluxes in the primary and secondary circuits respectively, and the shunt admittance $y_m := 1/(\omega L_m) = R/(\omega N_1^2)$ models the finite permeability of the core. The model (3.4) can be interpreted as the circuit in Figure 3.2(b). Variables with hats denote internal variables.

The *end-to-end behavior* of the nonideal transformer can be described by a transmission matrix that maps (V_2, I_2) to (V_1, I_1) (see Chapter 2.2.1 for the transmission matrix of a transmission line). Eliminating the internal variables (with hats) from (3.4), the transmission matrix is given by (Exercise 3.1)

$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = \begin{bmatrix} a' & a' z_s + n z_p \\ a y_m & n + a z_s y_m \end{bmatrix} \begin{bmatrix} V_2 \\ I_2 \end{bmatrix} \tag{3.5}$$

where $n := N_2/N_1$, $a := N_1/N_2$, and $a' := a(1 + z_p y_m)$. We will refer to such a model that describes the end-to-end behavior as an *external model*. An equivalent external model to the transmission matrix is an *admittance matrix* that maps (V_1, V_2) to $(I_1, -I_2)$:

$$\begin{bmatrix} I_1 \\ -I_2 \end{bmatrix} = \frac{1}{\eta}\begin{bmatrix} n + a z_s y_m & -1 \\ -1 & a' \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}$$

where $\eta := a' z_s + n z_p$. We will freely use either matrix for describing the end-to-end behavior of a two-terminal device such as a transformer or a transmission line.

In the following we present three circuit models derived from that in Figure 3.2(b). Their relation is shown in 3.3. The circuit model in Figure 3.2(b) is equivalent to a T equivalent circuit (Chapter 3.1.3). The T equivalent circuit can be approximated by a simplified model whose parameters can be determined by short-circuit and open-circuit tests (Chapter 3.1.4). The circuit model in Figure 3.2(b) is also equivalent to a circuit consisting of two ideal transformers connected by a unitary voltage network (Chapter 3.1.5). The unitary voltage network can be generalized to model nonstandard transformers with multiple windings, e.g., split-phase transformer. These models reduce to the same model when the shunt admittance y_m in Figure 3.2(b) is assumed zero (i.e., open-circuited). We emphasize that, by equivalence, we only mean that two circuits have the same end-to-end behavior, i.e., same transmission or admittance matrices,

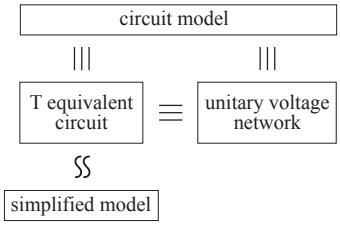


Figure 3.3 Relation between different circuit models of transformers.

but their internal variables may take different values. This is important, e.g., when we try to determine transformer parameter values from measurements using these circuit models; the derivation should use only terminal variables, not internal variables, as we discuss in Chapter 3.1.3.

3.1.3 T equivalent circuit

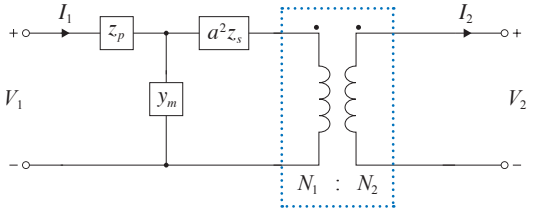


Figure 3.4 T equivalent circuit.

It is shown in Exercise 3.1 that the circuit model in Figure 3.2(b) has the same transmission matrix (3.5) and hence the same end-to-end behavior as what is called the *T equivalent circuit* of the transformer shown in Figure 3.4. The difference between the models in Figure 3.2(b) and in Figure 3.4 is the position and the scaling of the leakage impedance z_s; this is called referring z_s on the secondary side to the primary side and is discussed in Chapter 3.3.1.

Even though the circuit model in Figure 3.2(b) and the T equivalent circuit in Figure 3.4 have the same transmission matrix, their internal variables may not be equal because of the reference of z_s to the primary side. Indeed (3.4) describes the internal variables of the model in Figure 3.2(b), but not necessarily those in the T equivalent circuit in Figure 3.4. For instance, when the secondary circuit is shorted, i.e., setting $V_2 = 0$, the internal variables $\hat{V}_1$ and $\hat{V}_2$ are nonzero in general in Figure 3.2(b), as determined by (3.4), but these voltages are zero in Figure 3.4. This has implications on parameter determination as we now explain.

Parameter determination.

Two simple tests are often used to determine transformer model parameters:

1 *Short-circuit test* ($V_2 = 0$). With the secondary circuit short-circuited, the primary voltage V_{sc} and primary current I_{sc} are measured. The primary short-circuit voltage V_{sc} is called the *impedance voltage*.
2 *Open-circuit test* ($I_2 = 0$). With the secondary circuit open, the primary voltage V_{oc} and primary current I_{oc} are measured.

To determine the parameters (z_p, z_s, y_m) of the transmission matrix T in (3.5), note that during the short-circuit test, the voltage on the primary side of the ideal transformer is zero. Hence

$$V_{sc} = \left(z_p + \left(y_m + \frac{1}{a^2 z_s} \right)^{-1} \right) I_{sc} \tag{3.6a}$$

During the open-circuit test, the secondary current $I_2 = 0$ and hence there is zero current on the primary side of the ideal transformer. Hence

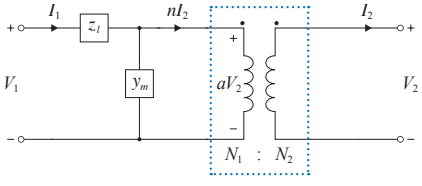
$$V_{oc} = \left(z_p + \frac{1}{y_m} \right) I_{oc} \tag{3.6b}$$

Since there are three unknowns (z_p, z_s, y_m), they cannot be uniquely determined from the two equations in (3.6). Additional measurements will be needed to determine (z_p, z_s, y_m), e.g. measurements of separate dc resistances in the primary and secondary circuits. Sometimes y_m is assumed to be zero (open-circuited) so that (3.6a) becomes $V_{sc} = (z_p + a^2 z_s) I_{sc}$, yielding the total leakage impedance $z_p + z_s$. Alternatively assuming $z_p = \eta z_s$ with known η results in two nonlinear equations in two unknowns (z_s, y_m).

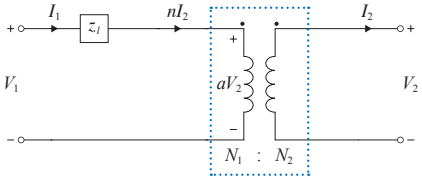
It may seem that we can measure the current I_2 in the T equivalent circuit in Figure 3.4 during a short-circuit test and use it to determine (z_p, z_s, y_m), but this is not the case because it will involve internal variables. Even though we have informally justified (3.6) using internal variables in the T equivalent circuit, e.g., the voltage and current on the primary side of the ideal transformer, we should be careful with this line of reasoning. A more rigorous derivation of (3.6) uses the circuit model in Figure 3.2(b), by setting $V_2 = 0$ in (3.4) (Exercise 3.2). In this case, even if the short-circuit current I_2 is also measured, there are 6 unknowns $(\hat{V}_1, \hat{V}_2, \hat{I}_m; z_p, z_s, y_m)$ but only 5 equations in (3.4) and hence these unknowns cannot be uniquely determined from just the short-circuit and open-circuit tests either. This implies that we cannot apply the measured value of short-circuit current I_2 to determine (z_p, z_s, y_m).

3.1.4 Simplified model

In practice the shunt admittance y_m is much smaller than the leakage admittances (see Example 3.1). Specifically when $|y_m| \ll 1/|a^2 z_s|$ or $|\epsilon| := |a^2 z_s y_m| \ll 1$, we interchange y_m and $a^2 z_s$ to obtain the simplified model in Figure 3.5(a) with $z_l = z_p + a^2 z_s$. An even simpler model assumes $y_m = 0$, as shown in Figure 3.5(b).



(a) Simplified model



(b) $y_m = 0$

Figure 3.5 (a) Simplified model of nonideal transformer including power losses, leakage flux and finite permeability of magnetic core with $z_l := z_p + a^2 z_s$. (b) Simplified model assuming infinite permeabilitiy.

Transmission matrix.

Apply KCL, KVL and Ohm's law to the model in Figure 3.5(a) to get:

$$V_1 = z_l I_1 + aV_2, \qquad I_1 = y_m(aV_2) + nI_2$$

Hence the transmission matrix $\hat{M}$ is given by

$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = \underbrace{\begin{bmatrix} a\,(1 + z_l y_m) & n z_l \\ a y_m & n \end{bmatrix}}_{\hat{M}} \begin{bmatrix} V_2 \\ I_2 \end{bmatrix} \tag{3.7a}$$

We mostly use the simplified model in Figure 3.5, or equivalently, in (3.7a). When $y_m = 0$ the relation (3.7a) can be equivalently expressed in terms of an admittance matrix Y:

$$\begin{bmatrix} I_1 \\ -I_2 \end{bmatrix} = \underbrace{\frac{1}{z_l}\begin{bmatrix} 1 & -a \\ -a & a^2 \end{bmatrix}}_{Y} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \tag{3.7b}$$

When $z_l = y_m = 0$ the model (3.7a) reduces to (3.1) for an ideal transformer.

Approximation to T equivalent circuit.

We now justify the model in Figure 3.5(a) with $z_l = z_p + a^2 z_s$ as a reasonable approximation of the T equivalent circuit in Figure 3.4(b) when y_m is small. Let M and $\hat{M}$ denote that transmission matrices in (3.5) and (3.7a) respectively. Their difference is

$$\hat{M} - M = \begin{bmatrix} a(1+z_l y_m) & n z_l \\ a y_m & n \end{bmatrix} - \begin{bmatrix} a(1+z_p y_m) & a(1+z_p y_m) z_s + n z_p \\ a y_m & n + a z_s y_m \end{bmatrix} = \epsilon \begin{bmatrix} a & -n z_p \\ 0 & -n \end{bmatrix}$$

where $\epsilon := a^2 z_s y_m$. The conductance in the shunt admittance is negligible in practice and hence the shunt admittance y_m due to the primary magnetizing current takes the form $y_m = (\mathbf{i} x_m)^{-1} = -\mathbf{i} b_m$ with $b_m > 0$. The leakage impedance z_p takes the form $z_p = r_p + \mathbf{i} x_p$ with $r_p > 0$ and $x_p > 0$; similarly for z_s. Suppose $z_p = \eta z_s$ for some real number $\eta > 0$ and $|\epsilon| \ll 1$. Then the relative error can be shown to satisfy (Exercise 3.3)

$$\frac{\|\hat{M} - M\|}{\|M\|} < |\epsilon| \ll 1$$

where the matrix norm $\|A\|$ is the sum norm $\|A\| := \sum_{i,j} |A_{ij}|$, or the l_1 vector norm when the $n \times n$ matrix A is treated as a vector in $\mathbb{C}^{n^2}$ (see Appendix A.8.3 for matrix norms). Note that for $a < 1$, the model parameters (z_l, y_m) should be on the high voltage side. When the shunt admittance is neglected $y_m = 0$, these two models are the same, i.e., $\hat{M} = M$.

Parameter determination.

The parameters (z_l, y_m) of the simplified model in Figure 3.5(a), or equivalently, in (3.7a), can be uniquely determined from two simple tests:

1 *Short-circuit test* ($V_2 = 0$). With the secondary circuit short-circuited, the primary voltage V_{sc} and current I_{sc} are measured. Then, from Figure 3.5,

$$z_l = \frac{V_{sc}}{I_{sc}}$$

The primary short-circuit voltage V_{sc} is called the impedance voltage.

2 *Open-circuit test* ($I_2 = 0$). With the secondary circuit open, the primary voltage V_{oc} and current I_{oc} are measured. Then $V_{oc} = (z_l + 1/y_m) I_{oc}$ and hence

$$\frac{1}{y_m} = \frac{V_{oc}}{I_{oc}} - \frac{V_{sc}}{I_{sc}}$$

Example 3.1 (Parameter determination)**.** Consider a single-phase distribution (step-down) transformer with the following ratings: 2.9 MVA, 7.2 kV / 240 V. Construct the equivalent circuit model in Figure 3.5 from the following test results:

1 *Short-circuit test* ($V_2 = 0$). With the secondary circuit short-circuited, a voltage $|V_{sc}| = 500\,\text{V}$ is applied to the primary circuit that causes the rated primary current $|I_1^s|$ to flow.

2 *Open-circuit test* ($I_2 = 0$). With the secondary circuit open, the rated voltage $|V_{oc}| = 7.2\,\text{kV}$ is applied to the primary circuit. This caused a current of $|I_{oc}| = 7\text{A}$ to flow in the primary circuit.

Assume $z_l = \mathbf{i}x_l$ and $y_m = (\mathbf{i}x_m)^{-1}$. Determine x_l and x_m.

Solution. In the short-circuit test the secondary voltage $V_2 = 0$. Hence the voltage on the primary side of the *ideal* transformer is zero and the shunt reactance x_m is effectively short-circuited, leaving only the leakage reactance x_l in the primary circuit. Since the rated primary current is $|I_{sc}| = 2.9\,\text{MVA}/7.2\,\text{kV} = 403\text{A}$, we have $|V_{sc}| = |I_{sc}z_l| = |I_{sc}|x_l$. Hence $x_l = 500\,\text{V}/403\,\text{A} = 1.24\ \Omega$.

In the open-circuit test the secondary current $I_2 = 0$ and hence there is zero current on the primary side of the *ideal* transformer (see Figure 3.5). Hence $|V_{oc}| = |I_{oc}(z_l + 1/y_m)| = |I_{oc}|(x_l + x_m)$, and $x_m = |V_{oc}|/|I_{oc}| - x_l = 7.2\,\text{kV}/7\,\text{A} - 1.24 = 1.03\,\text{k}\Omega$.

As expected, $|y_m| \ll 1/|z_l|$. □

In transformer ratings, the ratio of secondary open-circuit voltage to the primary open-circuit voltage is usually taken to be the voltage gain n, even though more precisely it should be

$$\frac{V_2}{V_1} = n \cdot \frac{1/y_m}{z_l + 1/y_m}$$

In practice the resistances due to core losses are much smaller than the reactances due to leakage fluxes and finite permeability of the core so that $z_l \approx \mathbf{i}x_l$ and $y_m \approx -\mathbf{i}b_m$. Moreover $b_m \ll 1/x_l$. For Example 3.1

$$\frac{V_2}{V_1} = n\,\frac{x_m}{x_l + x_m} \;=\; \frac{1.03\,\text{k}\Omega}{1.03\,\text{k}\Omega + 1.24\,\Omega} \;\simeq\; n$$

Parameter determination from transformer ratings when $y_m := 0$.

If $y_m := 0$ then the model parameter is just the leakage impedance z_l in the primary circuit, which can be determined from the short-circuit test, $z_l = V_{sc}/I_{sc}$. Moreover its magnitude can be determined from typical transformer ratings, as follows.

A typical specification of a three-phase transformer includes:

- Three-phase power rating $|S_{3\phi}|$.
- Rated primary line-to-line voltage $|V_{\text{pri}}|$ and rated primary line current $|I_{\text{pri}}|$.
- Rated secondary line-to-line voltage $|V_{\text{sec}}|$ and rated secondary line current $|I_{\text{sec}}|$.
- Impedance voltage β on the primary side, per phase, as a percentage of the rated primary voltage. The shunt admittance is assumed zero.

As mentioned above, the impedance voltage is the voltage drop across the leakage impedance z_l on the primary side of each *single-phase* transformer in a short-circuit test.

The β specification means that the voltage needed on the primary side to produce the rated primary current across each single-phase transformer is β, as a percentage of the rated primary voltage. We emphasize that the short-circuit voltage and current needed to derive z_l should be those across each single-phase transformer, which depends on the configuration of the primary circuit. If the primary circuit is in Δ configuration then the short-circuit voltage and current on the primary side of the single-phase transformer are (assuming balanced positive sequence):

$$\Delta \text{ configuration:} \quad |V_{sc}| = |V_{ab}| = \beta |V_{\text{pri}}|, \qquad |I_{sc}| = |I_{ab}| = \left| \frac{I_{\text{pri}}}{\sqrt{3}} e^{\mathbf{i}\pi/6} \right|$$

If the primary circuit is in Y configuration then the short-circuit voltage and current on the primary side of the single-phase transformer are:

$$Y \text{ configuration:} \quad |V_{sc}| = |V_{an}| = \beta \left| \frac{V_{\text{pri}}}{\sqrt{3}\, e^{\mathbf{i}\pi/6}} \right|, \qquad |I_{sc}| = |I_{an}| = |I_{\text{pri}}|$$

Since $z_l = V_{sc}/I_{sc}$ we therefore have,

$$\Delta \text{ configuration:} \quad |z_l| = \frac{\sqrt{3}\beta |V_{\text{pri}}|}{|I_{\text{pri}}|}; \quad Y \text{ configuration:} \quad |z_l| = \frac{\beta |V_{\text{pri}}|}{\sqrt{3}|I_{\text{pri}}|} \tag{3.8a}$$

We reiterate that V_{pri} denotes the line-to-line voltage even for Y configuration; otherwise $|z_l| = \beta |V_{\text{pri}}|/|I_{\text{pri}}|$ for Y configuration if the rated voltage V_{pri} is line-to-neutral.

Sometimes the primary line current $|I_{\text{pri}}|$ is not specified directly. In that case z_l can be determined from the power and voltage ratings ($|S_{3\phi}|, |V_{\text{pri}}|$), as follows. If the primary circuit is in Δ configuration then the short-circuit voltage and current on the primary side of the single-phase transformer are (assuming balanced positive sequence):

$$\begin{aligned} \Delta \text{ configuration:} \quad & |S_{3\phi}| = 3|S_{\phi}| = 3|V_{ab}||I_{ab}| \\ & |V_{sc}| = |V_{ab}| = \beta |V_{\text{pri}}|, \qquad |I_{sc}| = |I_{ab}| = \frac{|S_{3\phi}|}{3|V_{\text{pri}}|} \end{aligned}$$

Note that $\frac{|S_{3\phi}|}{3|V_{\text{pri}}|}$ is the rated primary current produced in the short-circuit test. If the primary circuit is in Y configuration then the short-circuit voltage and current on the primary side of the single-phase transformer are:

$$\begin{aligned} Y \text{ configuration:} \quad & |S_{3\phi}| = 3|S_{\phi}| = 3|V_{an}||I_{an}| \\ & |V_{sc}| = |V_{an}| = \beta \left| \frac{V_{\text{pri}}}{\sqrt{3}\, e^{\mathbf{i}\pi/6}} \right|, \qquad |I_{sc}| = |I_{an}| = \frac{|S_{3\phi}|}{3\left| \frac{V_{\text{pri}}}{\sqrt{3}\, e^{\mathbf{i}\pi/6}} \right|} = \frac{|S_{3\phi}|}{\sqrt{3}|V_{\text{pri}}|} \end{aligned}$$

Since $z_l = V_{sc}/I_{sc}$ we therefore have,

$$\Delta \text{ configuration:} \quad |z_l| = \frac{3\beta |V_{\text{pri}}|^2}{|S_{3\phi}|}; \quad Y \text{ configuration:} \quad |z_l| = \frac{\beta |V_{\text{pri}}|^2}{|S_{3\phi}|} \tag{3.8b}$$

As mentioned above, V_{pri} denotes the line-to-line voltage even for Y configuration;

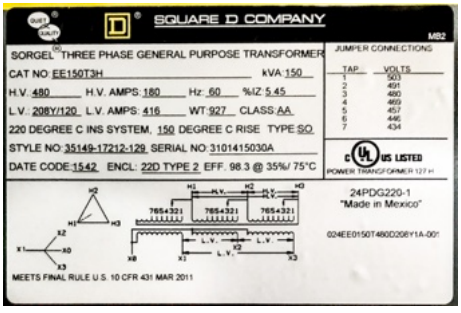


Figure 3.6 The transformer ratings.

otherwise $|z_l| = 3\beta|V_{\text{pri}}|^2/|S_{3\phi}|$ for Y configuration if the rated voltage V_{pri} is line-to-neutral.

Example 3.2 (Transformer ratings)**.** Figure 3.6 shows a typical specification of a three-phase transformer in ΔY configuration:

- Three-phase power rating $|S_{3\phi}| = 150\,\text{kVA}$.
- Rated primary line-to-line (high) voltage $|V_{\text{pri}}| = 480\,\text{V}$ in Δ configuration with rated primary line current $|I_{\text{pri}}| = 180\,\text{A}$.
- Rated secondary line-to-line (low) voltage $|V_{\text{sec}}| = 208Y/120\,\text{V}$ in Y configuration with rated secondary line current $|I_{\text{sec}}| = 416\,\text{A}$. This notation means that the secondary side is Y-configured with a line-to-line voltage of 208 V and line-to-neutral voltage of 120 V.
- Impedance voltage $\beta = 5.45\%$ on the primary side (the shunt admittance is assumed zero).

Verify that the rated line currents on the primary and secondary sides are consistent with the power rating and voltage ratings. Determine the magnitude $|z_l|$ of the leakage impedance of the transformer.

Solution. The primary side is in Δ configuration and hence we have

$$|S_{3\phi}| = 3|S_{ab}| = 3|V_{ab}\,\bar{I}_{ab}| = 3|V_{\text{pri}}||I_{ab}|$$

Since (assuming balanced positive sequence)

$$I_a = I_{ab} - I_{ca} = I_{ab}\left(1 - e^{\mathrm{i}2\pi/3}\right) = I_{ab}\cdot\sqrt{3}\,e^{-\mathrm{i}\pi/6}$$

we have $|I_{\text{pri}}| = \sqrt{3}\,|I_{ab}|$. Hence

$$|S_{3\phi}| = \sqrt{3}\,|V_{\text{pri}}||I_{\text{pri}}|$$

The rated line-to-line voltage $|V_{\text{pri}}| = |V_{ab}| = 480\text{V}$. The rated line current $|I_{\text{pri}}| = |I_a| =$

180A. Hence

$$\sqrt{3}\,|V_{\text{pri}}|\,|I_{\text{pri}}| = \sqrt{3}\cdot 480\cdot 180 \;=\; 149.65\ \text{kVA}$$

which is approximately the power rating $|S_{3\phi}| = 150$ kVA.

The secondary side is in Y configuration and hence we have

$$|S_{3\phi}| = 3\,|S_{an}| \;=\; 3\,|V_{an}\,\bar{I}_{an}| \;=\; 3\left|\frac{V_{\text{sec}}}{\sqrt{3}e^{\mathbf{i}\pi/6}}\right||I_{\text{sec}}| \;=\; \sqrt{3}\,|V_{\text{sec}}|\,|I_{\text{sec}}|$$

where the third equality follows since $V_{\text{sec}} = V_{ab} = V_{an}\left(\sqrt{3}e^{\mathbf{i}\pi/6}\right)$ is the line-to-end voltage. The rated secondary line-to-line voltage is $|V_{\text{sec}}| = 208\,\text{V}$ and the line current $|I_{\text{sec}}| = 416\,\text{A}$, and hence

$$\sqrt{3}\,|V_{\text{sec}}|\,|I_{\text{sec}}| \;=\; \sqrt{3}\cdot 208\cdot 416 \;=\; 149.87\text{kVA}$$

which is approximately the power rating 150 kVA.

From (3.8a) the magnitude $|z_l|$ of the leakage impedance of each single-phase transformer is (β is the impedance voltage on the *primary* side)

$$|z_l| \;=\; \frac{\sqrt{3}\beta|V_{\text{pri}}|}{|I_{\text{pri}}|} \;=\; \frac{\sqrt{3}\cdot 5.45\%\cdot 480\,\text{V}}{180\,\text{A}} \;=\; 0.2517\,\Omega$$

□

Distribution system transformers.

In the US, single-phase or three-phase stepdown transformers are typical in the distribution system. The most common three-phase system voltage on the primary side is 12.47 kV, serving more than 50% of loads. This is the line-to-line voltage (magnitude) and hence the line-to-neutral voltage is $|V_{an}| = 12.47/\sqrt{3} = 7.2\,\text{kV}$. A typical primary side current rating is $|I_{an}| = 400$A. Hence the total (three-phase) rated apparent power is $|S_{3\phi}| = 3|V_{an}||I_{an}| = (3)(7.2)(400) = 8.6\,\text{MVA}$. Other common distribution system voltages and their total power at 400A are shown in Table 3.1. The advantages of a

line-to-line voltage (kV) $\lvert V_{ab}\rvert$	phase voltage (kV) $\lvert V_{an}\rvert$	total power (MVA) $\lvert S_{3\phi}\rvert$
4.8	2.8	3.3
12.47	7.2	8.6
22.9	13.2	15.9
34.5	19.9	23.9

Table 3.1 Typical distribution system voltages (line-to-line) and their total (three-phase) power rating at 400A current.

higher-voltage system include:

- It can carry more power for a given ampacity.
- It has a smaller voltage drop for a given level of power flow, requiring fewer voltage regulators and capacitor banks for voltage support (see Exercise 2.7.5).
- It has a smaller line loss for a given level of power flow (see Exercise 2.7).
- It can cover a larger service area since it has a smaller voltage drop and a smaller line loss. Roughly, for the same load density, the area covered increases linearly with voltage.
- It requires fewer substations since it covers a larger service area, which can be a big cost saving.

The disadvantages of a higher-voltage system include:

- It may be less reliable, since a longer circuit can lead to more customer interruptions.
- Crew safety is a bigger concern with a higher voltage.
- Higher voltage equipment costs more, from transformers to cables to voltage regulators.

The 12.47 kV system seems to strike a good balance.

On the primary side, one end of the winding typically connects to one of the primary phases and the other end connects to the transformer case which is connected to the neutral wire of the three-phase system and also earth ground. On the secondary side, the 240V is center-tapped and the center neutral wire is grounded, making the two ends "hot" with respect to the center tap. These three wires run down the service drop to the meter and electric panel of a house. This is shown in Figure 3.7. Connecting a load

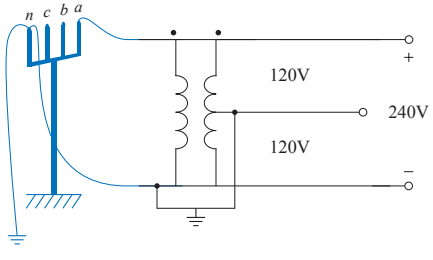


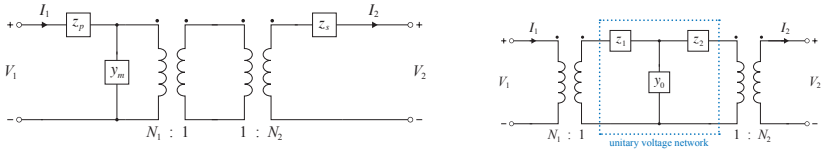
Figure 3.7 A common single-phase distribution transformer in the US.

between either hot wire and the neutral gives 120V while connecting it between both hot wires gives 240V. Note that the transformer is single-phase. This is the split-phase 120/240 V system typical in the US.

3.1.5 Model with unitary voltage network

Single-phase two-winding transformer.

As far as the end-to-end behavior is concerned, the transformer model in Figure 3.2(b) is equivalent to the model in Figure 3.8(a) where the ideal transformer with turns ratio N_1/N_2 is replaced by two ideal transformers in series with turns ratios N_1 and $1/N_2$. Referring the leakage impedances (z_p, z_s) and shunt admittance y_m to the other sides



(a) Equivalent model (b) Unitary voltage network

Figure 3.8 Models of nonideal transformer with unitary voltage network.

of the ideal transformers using (3.14) in Chapter 3.3, this model is equivalent to the one in Figure 3.8(b) where

$$y_0 := N_1^2 y_m, \qquad z_1 := \frac{z_p}{N_1^2}, \qquad z_2 := \frac{z_s}{N_2^2} \tag{3.9}$$

The network between the two ideal transformers is sometimes referred to as a unitary voltage network because the nominal voltage of the network is 1 pu if the scaled nominal voltages $V_1^{\text{nom}}/N_1 = V_2^{\text{nom}}/N_2$ on both sides of the (nonideal) transformer is used as the voltage base for per-unit normalization (per-unit normalization is studied in Appendix 3.5). Note that no nodes in the transformer models may be grounded. The main advantage of modeling a nonideal transformer this way is that the unitary voltage network can be generalized from the simple network in Figure 3.8(b) to a more general network that can be used to model nonstandard transformers with multiple windings; see below.

We now derive the admittance matrix that maps (V_1, V_2) to $(I_1, -I_2)$. First focus on the unitary voltage network, shown in Figure 3.9, where $y_1 := 1/z_1 = N_1^2 y_p$, $y_2 := 1/z_2 = N_2^2 y_s$ with $y_p := 1/z_p$, $y_s := 1/z_s$. Variables with hats denote internal variables.[2] The

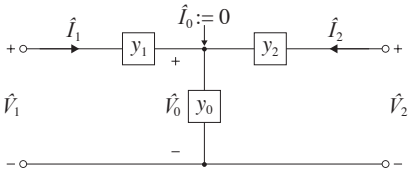


Figure 3.9 Unitary voltage network of the model in Figure 3.8(b).

[2] The explicit separation of internal variables (e.g., $\hat{V}_i, \hat{I}_i$) and terminal variables (e.g., V_i, I_i) may not be

variables $(\hat{V}_0, \hat{V}_1, \hat{V}_2)$ are defined as voltage drops as shown in the figure and $(\hat{I}_0, \hat{I}_1, \hat{I}_2)$ are the current injections at these nodes with $\hat{I}_0 := 0$. Then

$$\hat{I}_1 = y_1(\hat{V}_1 - \hat{V}_0), \qquad \hat{I}_2 = y_2(\hat{V}_2 - \hat{V}_0), \qquad \hat{I}_0 + \hat{I}_1 + \hat{I}_2 = y_0\hat{V}_0 \tag{3.10}$$

or in terms of admittance matrix (we will study admittance matrices in detail in Chapter 4)

$$\begin{bmatrix} \hat{I}_0 \\ \hat{I}_1 \\ \hat{I}_2 \end{bmatrix} = \begin{bmatrix} y_0 + y_1 + y_2 & -y_1 & -y_2 \\ -y_1 & y_1 & 0 \\ -y_2 & 0 & y_2 \end{bmatrix} \begin{bmatrix} \hat{V}_0 \\ \hat{V}_1 \\ \hat{V}_2 \end{bmatrix}$$

Since $\hat{I}_0 = 0$ we can eliminate $\hat{V}_0$ and derive the Kron-reduced admittance matrix Y_{uvn} that maps $(\hat{V}_1, \hat{V}_2)$ to $(\hat{I}_1, \hat{I}_2)$. Let $\hat{I} := (\hat{I}_1, \hat{I}_2)$ and $\hat{V} := (\hat{V}_1, \hat{V}_2)$. Then $\hat{I} = Y_{\text{uvn}}\hat{V}$ where Y_{uvn} is the Schur complement of $y_0 + y_1 + y_2$ (see Appendix A.3.1 for details of Schur complement):

$$Y_{\text{uvn}} := \begin{bmatrix} y_1 & 0 \\ 0 & y_2 \end{bmatrix} - \frac{1}{\sum_{i=0}^{2} y_i} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \begin{bmatrix} y_1 & y_2 \end{bmatrix} = \frac{1}{\sum_i y_i} \begin{bmatrix} y_1(y_0 + y_2) & -y_1 y_2 \\ -y_1 y_2 & y_2(y_0 + y_1) \end{bmatrix} \tag{3.11a}$$

Next connect the two ideal transformers to each side of the unitary voltage network; see Figure 3.8(b). Let $I := (I_1, -I_2)$ and $V := (V_1, V_2)$. The conversion between internal variables $(\hat{V}, \hat{I})$ and terminal variables (V, I) is $\hat{V} = MV$ and $\hat{I} = M^{-1}I$ where

$$M := \begin{bmatrix} 1/N_1 & 0 \\ 0 & 1/N_2 \end{bmatrix} \tag{3.11b}$$

Substituting into $\hat{J} = Y_{\text{uvn}}\hat{U}$ we obtain the relation between the terminal variables V to I:

$$I = (MY_{\text{uvn}}M)V \tag{3.11c}$$

where $MY_{\text{uvn}}M$ is called the admittance matrix of the transformer. It can be shown that (3.11) is equivalent to the T equivalent circuit (3.5) (Exercise 3.4). As a consequence the model parameters (y_0, y_1, y_2) cannot be uniquely determined by just the short-circuit and open-circuit tests.

We often do not know the numbers N_1, N_2 of turns of the primary and secondary windings respectively, but can determine the turns ratio $a := N_1/N_2$ from the specified rated voltages. The admittance matrix $MY_{\text{uvn}}M$ can also be written in terms of the turns ratio a (Exercise 3.5):

$$Y_{YY} := MY_{\text{uvn}}M = \frac{y_p y_s}{a^2 y_m + a^2 y_p + y_s} \begin{bmatrix} 1 + a^2 y_m / y_s & -a \\ -a & a^2(1 + y_m/y_p) \end{bmatrix} \tag{3.11d}$$

If $y_0 = y_m = 0$ then both (3.5) and (3.11) are equivalent to the simplified model in Figure 3.5(b). In this case the model parameter is just the leakage impedance z_l in the

significant for single-phase devices but turns out to be crucial in modeling three-phase devices; see Chapters 14 and 15.

primary circuit, which can be determined from standard power ratings as described above. Recall that $z_l = z_p + a^2 z_s$ and hence the leakage admittance in the simplified model is

$$y_l = \frac{1}{z_l} = \frac{1}{1/y_p + a^2 1/y_s} = \frac{y_p y_s}{a^2 y_p + y_s}$$

Indeed, when $y_m = 0$, the admittance matrix Y_{YY} is the same for both the simplified model and the unitary voltage network model, from (3.11d):

$$Y_{YY} = M Y_{\text{uvn}} M = y_l \begin{bmatrix} 1 & -a \\ -a & a^2 \end{bmatrix}$$

which is the same as (3.7b).

Multi-winding transformers.

The single-phase circuit model in Figure 3.8(b) can be generalized in two ways, or a combination. First, multiple copies of the single-phase model can be connected in Δ or Y configuration on each side to create models for three-phase transformers. This is derived in detail in Chapter 15.3 for unbalanced three-phase systems. Second, the unitary voltage network can be generalized to model nonstandard transformers with more than two windings. As an illustration we now use this approach to model a split-phase transformer.

Figure 3.10 shows a single-phase split-phase transformer. The internal voltages

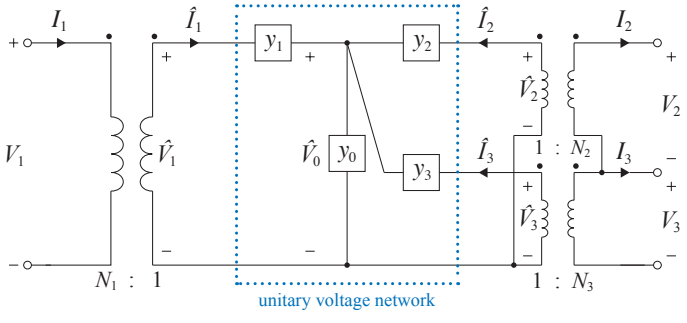


Figure 3.10 Single-phase split-phase transformer.

$(\hat{V}_0, \hat{V}_1, \hat{V}_2, \hat{V}_3)$ and currents $(\hat{I}_0, \hat{I}_1, \hat{I}_2, \hat{I}_3)$ on the unitary voltage network are defined in the figure. The admittance matrix that maps these voltages to currents is given by:

$$\begin{bmatrix} \hat{I}_0 \\ \hat{I}_1 \\ \hat{I}_2 \\ \hat{I}_3 \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{3} & -y_1 & -y_2 & -y_3 \\ -y_1 & y_1 & 0 & 0 \\ -y_2 & 0 & y_2 & 0 \\ -y_3 & 0 & 0 & y_3 \end{bmatrix} \begin{bmatrix} \hat{V}_0 \\ \hat{V}_1 \\ \hat{V}_2 \\ \hat{V}_3 \end{bmatrix}$$

Let $\hat{V} := (\hat{V}_1, \hat{V}_2, \hat{V}_3)$ and $\hat{I} := (\hat{I}_1, \hat{I}_2, \hat{I}_3)$. Since $\hat{I}_0 = 0$ we can eliminate $\hat{V}_0$ to relate $\hat{I} = Y_{\text{uvn}}\hat{V}$ where Y_{uvn} is the Kron-reduced admittance matrix:

$$\begin{aligned} Y_{\text{uvn}} &:= \begin{bmatrix} y_1 & 0 & 0 \\ 0 & y_2 & 0 \\ 0 & 0 & y_3 \end{bmatrix} - \frac{1}{\sum_{i=0}^{3} y_i} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} \\ &= \frac{1}{\sum_i y_i} \begin{bmatrix} y_1(y_0+y_2+y_3) & -y_1y_2 & -y_1y_3 \\ -y_2y_1 & y_2(y_0+y_1+y_3) & -y_2y_3 \\ -y_3y_1 & -y_3y_2 & y_3(y_0+y_1+y_2) \end{bmatrix} \end{aligned} \tag{3.12a}$$

This extends in a straightforward manner Y_{uvn} in (3.11) from two to three windings.

Next we connect ideal transformers to the unitary voltage network as shown in Figure 3.10. The terminal voltages $V := (V_1, V_2, V_3)$ and currents $I := (I_1, -I_2, -I_3)$, as well as the internal current $\hat{I}_3$ into the third winding, are defined in the figure. Let $M := \text{diag}(1/N_1, 1/N_2, 1/N_3)$. Then $\hat{V} = MV$ and, using $I_2 + I_3 + \hat{I}_3 = 0$,

$$\hat{I} \;=\; M^{-1} \begin{bmatrix} I_1 \\ -I_2 \\ \hat{I}_3 \end{bmatrix} \;=\; M^{-1} \begin{bmatrix} I_1 \\ -I_2 \\ -I_2 - I_3 \end{bmatrix} \;=:\; M^{-1} A I$$

where

$$A \;:=\; \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \tag{3.12b}$$

Substituting into $\hat{I} = Y_{\text{uvn}}\hat{V}$ we obtain the relation between the terminal variables V to I:

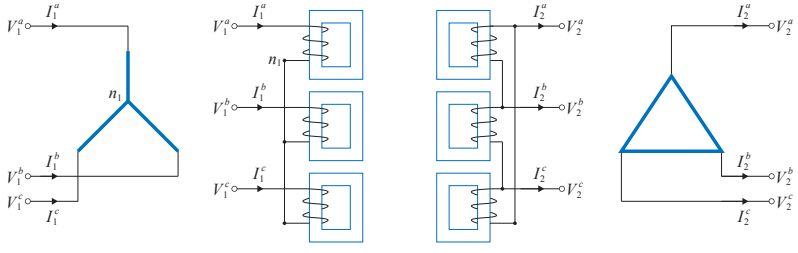
$$I \;=\; A^{-1}\,(M Y_{\text{uvn}} M)\, V \tag{3.12c}$$

3.2 Balanced three-phase transformers

In this section we develop models for a balanced three-phase transformer and derive its per-phase equivalent.

3.2.1 Ideal transformers

The primary and secondary circuits of a three-phase transformer can be arranged in four different configurations: YY, $\Delta\Delta$, ΔY, $Y\Delta$. Figure 3.11(a) shows a primary three-phase winding in Y configuration and its schematic diagram. The winding on the first magnetic core goes from terminal a to neutral n and then connects with the neutral terminals on the second and third magnetic cores. It matches the connectivity in the schematic diagram where the windings are indicated by the thick lines. Figure 3.11(b)



(a) Primary winding in Y configuration (b) Secondary winding in Δ configuration

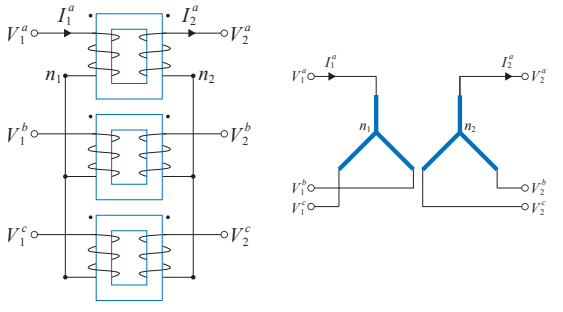
Figure 3.11 Primary and secondary windings in Y and Δ configurations respectively. The thick lines in the schematic diagrams represent transformer windings.

shows a secondary three-phase winding in Δ configuration and its schematic diagram. In both diagrams, the windings go from terminal a on the first magnetic core to terminal b on the second magnetic core to terminal c on the third magnetic core. The winding of an ideal three-phase transformer in YY configuration and its schematic digram are shown in Figure 3.12(a). The parallel lines in the schematic diagram indicate corresponding primary and secondary windings in the single-phase transformers. Similarly the winding of an ideal three-phase transformer in $\Delta\Delta$ configuration and its schematic digram are shown in Figure 3.12(b), and those for ΔY and $Y\Delta$ configurations are shown in Figure 3.13. The different configurations of three-phase transformer banks can also be represented compactly as in Figure 3.14 (see its caption for details).

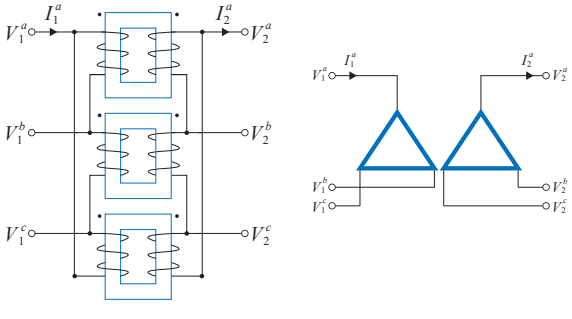
Recall that the internal voltages and currents are denoted by $V_j^Y := (V_j^{an}, V_j^{bn}, V_j^{cn}) \in \mathbb{C}^3$. $I_j^Y := (I_j^{an}, I_j^{bn}, I_j^{cn}) \in \mathbb{C}^3$ for Y configuration and $V_j^\Delta := (V_j^{ab}, V_j^{bc}, V_j^{ca}) \in \mathbb{C}^3$, $I_j^\Delta := (I_j^{ab}, I_j^{bc}, I_j^{ca}) \in \mathbb{C}^3$ for Δ configuration (see Figure 3.11). The terminal voltages and currents are denoted by $V_j := (V_j^a, V_j^b, V_j^c) \in \mathbb{C}^3$ and $I_j := (I_j^a, I_j^b, I_j^c) \in \mathbb{C}^3$, with the current I_1 flowing into the primary side of the transformer and I_2 flowing out of its secondary side. The external behavior of an ideal three-phase transformer is defined by the ratio of the line-to-line voltages on the secondary and the primary sides, and the ratio of the line currents on the secondary and the primary sides. We refer to these ratios as its *external model*. The phases of a balanced transformer are decoupled and therefore it can be represented by its phase a model, called its *per-phase equivalent*.

The external model of an ideal balanced three-phase transformer and its per-phase equivalent can be derived using the following procedure:

1 *Internal model*. Derive the internal voltage and current gains based on the pairing of primary and secondary windings in different configurations (see Figures 3.12



(a) YY configuration



(b) $\Delta\Delta$ configuration

Figure 3.12 Ideal three-phase transformers in YY and $\Delta\Delta$ configurations. The parallel lines in the schematic diagram indicate corresponding primary and secondary windings.

and 3.13):

$$YY: \quad V_2^Y = nV_1^Y, \quad -I_2^Y = aI_1^Y \tag{3.13a}$$

$$\Delta\Delta: \quad V_2^\Delta = nV_1^\Delta, \quad -I_2^\Delta = aI_1^\Delta \tag{3.13b}$$

$$\Delta Y: \quad V_2^Y = nV_1^\Delta, \quad -I_2^Y = aI_1^\Delta \tag{3.13c}$$

$$Y\Delta: \quad V_2^\Delta = nV_1^Y, \quad -I_2^\Delta = aI_1^Y \tag{3.13d}$$

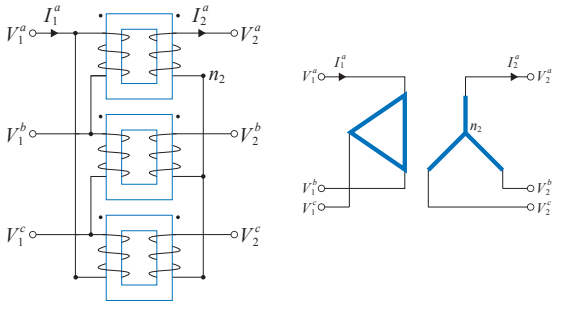
2 *Conversion rules*. Apply the conversion rules (1.13) (1.14) to express line-to-line voltages and line currents on both sides in terms of the internal voltages and currents respectively:

$$Y \text{ config:} \quad V_j^{\text{line}} = \Gamma V_j^Y = (1-\alpha)V_j^Y = \sqrt{3}e^{\mathbf{i}\pi/6}V_j^Y, \qquad I_j = \pm I_j^Y \tag{3.13e}$$

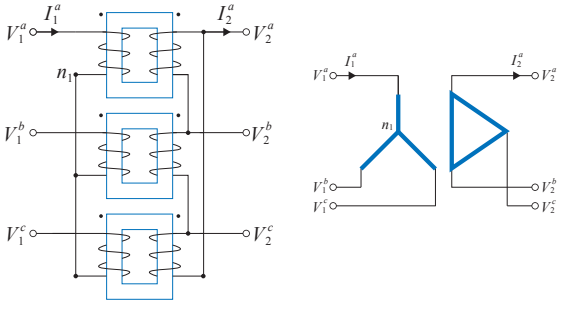
$$\Delta \text{ config:} \quad I_j = \pm \Gamma^{\mathsf{T}} I_j^\Delta = \pm (1-\alpha^2) I_j^\Delta = \pm \sqrt{3}e^{-\mathbf{i}\pi/6} I_j^\Delta, \quad V_j^{\text{line}} = V_j^\Delta \tag{3.13f}$$

where we have assumed the balanced voltages V_j^Y and currents I_j^Δ are in positive sequence, i.e., in $\text{span}(\alpha_+)$, and used Corollary 1.3.

3 *External model*. Derive the line-to-line voltage gains $K(n) \in \mathbb{C}$ and line current gains $1/\bar{K}(n) \in \mathbb{C}$ for the three-phase transformer by eliminating the internal



(a) ΔY configuration



(b) $Y\Delta$ configuration

Figure 3.13 Ideal three-phase transformers in ΔY and $Y\Delta$ configurations. The parallel lines in the schematic diagram indicate corresponding primary and secondary windings.

variables from the internal model in Step 1 and the conversion rule in Step 2:

$$V_2^{\text{line}} = K(n)V_1^{\text{line}}, \qquad I_2 = \frac{1}{\bar{K}(n)} I_1 \tag{3.13g}$$

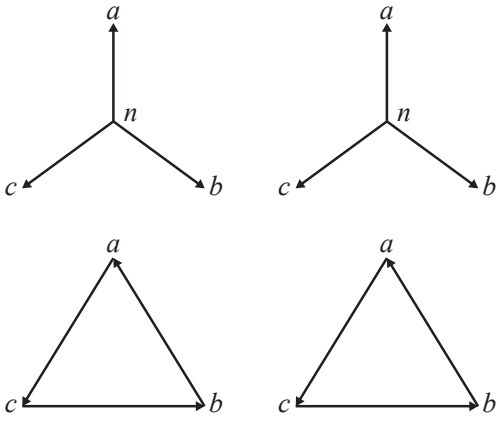
The fact that the voltage gain $K(n)$ is a scalar means that the phases of a balanced three-phase transformer are decoupled. The results for different configurations are given in Table 3.2 (see Example 3.3 for derivation).

Property	Gain
Voltage gain	$K(n)$
Current gain	$\frac{1}{\bar{K}(n)}$
Power gain	1
Sec z_l referred to pri	$\frac{z_l}{\lvert K(n)\rvert^2}$

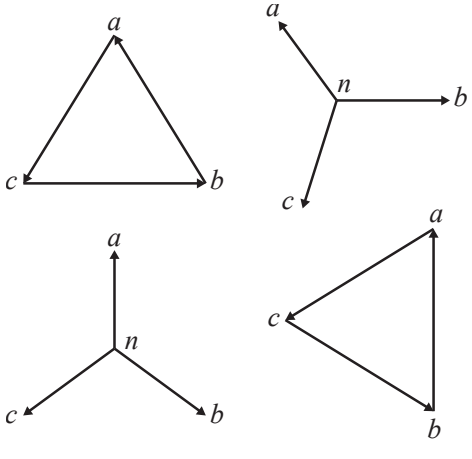
Configuration	Gain
YY	$K_{YY}(n) := n$
$\Delta\Delta$	$K_{\Delta\Delta}(n) := n$
ΔY	$K_{\Delta Y}(n) := \sqrt{3}n\, e^{\mathbf{i}\pi/6}$
$Y\Delta$	$K_{Y\Delta}(n) := \frac{n}{\sqrt{3}}\, e^{-\mathbf{i}\pi/6}$

Table 3.2 Ideal complex transformer properties.

4 *Per-phase equivalent*. The YY-equivalent of a balanced three-phase transformer is a balanced transformer in YY configuration that has the same external model, i.e., they have the same voltage gain $K(n)$ and current gain $1/\bar{K}(n)$ given in (3.13g).



(a) YY and $\Delta\Delta$ configurations



(b) ΔY and $Y\Delta$ configurations

Figure 3.14 Compact representation of ideal three-phase transformers in (a) YY, $\Delta\Delta$ configurations and (b) ΔY, $Y\Delta$ configurations. For instance, in the YY configuration, the vertical arrow represents the vector V^{an} in the complex plane. The arrow from b to a (not shown) represents the vector V^{ab}. The parallel lines in the diagram indicate corresponding primary and secondary windings.

Since the phases are decoupled, the per-phase equivalent is the phase a model of the YY-equivalent, i.e., a single-phase transformer with voltage gain $K(n)$. See Example 3.3.

Example 3.3 (External models and per-phase equivalents)**.** In this example we apply the method outlined above to derive the external models of ideal balanced three-phase transformers in YY, $\Delta\Delta$, ΔY and $Y\Delta$ configurations as well as their per-phase equivalents.

1 *YY configuration*. To derive the external model, eliminate the internal variables from (3.13a)–(3.13f):

$$\begin{aligned} V_2^{\text{line}} &= (1-\alpha)V_2^Y = (1-\alpha)nV_1^Y = nV_1^{\text{line}} \\ I_2 &= -I_2^Y = aI_1^Y = aI_1^Y \end{aligned}$$

giving the voltage gain $K_{YY}(n) := n$ and the current gain $1/\bar{K}_{YY}(n) := 1/n =: a$. The per-phase equivalent is simply an ideal single-phase transformer with voltage gain $K_{YY}(n) := n$.

2 $\Delta\Delta$ *configuration*. Similarly the external model is, from (3.13a)–(3.13f):

$$\begin{aligned} V_2^{\text{line}} &= V_2^\Delta = nV_1^\Delta = nV_1^{\text{line}} \\ I_2 &= -(1-\alpha^2)I_2^\Delta = (1-\alpha^2)aI_1^\Delta = aI_1 \end{aligned}$$

giving the same gains $K_{\Delta\Delta}(n) := n$ and $1/\bar{K}_{\Delta\Delta}(n) := a$ as those for the YY configuration. Hence the per-phase equivalent is also an ideal single-phase transformer with voltage gain $K_{\Delta\Delta} := n$.

3 ΔY *configuration*. The external model is, from (3.13a)–(3.13f):

$$\begin{aligned} V_2^{\text{line}} &= (1-\alpha)V_2^Y = (1-\alpha)nV_1^\Delta = (1-\alpha)nV_1^{\text{line}} \\ I_2 &= -I_2^Y = aI_1^\Delta = \frac{a}{1-\alpha^2}I_1 = \frac{a}{1-\bar{\alpha}}I_1 \end{aligned}$$

giving the voltage gain $K_{\Delta Y}(n) := (1-\alpha)n$ and current gain $1/\bar{K}_{\Delta Y}(n) := a(1-\alpha)^{-1}$. Hence the per-phase equivalent is an ideal single-phase transformer with voltage gain $K_{\Delta Y}(n) := (1-\alpha)n = \sqrt{3}e^{\mathbf{i}\pi/6}n$. The ΔY configuration has several advantages (e.g., a gain of $\sqrt{3}$ in addition to the gain n due to turns ratio) and is the most commonly adopted transformer in practice.

4 $Y\Delta$ *configuration*. The external model is, from (3.13a)–(3.13f):

$$\begin{aligned} V_2^{\text{line}} &= V_2^\Delta = nV_1^Y = \frac{n}{1-\alpha}V_1^{\text{line}} \\ I_2 &= -(1-\alpha^2)I_2^\Delta = (1-\alpha^2)aI_1^Y = (1-\alpha^2)aI_1 = (1-\bar{\alpha})aI_1 \end{aligned}$$

giving the voltage gain $K_{Y\Delta}(n) := n/(1-\alpha)$ and current gain $1/\bar{K}_{Y\Delta}(n) := (1-\bar{\alpha})a$. Hence the per-phase equivalent is an ideal single-phase transformer with voltage gain $K_{Y\Delta}(n) := n/(1-\alpha) = n/(\sqrt{3}e^{\mathbf{i}\pi/6})$. □

Hence the voltage gain $K(n)$ and the current gain $1/\bar{K}(n)$ given in Table 3.2 apply to line voltages/currents in both the original transformer and its YY equivalent. For Δ configuration on the primary or secondary side, its Y-equivalent in terms of the line voltage V_j^{line} and line current I_j can be derived from (3.13e)(3.13f) (also explained in (1.32a)). Specifically the Y-equivalent of (V_j^Δ, I_j^Δ) is

$$V_j^{Y\text{eq}} = \frac{1}{1-\alpha}V_j^\Delta = \frac{1}{\sqrt{3}\,e^{\mathbf{i}\pi/6}}V_j^\Delta, \qquad I_j^{Y\text{eq}} = \pm\left(1-\alpha^2\right)I_j^\Delta = \pm\frac{\sqrt{3}}{e^{\mathbf{i}\pi/6}}I_j^\Delta$$

Using the per-phase equivalent of an ideal balanced transformer (i.e., phase a model

of an equivalent transformer in YY configuration), we conclude that its complex power gain is 1:

$$\frac{-S_2}{S_1} := \frac{V_2^{an}(-\bar{I}_2^{an})}{V_1^{an}(\bar{I}_1^{an})} = K(n)\frac{1}{K(n)} = 1$$

It often simplifies per-phase analysis of a balanced system to refer series impedances and shunt admittances on one side to the other side of a transformer. This is explained in Chapter 3.3. In particular, a secondary series impedance z_l is referred to the primary as $z_l/|K(n)|^2$ according to (3.14) below. When terminated in a symmetric three-phase impedance load z_{load} on the secondary side so that $V_2^{an} = z_{\text{load}} I_2^{an}$ (using YY-equivalent), the per-phase driving-point impedance on the primary side is:

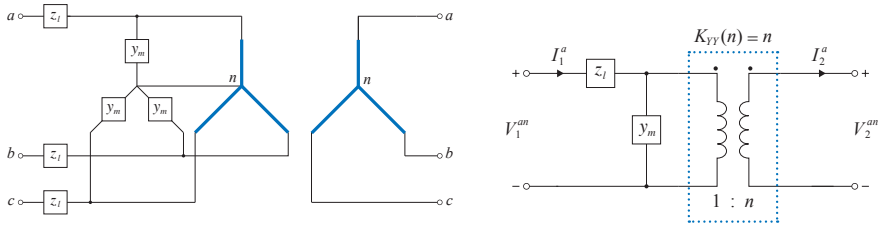
$$\frac{V_1^{an}}{I_1^{an}} = \frac{V_2^{an}/K(n)}{I_2^{an}\bar{K}(n)} = \frac{z_{\text{load}}}{|K(n)|^2}$$

These relations are also summarized in Table 3.2.

3.2.2 Nonideal transformers

In this section we first present circuit models of (nonideal) three-phase transformers and then their per-phase equivalent circuits after all Δ-configured transformers have been converted into their Y-equivalents. Each non-ideal single-phase transformer is modeled using the simplified model studied in Chapter 3.1.4.

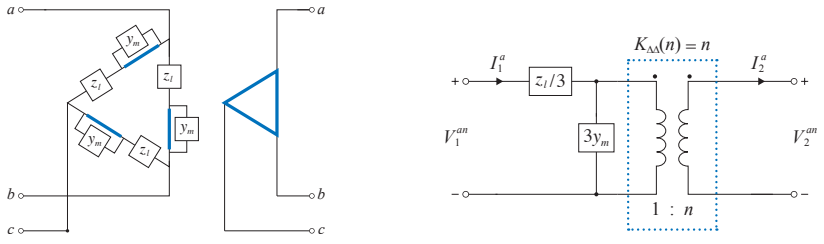
Per-phase equivalent circuits. Figure 3.15(a) shows a model of balanced three-phase (nonideal) transformers in YY configuration and Figure 3.15(b) shows its per-phase equivalent circuit. The per-phase circuit is identical to that in Figure 3.5(a). Figure



(a) YY configuration (b) Per-phase circuit

Figure 3.15 Model of three-phase transformers in YY configuration and its per-phase equivalent circuit.

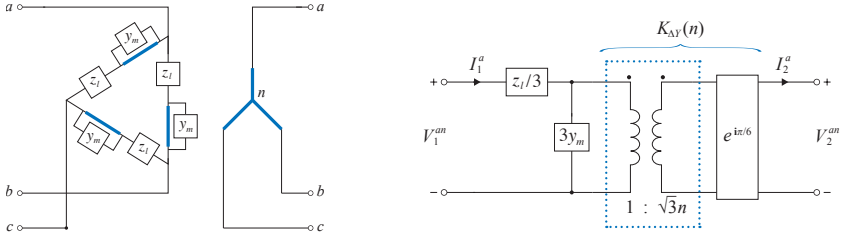
3.16(a) shows a model of balanced three-phase transformers in $\Delta\Delta$ configuration. Its YY equivalent and per-phase circuit are identical to those in Figure 3.15 except that the equivalent leakage impedance $z_l/3$ is one-third of the value in the original $\Delta\Delta$ circuit and the shunt admittance $3y_m$ is three times the value in the original $\Delta\Delta$



(a) ΔΔ configuration (b) Per-phase circuit

Figure 3.16 Model of three-phase transformers in ΔΔ configuration and its per-phase equivalent circuit.

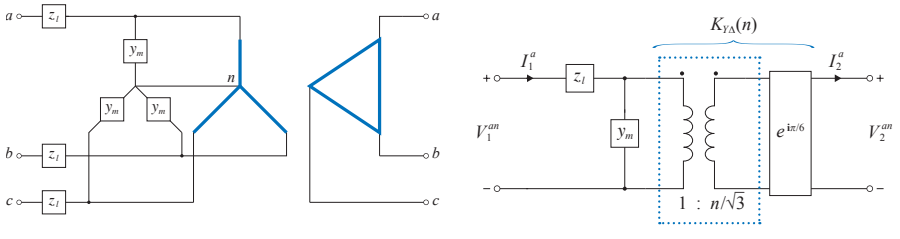
circuit. This can be verified by checking the secondary open-circuit equivalent and the secondary short-circuit equivalent of the original ΔΔ circuit. Figure 3.17 shows a model of balanced three-phase transformers in ΔY configuration and its per-phase equivalent circuit. Finally Figure 3.18 shows the model for $Y\Delta$ configuration and its



(a) ΔY configuration (b) Per-phase circuit

Figure 3.17 Model of three-phase transformers in ΔY configuration and its per-phase equivalent circuit.

per-phase circuit.



(a) $Y\Delta$ configuration (b) Per-phase circuit

Figure 3.18 Model of three-phase transformers in $Y\Delta$ configuration and its per-phase equivalent circuit.

Hence balanced three-phase transformers in YY, ΔΔ, ΔY and $Y\Delta$ configurations all

have the same per-phase equivalent circuit, with appropriate values for their leakage impedance and shunt admittance and the corresponding (complex) transformer gains $K(n)$.

3.3 Equivalent impedance in transformer circuit

In this subsection we explain how to derive an "equivalent" impedance when looking into the terminal, either on the primary side or on the secondary side of a transformer. Consider the singe-phase equivalent circuit of a balanced three-phase transformer. A series impedance z_s in the secondary circuit of the transformer can be equivalently replaced by a series impedance z_p in the primary circuit, and vice versa, provided they are related by:

$$z_p = \frac{z_s}{|K(n)|^2} \quad \text{or equivalently} \quad z_s = |K(n)|^2 z_p \tag{3.14a}$$

The first operation in (3.14a) is called *referring z_s in the secondary to the primary*. The second operation is called *referring z_p in the primary to the secondary*. A shunt admittance y_s in the secondary circuit of the transformer can be equivalently replaced by a shunt admittance y_p in the primary circuit, and vice versa, provided they are related by:

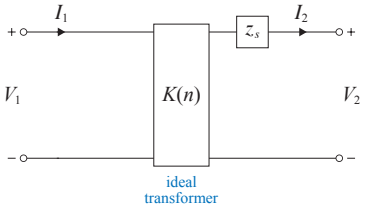
$$y_p = |K(n)|^2 y_s \quad \text{or equivalently} \quad y_s = \frac{y_p}{|K(n)|^2} \tag{3.14b}$$

These operations will be used as a shortcut in the analysis of circuits that contain transformers the same way we use the Thévenin equivalent of impedances in series or in parallel; see Chapter 3.4.

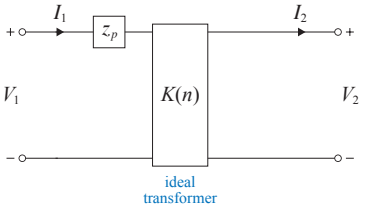
Here "equivalence" means that the external behavior remains unchanged when a series impedance or a shunt admittance on one side is referred to the other. Specifically we consider two kinds of external behavior. In the first case, explained in Chapter 3.3.1, the external behavior is the transmission matrix that maps (V_2, I_2) to (V_1, I_1). In the second case, explained in Chapter 3.3.2, the external behavior is the driving-point impedance on one side of the transformer when the other side is connected to an impedance. We next derive (3.14) as a simple consequence of Kirchhoff's and Ohm's laws.

3.3.1 Transmission matrix

Consider the per-phase transformer circuits in Figure 3.19 of a balanced three-phase system, one with a series impedance in the secondary circuit and the other in the primary circuit. Let T_s and T_p denote the transmission matrices that maps (V_2, I_2) to (V_1, I_1) in Figure 3.19(a) and Figure 3.19(b) respectively. We claim that the relation



(a) Series impedance z_s in the secondary circuit.



(b) Series impedance z_p in the primary circuit.

Figure 3.19 Referring series impedance in the secondary to the primary.

(3.14a) between series impedances z_p and z_s ensures that $T_s = T_p$. It is in this sense that we say these two circuits are equivalent.

To show that $T_s = T_p$ let (V, I) denote the voltage and current at the secondary terminal of the ideal transformer in Figure 3.19(a). Then $V = V_2 + z_s I$ and $I = I_2$, or

$$\begin{bmatrix} V \\ I \end{bmatrix} = \begin{bmatrix} 1 & z_s \\ 0 & 1 \end{bmatrix} \begin{bmatrix} V_2 \\ I_2 \end{bmatrix}$$

Hence

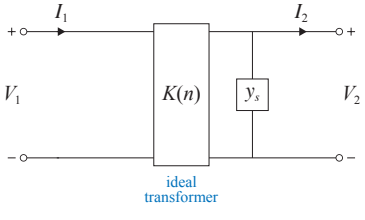
$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = \begin{bmatrix} K^{-1}(n) & 0 \\ 0 & \bar{K}(n) \end{bmatrix} \begin{bmatrix} 1 & z_s \\ 0 & 1 \end{bmatrix} \begin{bmatrix} V_2 \\ I_2 \end{bmatrix} = \underbrace{\begin{bmatrix} K^{-1}(n) & K^{-1}(n) z_s \\ 0 & \bar{K}(n) \end{bmatrix}}_{T_s} \begin{bmatrix} V_2 \\ I_2 \end{bmatrix}$$

Similarly, for the circuit in Figure 3.19(b), we have

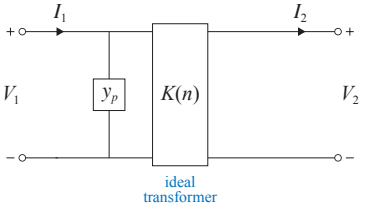
$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = \begin{bmatrix} 1 & z_p \\ 0 & 1 \end{bmatrix} \begin{bmatrix} K^{-1}(n) & 0 \\ 0 & \bar{K}(n) \end{bmatrix} \begin{bmatrix} V_2 \\ I_2 \end{bmatrix} = \underbrace{\begin{bmatrix} K^{-1}(n) & \bar{K}(n) z_p \\ 0 & \bar{K}(n) \end{bmatrix}}_{T_p} \begin{bmatrix} V_2 \\ I_2 \end{bmatrix}$$

Hence $T_s = T_p$ if and only if (3.14a) holds.

The relation (3.14b) between shunt admittances y_p and y_s ensures that the transmission matrix for the circuit in Figure 3.20(a) is the same as that in Figure 3.20(b). This



(a) Shunt admittance y_s in the secondary circuit.

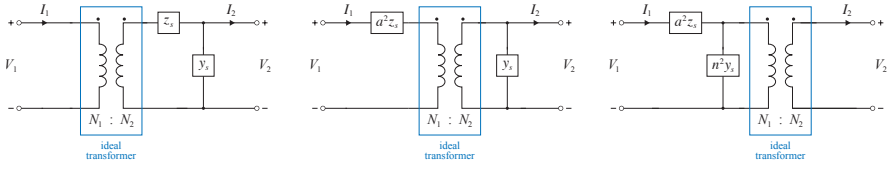


(b) Shunt admittance y_p in the primary circuit.

Figure 3.20 Referring shunt admittance in the secondary to the primary.

is left as Exercise 3.8. The operations in (3.14) can be repeatedly applied to a circuit involving multiple impedances and admittances, as illustrated in the next example.

Example 3.4. A combination of a series impedance z_s and a shunt admittance y_s in the secondary circuit, as shown in Figure 3.21(a), can be referred to the primary one element at a time, starting from the element that is *closest to* the ideal transformer. The



(a) (z_s, y_s) in the secondary circuit. (b) Refer z_s to the primary. (c) Refer y_s to the primary.

Figure 3.21 Referring (z_s, y_s) in the secondary to the primary.

transformer gain is $K(n) = n = 1/a := N_2/N_1$. Referring the series impedance z_s to the primary yields the equivalent circuit in Figure 3.21(b) with an equivalent primary impedance $a^2 z_s$. Referring then the shunt admittance y_s to the primary yields the equivalent circuit in Figure 3.21(c) with an equivalent shunt admittance $n^2 y_s$. □

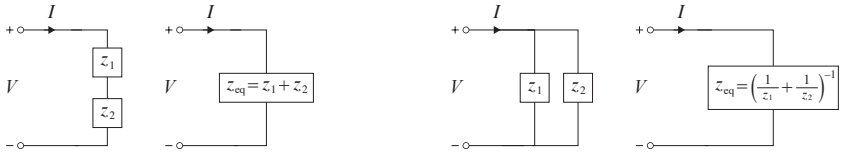
3.3.2 Driving-point impedance

In the second case the external behavior is the driving point impedances on one side of the transformer when the other side is connected to an impedance. In general suppose we apply a voltage V across two terminals that are connected to a network of impedances and transformers. Suppose a current I flows between these two terminals through the network. The ratio V/I is called the driving-point impedance at these terminals. For networks consisting of a cascade of impedances in series and in parallel, the driving-point impedance is also called the Thévenin equivalent impedance. The Thévenin equivalent impedance of such a network can be derived by repeatedly applying simple reduction rules for the two basic configurations shown in Figure 3.22. For two impedances z_1, z_2 in series depicted in Figure 3.22(a), the Thévenin equivalent impedance z_{eq} is defined such that the two networks in Figure 3.22(a) have the same driving-point impedance:

$$\frac{V}{I} = z_1 + z_2 \quad =: \quad z_{\text{eq}} \tag{3.15a}$$

Similarly the Thévenin equivalent impedance of two impedances in parallel depicted in Figure 3.22(b) is defined to be:

$$\frac{V}{I} = \left(\frac{1}{z_1} + \frac{1}{z_2}\right)^{-1} \quad =: \quad z_{\text{eq}} \tag{3.15b}$$



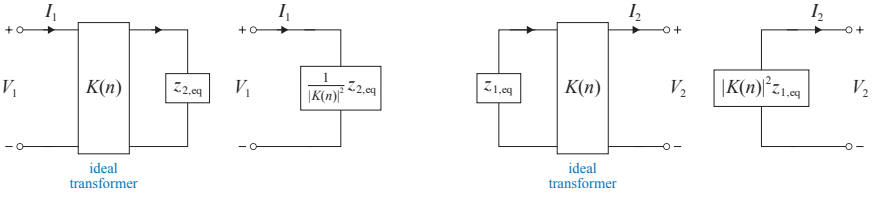
(a) Impedances in series (b) Impedances in parallel

Figure 3.22 (a) Thévenin equivalent z_{eq} of two impedances z_1, z_2 in series. (b) Thévenin equivalent z_{eq} of two impedances z_1, z_2 in parallel.

These are simple consequences of Kirchhoff's and Ohm's laws. Repeated application of (3.15) reduces a cascade of impedances in parallel and series into a single equivalent impedance that preserves the driving-point impedance.

When such a network contains not just impedances, but also transformers, the relation (3.14) allows us to reduce it to a single Thévenin equivalent impedance with the same driving-point impedance. As we explain below, the key element of this procedure is the driving-point impedance seen from two terminals of one side of a single-phase transformer when the other side is connected to an impedance z_{eq} that may be the Thévenin equivalent of a network of impedances. This yields an equivalent network where the transformer and z_{eq} is replaced by a scaled impedance and the number of transformer is reduced by 1. Repeated application of (3.14) and (3.15) can then be used to remove all transformers from the equivalent network, allowing the derivation of the Thévenin equivalent impedance of the original network. When applicable, this technique greatly simplifies per-phase analysis of a balanced system as we will see in Chapter 3.4.

We now explain the key building block of this procedure. When the secondary side of an ideal transformer is connected to an impedance $z_{2,eq}$ as shown in Figure 3.23(a), the transformer and the impedance $z_{2,eq}$ can be replaced by the Thévenin equivalent impedance $z_{2,eq}/|K(n)|^2$ in the sense that the driving-point impedance V_1/I_1 on the primary side is the same in both circuits in Figure 3.23(a). This is the same operation that refers $z_{2,eq}$ in the secondary to the primary expressed in (3.14a). It is a consequence of the Kirchhoff's and Ohm's laws and is derived in Exercise 3.10. Similarly when



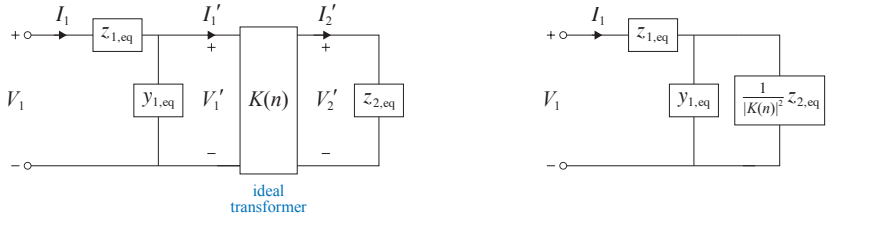
(a) V_1/I_1 on the primary side (b) V_2/I_2 on the secondary side

Figure 3.23 Driving-point impedances

the primary side is connected to an impedance $z_{1,\text{eq}}$ as shown in Figure 3.23(b), the transformer and the impedance $z_{1,\text{eq}}$ can be replaced by the Thévenin equivalent impedance $|K(n)|^2\, z_{1,\text{eq}}$ in the sense that the driving-point impedance V_2/I_2 on the secondary side is the same in both circuits in Figure 3.23(b). This is the same operation that refers $z_{1,\text{eq}}$ in the primary to the secondary expressed in (3.14a) (Exercise 3.10).

We caution that the shortcut (3.14) and (3.15) are not always applicable. For example they may not be applied to a circuit that contains parallel paths; see Example 3.8 in Chapter 3.4.2. In that case we analyze the circuit using Kirchhoff's and Ohm's laws. The shortcut is usually applicable to a radial system that does not contain parallel paths. We now illustrate its application in the derivation of the driving-point impedances on the primary and the secondary side.

Example 3.5 (V_1/I_1 on the primary side.)**.** Consider the network in Figure 3.24(a) where the secondary side is connected to a network whose Thévenin equivalent is $z_{2,\text{eq}}$. What is the driving-point impedance V_1/I_1? We first derive the driving-point



(a) Transformer circuit (b) Equivalent circuit seen on the primary side

Figure 3.24 Driving-point impedance V_1/I_1 on the primary side.

impedance directly using Kirchhoff's and Ohm's laws. We then use the result to verify the shortcut expressed in (3.14) and (3.15).

Circuit analysis. We have for the primary circuit

$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = \begin{bmatrix} 1 + z_{1,\text{eq}}\, y_{1,\text{eq}} & z_{1,\text{eq}} \\ y_{1,\text{eq}} & 1 \end{bmatrix} \begin{bmatrix} V_1' \\ I_1' \end{bmatrix}$$

Hence

$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = \begin{bmatrix} 1 + z_{1,\text{eq}}\, y_{1,\text{eq}} & z_{1,\text{eq}} \\ y_{1,\text{eq}} & 1 \end{bmatrix} \begin{bmatrix} K^{-1}(n) & 0 \\ 0 & \bar{K}(n) \end{bmatrix} \begin{bmatrix} V_2' \\ I_2' \end{bmatrix}$$

Substituting $V_2' = z_{2,\text{eq}}\, I_2'$ we have

$$\begin{aligned} \begin{bmatrix} V_1 \\ I_1 \end{bmatrix} &= \begin{bmatrix} 1 + z_{1,\text{eq}}\, y_{1,\text{eq}} & z_{1,\text{eq}} \\ y_{1,\text{eq}} & 1 \end{bmatrix} \begin{bmatrix} |K(n)|^{-2} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} z_{2,\text{eq}} \\ 1 \end{bmatrix} \bar{K}(n)\, I_2' \\ &= \begin{bmatrix} 1 + z_{1,\text{eq}}\, y_{1,\text{eq}} & z_{1,\text{eq}} \\ y_{1,\text{eq}} & 1 \end{bmatrix} \begin{bmatrix} z_{2,\text{eq}}/|K(n)|^2 \\ 1 \end{bmatrix} \bar{K}(n)\, I_2' \end{aligned}$$

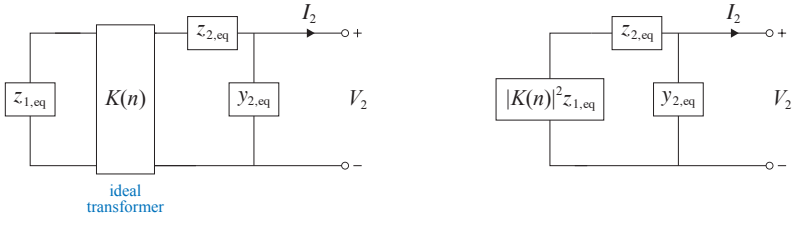
Hence the driving-point impedance is

$$\frac{V_1}{I_1} = \frac{(1+z_{1,\mathrm{eq}}\,y_{1,\mathrm{eq}})\,(z_{2,\mathrm{eq}}/|K(n)|^2)+z_{1,\mathrm{eq}}}{y_{1,\mathrm{eq}}\,(z_{2,\mathrm{eq}}/|K(n)|^2)+1} = z_{1,\mathrm{eq}} + \left(y_{1,\mathrm{eq}} + \frac{1}{z_{2,\mathrm{eq}}/|K(n)|^2}\right)^{-1} \tag{3.16}$$

It is the Thévenin equivalent on the primary side of a network consisting of impedances, admittances, as well as an ideal transformer. The Thévenin equivalent (3.16) has a simple interpretation, as we now explain.

Shortcut.. Use (3.14a) to refer $z_{2,\mathrm{eq}}$ in the secondary to the primary, we can replace the ideal transformer and $z_{2,\mathrm{eq}}$ by the equivalent impedance $z_{2,\mathrm{eq}}/|K(n)|^2$ and arrive at the equivalent circuit in Figure 3.24(b) seen from the primary side. The application of (3.15) then yields the driving-point impedance (3.16). □

Example 3.6 (V_2/I_2 on the secondary side.)**.** Consider the circuit in Figure 3.25(a) where the primary side is connected to the impedance $z_{1,\mathrm{eq}}$. Use (3.14a) to refer $z_{1,\mathrm{eq}}$



(a) Transformer circuit

(b) Equivalent circuit seen on the secondary side

Figure 3.25 Driving-point impedance V_2/I_2 on the secondary side.

in the primary to the secondary, we can replace the ideal transformer and $z_{1,\mathrm{eq}}$ by the equivalent impedance $|K(n)|^2\, z_{1,\mathrm{eq}}$ and arrive at the equivalent circuit in Figure 3.25(b) seen from the secondary side. The application of (3.15) then yields the driving-point impedance:

$$\frac{V_2}{I_2} = \left(y_{2,\mathrm{eq}} + \frac{1}{z_{2,\mathrm{eq}} + |K(n)|^2 \cdot z_{1,\mathrm{eq}}}\right)^{-1} \tag{3.17}$$

□

3.4 Per-phase analysis

In this section we apply the techniques developed in the previous sections in the analysis of a balanced three-phase power system consisting of generators, transformers, transmission lines, and loads, in a mix of Y and Δ configurations. We first explain how

to obtain a per-phase equivalent circuit of the system and then illustrate, through an example, the per-phase analysis using the shortcut (3.14) and (3.15). Finally we discuss a circuit that contains parallel paths to which the shortcut is not applicable. We explain why the end-to-end complex transformer gains on these paths should be equal.

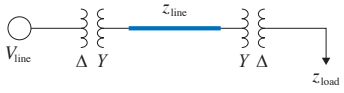
3.4.1 Analysis procedure

We have explained in Chapter 1.2.5 how to convert all sources, series impedances, shunt admittances in Δ configurations into their equivalent Y configurations and obtain a per-phase equivalent circuit. Chapter 3.2.1 shows that an ideal balanced three-phase transformer has a per-phase equivalent model specified by a complex voltage gain $K(n)$ that relates the voltages and currents on two sides of the transformer. Chapter 3.2.2 shows how to incorporate the transformer series impedance and shunt admittance into the per-phase model for both Y and Δ configurations. Chapter 3.3.1 explains how to refer series impedances and shunt admittances on one side to the other and Chapter 3.3.2 explains how to use this shortcut to simplify circuit analysis the same way we use Thévenin equivalent of impedances in series or in parallel. Putting everything together the procedure for per-phase analysis of a balanced three-phase system is as follows:

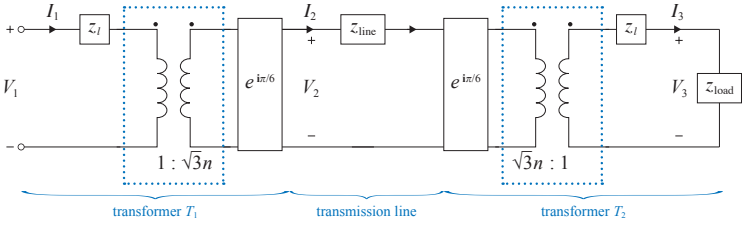
1 Convert all sources and loads in Δ configuration into their Y equivalents using (1.32a) for sources and (1.32b) for loads.
2 Convert all ideal transformers in Δ configuration into their Y equivalents with voltage gains $K(n)$ given in Table 3.2.
3 Obtain the phase a equivalent circuit by connecting all neutrals.
4 Solve for the desired phase a variables. Use Thévenin equivalent of series impedances and shunt admittances in a network containing transformers to simplify the analysis when applicable, e.g., for a radial system.
5 Obtain variables for phases b and c by subtracting (or adding) 120° and 240° from the phase a variables for positive-sequence (negative-sequence) sources. If variables in the internal of the Δ configurations are desired, derive them from the original circuits.

We illustrate this procedure in the next example.

Example 3.7. Consider the balanced system described by the one-line diagram in Figure 3.26(a) where a three-phase generator is connected to a stepup three-phase transformer bank (primary on the left) in ΔY configuration, which is connected through a three-phase transmission line to a stepdown transformer bank (primary on the right) in ΔY configuration, and then to a load. The terminal line voltage of the generator is V_{line}. The transmission line is modeled by a series impedance z_{line} and the load is assumed to be an impedance z_{load}. The transformer banks are made up of identical single-phase transformers each specified by a series impedance of $3z_l$ and a turns ratio of $a := 1/n$.



(a) One-line diagram



(b) Per-phase circuit

Figure 3.26 Example 3.7.

Find the generator current, the transmission line current, the load current, the load voltage, and the complex power delivered to the load in terms of the given parameters.

Solution. The per-phase equivalent circuit is shown in Figure 3.26(b). Note that the stepdown ΔY transformer near the load has its primary side on the right and secondary side on the left so that, going from left to right, the voltage (current) angle is shifted *down (down)* by 30° and their magnitudes scaled *down (up)* by $\sqrt{3}n$; see Exercise 3.6. The primary sides of both the stepup and stepdown transformers have been converted from Δ to its Y equivalent, with an equivalent series impedance z_l that is $1/3$ of the original impedance $3z_l$. The phase voltage of the generator in the per-phase equivalent circuit is

$$V_1 := \frac{V_{\text{line}}}{\sqrt{3}\, e^{\mathbf{i}\pi/6}}$$

Our solution strategy is as follows. We will use (3.14) and (3.15) to refer all the (load, transformer, and transmission line) impedances to the primary side of the stepup transformer. This calculates the driving-point impedance seen at the generator. Given generator phase voltage V_1, we can derive the generator current I_1. We then propagate this towards the load to calculate the other quantities.

Let $K(n) := \sqrt{3}n\, e^{\mathbf{i}\pi/6}$. Going from right to left, we cross the stepdown transformer T_2 from the primary to the secondary. Referring the impedance $z_{1,\text{eq}} := z_{\text{load}} + z_l$ on the primary to the secondary (see Figure 3.23(b)), the equivalent impedance at the right-end of the transmission line is

$$|K(n)|^2 \, (z_{\text{load}} + z_l)$$

Hence the equivalent impedance at the secondary side of the stepup transformer T_1 is

$$z_{2,\text{eq}} := z_{\text{line}} + |K(n)|^2 \, (z_{\text{load}} + z_l)$$

Referring this impedance to the primary side of T_1 (see Figure 3.23(a)), the driving point impedance at the generator is:

$$\begin{aligned}\frac{V_1}{I_1} &= z_l + \frac{1}{|K(n)|^2}\cdot\left(z_{\text{line}} + |K(n)|^2\,(z_{\text{load}} + z_l)\right)\\ &= 2z_l + \frac{z_{\text{line}}}{|K(n)|^2} + z_{\text{load}}\end{aligned}$$

Hence the primary side of T_1 sees the series impedance z_l of the two transformers, a scaled down version of the line impedance z_{line}, and the load z_{load}, all in series. Note that, seen from the generator, the load z_{load} goes through a stepdown transformer and a stepup transformer and therefore the scaling effects of these two transformers are canceled out.

Given the bus voltage V_1 of the generator, the generator current is then

$$I_1 = \frac{V_1}{2z_l + \frac{z_{\text{line}}}{|K(n)|^2} + z_{\text{load}}}$$

The transmission line current is

$$I_2 = \frac{I_1}{\bar{K}(n)} \;=\; \frac{V_1}{\bar{K}(n)\left(2z_l + \frac{z_{\text{line}}}{|K(n)|^2} + z_{\text{load}}\right)}$$

The load current is

$$I_3 = \bar{K}(n)\,I_2 \;=\; I_1$$

i.e., the effects of stepup and stepdown transformers cancel each other and the load current is equal to the generator current. The load voltage is

$$V_3 = z_{\text{load}}\,I_3 \;=\; z_{\text{load}}\,I_1 \;=\; V_1\cdot\frac{z_{\text{load}}}{2z_l + \frac{z_{\text{line}}}{|K(n)|^2} + z_{\text{load}}}$$

Hence V_3 relates to V_1 according to the voltage-divider rule where V_1 is the voltage drop across the series of impedances $2z_l + \frac{z_{\text{line}}}{|K(n)|^2} + z_{\text{load}}$ and V_3 is the voltage drop across z_{load}. The complex power delivered to the load is

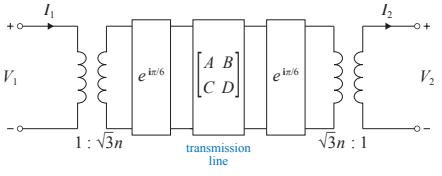
$$V_3\,\bar{I}_3 = z_{\text{load}}\cdot\left|\frac{V_1}{2z_l + \frac{z_{\text{line}}}{|K(n)|^2} + z_{\text{load}}}\right|^2 \;=\; z_{\text{load}}\cdot\frac{|V_{\text{line}}|^2}{3\left|2z_l + \frac{z_{\text{line}}}{|K(n)|^2} + z_{\text{load}}\right|^2}$$

□

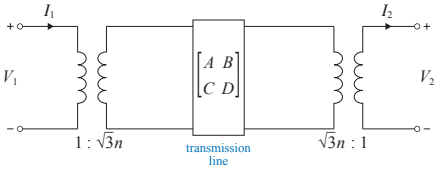
Simplified per-phase diagram for external behavior.

In Example 3.7, only the transmission line current I_2 that is in between the pair of transformers depends on the *connection-induced phase shift* $e^{\mathbf{i}\pi/6}$ in the complex transformer gain $K(n)$. Outside the pair of transformers, the driving point impedance V_1/I_1, the generator current I_1, the load current I_3, the load voltage V_3, and the power delivered to the load do not. They depend only on $|K(n)|^2$. This is the case even if we use the more detailed Π model of the transmission line instead of the short-line model

used here. Indeed, suppose the series impedance z_{line} in Figure 3.26(b) is replaced by the transmission matrix in (2.9) or (2.13)(2.14) as in Figure 3.27(a). Then the voltage



(a) Transmission line Π-model



(b) Equivalent circuit without connection-induced phase shift

Figure 3.27 Π-model of transmission line in place of the series impedance z_{line} model in Figure 3.26(b).

and current (V_1, I_1) on the left is related to the voltage and current (V_2, I_2) by

$$\begin{bmatrix} V_1 |K(n)| e^{\mathbf{i}\pi/6} \\ I_1 |K(n)|^{-1} e^{\mathbf{i}\pi/6} \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \cdot \begin{bmatrix} V_2 |K(n)| e^{\mathbf{i}\pi/6} \\ I_2 |K(n)|^{-1} e^{\mathbf{i}\pi/6} \end{bmatrix}$$

$$\begin{bmatrix} V_1 |K(n)| \\ I_1 |K(n)|^{-1} \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \cdot \begin{bmatrix} V_2 |K(n)| \\ I_2 |K(n)|^{-1} \end{bmatrix}$$

Therefore the external behavior is as if the connection-induced phase shift $e^{\mathbf{i}\pi/6}$ is absent, as shown in Figure 3.27(b). This motivates a simplified per-phase diagram for external behavior that ignores all the connection-induced phase shifts of transformers as long as every path contains stepup and stepdown transforms in pairs and wired in opposite directions. This is generally true for radial networks in practice where no transmission lines nor transformers are in parallel. Radial networks are a special case of a normal system that we discuss next.

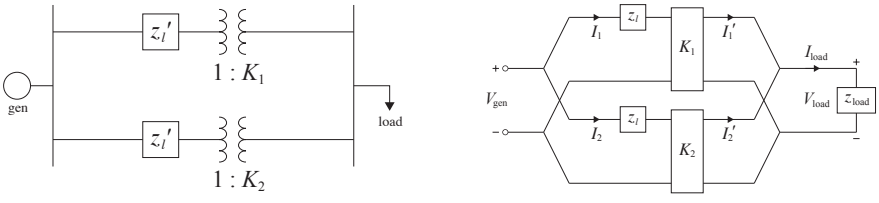
3.4.2 Normal system

A system is called *normal* if, in the per-phase equivalent circuit, the product of the *complex ideal* transformer gains around every loop is 1. Equivalently, on each parallel path,

1 the product of ideal transformer gain magnitudes is the same, and
2 the sum of ideal transformer phase shifts is the same.

Normal systems have a normalization that greatly simplifies analysis which we will discuss in Chapter 3.5. The following example motivates such a system.

Example 3.8 (Loop flows)**.** Consider a generator and a load connected by two three-phase transformer banks in parallel forming a loop as shown in Figure 3.28(a). The



(a) Transmission line Π-model (b) Equivalent circuit

Figure 3.28 Two buses connected in a loop with two parallel transformers.

transformer in the upper path is characterized by a series impedance and a complex gain K_1. The transformer in the lower path is characterized by the same series impedance and a possibly different complex gain K_2. Suppose line-to-neutral voltage of the generator bus is V_{gen}, the series impedance z_l of the transformer and the load impedance z_{load} in the per-phase equivalent circuit are given, as shown in Figure 3.28(b). Derive the currents $I_{\text{load}}, I_1', I_2'$ in terms of $V_{\text{gen}}, z_l, z_{\text{load}}$. Discuss the implications when

1 $K_2 = K_1$. This is the case if both transformer banks are YY-configured.
2 $K_2 = K_1\, e^{\mathbf{i}\theta}$. This is the case if the upper transformer bank is YY-configured with a voltage gain of n but the lower transformer bank is ΔY-configured with a voltage gain of $n/\sqrt{3}$ and $\theta = \pi/6$.
3 $K_2 = k \cdot K_1$, $k > 0$. This is the case if both transformer banks are YY-configured but with different turns ratios.

Solution. We cannot directly apply the shortcut (3.14) and (3.15) to refer the impedances z_{load} and z_l to the primary side because of the parallel paths, and must analyze the per-phase circuit using Kirchhoff's and Ohm's laws.

We have five unknowns currents $I_{\text{load}}, I_1', I_2', I_1, I_2$. The five equations that relate them are

$$\begin{aligned} I_{\text{load}} &= I_1' + I_2' \\ z_{\text{load}}\, I_{\text{load}} &= K_1 \cdot \left(V_{\text{gen}} - z_l\, I_1\right) \\ z_{\text{load}}\, I_{\text{load}} &= K_2 \cdot \left(V_{\text{gen}} - z_l\, I_2\right) \\ I_j' &= \frac{I_j}{\bar{K}_j}, \qquad j = 1,2 \end{aligned}$$

where the first equation expresses KCL, the second and third equations express the

load voltage seen on the upper and lower paths, respectively, and follow from the transformer equation and KVL, and the last equations express current gains of the transformers. Eliminating $I_{\text{load}}, I_1', I_2'$ we have

$$z_{\text{load}}\left(\frac{I_1}{\bar{K}_1} + \frac{I_2}{\bar{K}_2}\right) = K_1 \cdot \left(V_{\text{gen}} - z_l I_1\right)$$
$$z_{\text{load}}\left(\frac{I_1}{\bar{K}_1} + \frac{I_2}{\bar{K}_2}\right) = K_2 \cdot \left(V_{\text{gen}} - z_l I_2\right)$$

or

$$\begin{bmatrix} z_l + z_{\text{load}}|K_1|^{-2} & z_{\text{load}}(K_1\bar{K}_2)^{-1} \\ z_{\text{load}}(\bar{K}_1 K_2)^{-1} & z_l + z_{\text{load}}|K_1|^{-2} \end{bmatrix} \cdot \begin{bmatrix} I_1 \\ I_2 \end{bmatrix} = \begin{bmatrix} V_{\text{gen}} \\ V_{\text{gen}} \end{bmatrix}$$

Inverting the matrix, we obtain

$$I_1 = \frac{V_{\text{gen}}}{z_l + z_{\text{load}}\left(|K_1|^{-2} + |K_2|^{-2}\right)} \cdot \alpha_1$$
$$I_2 = \frac{V_{\text{gen}}}{z_l + z_{\text{load}}\left(|K_1|^{-2} + |K_2|^{-2}\right)} \cdot \alpha_2$$

where

$$\alpha_1 = 1 + \frac{z_{\text{load}}}{z_l} \cdot \frac{K_1 - K_2}{K_1 |K_2|^2}$$
$$\alpha_2 = 1 + \frac{z_{\text{load}}}{z_l} \cdot \frac{K_2 - K_1}{|K_1|^2 K_2}$$

Hence

$$I_1' = \frac{I_1}{\bar{K}_1} = \frac{V_{\text{gen}}}{z_l + z_{\text{load}}\left(|K_1|^{-2} + |K_2|^{-2}\right)} \cdot \frac{\alpha_1}{\bar{K}_1}$$
$$I_2' = \frac{I_2}{\bar{K}_2} = \frac{V_{\text{gen}}}{z_l + z_{\text{load}}\left(|K_1|^{-2} + |K_2|^{-2}\right)} \cdot \frac{\alpha_2}{\bar{K}_2}$$

and

$$I_{\text{load}} = I_1' + I_2' = \frac{V_{\text{gen}}}{z_l + z_{\text{load}}\left(|K_1|^{-2} + |K_2|^{-2}\right)} \cdot \left(\frac{1}{\bar{K}_1} + \frac{1}{\bar{K}_2}\right)$$

where we have used

$$\frac{\alpha_1}{\bar{K}_1} + \frac{\alpha_2}{\bar{K}_2} = \left(\frac{1}{\bar{K}_1} + \frac{z_{\text{load}}}{z_l} \cdot \frac{K_1 - K_2}{|K_1|^2 |K_2|^2}\right) + \left(\frac{1}{\bar{K}_2} + \frac{z_{\text{load}}}{z_l} \cdot \frac{K_2 - K_1}{|K_1|^2 |K_2|^2}\right) = \frac{1}{\bar{K}_1} + \frac{1}{\bar{K}_2}$$

1 When $K_2 = K_1$, then $\alpha_1 = \alpha_2 = 1$ and

$$I_1' = I_2' = \frac{V_{\text{gen}}}{z_l + z_{\text{load}}\left(2|K_1|^{-2}\right)} \cdot \frac{\alpha_1}{\bar{K}_1} = \frac{K_1 V_{\text{gen}}}{|K_1|^2 z_l + 2 z_{\text{load}}}$$

and

$$I_{\text{load}} = \underbrace{\frac{V_{\text{gen}}}{|K_1|^2 z_l + 2 z_{\text{load}}}}_{I_0} \cdot 2K_1 = I_0 \cdot 2K_1 \tag{3.18}$$

2 When $K_2 = K_1\, e^{\mathbf{i}\theta}$, then, for $i = 1, 2$,

$$I_i' = \frac{V_{\text{gen}}}{z_l \,+\, z_{\text{load}}\left(2\,|K_1|^{-2}\right)} \cdot \frac{\alpha_i}{\bar{K}_i} \;=\; \frac{V_{\text{gen}}}{|K_1|^2 z_l \,+\, 2z_{\text{load}}} \cdot (\alpha_i\, K_i)$$

Since $\alpha_1 K_1 \,+\, \alpha_2 K_2 \;=\; K_1 + K_2 \;=\; K_1\,(1 + e^{\mathbf{i}\theta})$ and $|K_1| = |K_2|$, we have

$$I_{\text{load}} = \frac{V_{\text{gen}}}{|K_1|^2 z_l \,+\, 2z_{\text{load}}} \cdot \left(1 + e^{\mathbf{i}\theta}\right) K_1 \;=\; I_0 \left(1 + e^{\mathbf{i}\theta}\right) K_1$$

Hence I_{load} reduces to the load current in (3.18) when the transformer gains are equal with $\theta = 0$. When the transformer gains K_1 and K_2 are not in phase, $\left(1 + e^{\mathbf{i}\theta}\right)$ can be much smaller than 2 and the current $|I_{\text{load}}|$ that enters the load can be much smaller than the currents $|I_i'|$, $i = 1, 2$. In particular

$$\frac{|I_{\text{load}}|}{|I_1'|} \;=\; \frac{|1 + e^{\mathbf{i}\theta}|}{|\alpha_1|} \;\text{ and }\; \frac{|I_{\text{load}}|}{|I_2'|} \;=\; \frac{|1 + e^{\mathbf{i}\theta}|}{|\alpha_2|}$$

To appreciate the issue, take $K_1 = 10$, $K_2 = 10\, e^{\mathbf{i}\pi/6}$, $V_{\text{gen}} = 8$ kV, $z_l = j0.05\,\Omega$, $z_{\text{load}} = 800\angle 0°\,\Omega$. Then

$$\begin{aligned} I_1' &= 3{,}754.99 \;\angle -164.85 \text{ A} \\ I_2' &= 4{,}527.24 \;\angle 14.88 \text{ A} \\ I_{\text{load}} = I_1' + I_2' \;&=\; 772.50 \;\angle 13.57 \text{ A} \\ \frac{|I_{\text{load}}|}{|I_1'|} = 20.57\%, \qquad & \frac{|I_{\text{load}}|}{|I_2'|} \;=\; 17.06\% \end{aligned}$$

Hence $|I_1'|$ and $|I_2'|$ are much larger than $|I_{\text{load}}|$. The interpretation is that most of the current loops between the two transformer banks without entering the load. This is undesirable because the circulating current serves no purpose and heats up the transformers. The problem arises because the connection-induced phase shifts in the two parallel paths are different. In practice we will not parallelize these transformers.

The complex generation power and load power are respectively

$$\begin{aligned} S_{\text{gen}} &:= V_{\text{get}}(\bar{I}_1 + \bar{I}_2) \;&=\; 182.98 \;\angle 70.97° \text{ MVA} \\ S_{\text{load}} &:= z_{\text{load}} |I_{\text{load}}|^2 \;&=\; 59.68 \;\angle 0° \text{ MVA} \end{aligned}$$

Again the apparent load power is a small fraction of the apparent generation power. However, since the transformers have zero resistance, their real powers are the same:

$$P_{\text{gen}} \;=\; P_{\text{load}} \;=\; 59.68 \text{ MW}$$

3 When $K_2 = k \cdot K_1$, we have

$$I_1' = \frac{K_1\, V_{\text{gen}}}{|K_1|^2 z_l \,+\, \left(1+k^{-2}\right) z_{\text{load}}} \cdot \alpha_1$$

$$I_2' = \frac{K_1\, V_{\text{gen}}}{|K_1|^2 z_l \,+\, \left(1+k^{-2}\right) z_{\text{load}}} \cdot \frac{\alpha_2}{k}$$

$$I_{\text{load}} = \frac{V_{\text{gen}}}{|K_1|^2 z_l \,+\, \left(1+k^{-2}\right) z_{\text{load}}} \cdot \left(1+\frac{1}{k}\right) K_1$$

Hence

$$\frac{|I_{\text{load}}|}{|I_1'|} \;=\; \frac{1+k^{-1}}{|\alpha_1|} \;\text{ and }\; \frac{|I_{\text{load}}|}{|I_2'|} \;=\; \frac{1+k}{|\alpha_2|}$$

If we take $K_1 = 10$, $K_2 = 20$, $V_{\text{gen}} = 8$ kV, $z_l = j0.05\ \Omega$, $z_{\text{load}} = 800\angle 0^\circ\ \Omega$. Then

$$I_1' = 3,260.76\ \angle 76.40 \text{ A}$$

$$I_2' = 3,213.39\ \angle -86.58 \text{ A}$$

$$I_{\text{load}} = I_1' + I_2' \;=\; 959.23\ \angle -2.29 \text{ A}$$

$$\frac{|I_{\text{load}}|}{|I_1'|} = 29.42\%, \qquad \frac{|I_{\text{load}}|}{|I_2'|} \;=\; 29.85\%$$

Again $|I_1'|$ and $|I_2'|$ are much larger than $|I_{\text{load}}|$ and there is a large loop flow between the transformer banks. This time the problem arises because the voltage gains in the two parallel paths are different. In practice we will not parallelize these transformers.

□

3.5 Appendix: Per-unit normalization

In this appendix we describe a normalization method that will simplify the analysis of balanced three-phase systems. For a normal system where all connection-induced phase shifts of transformers can be ignored in the per-phase equivalent circuit, the system after normalization will contain no transformers if there is no off-nominal transformer in the original system. For general systems, normalization may simplify the equivalent circuit and per-phase analysis, but the system after normalization may contain ideal transformers with real or complex voltage gains. Normalization was important before the widespread use of powerful computers because it simplifies computation significantly. It is less important today, and some people argue, sometimes more error-prone than worth the effort.

We are usually interested in four types of generally complex quantities: power S,

voltages V, currents I, and impedances Z and functions of these quantities. We will choose *base values* for these quantities and define the quantities in per unit as:

$$\text{quantity in p.u.} := \frac{\text{actual quantity}}{\text{base value of quantity}}$$

The base values are chosen to be real positive values and have the same units as the corresponding actual quantities. For example a power base S_B will be in unit VA when it serves as the base value for complex power, W for real power, var for reactive power. Hence the per-unit quantities generally have different magnitudes from, but always the same phase as, the corresponding actual quantities. Furthermore they are dimensionless. The base values are chosen so that the per-unit quantities behave exactly as the actual quantities do, as we now explain.

Consider a power network that consists of multiple areas connected by transformers. It represents either a single-phase system or the per-phase equivalent circuit of a balanced three-phase system. The nominal voltage magnitudes are the same within each area and those in neighboring areas are related by transformer turns ratios. It is common to choose the power base value S_{1B} for the entire network and the voltage base value V_{1B} for one of the areas, say, area 1. For example the base value V_{1B} can be chosen to be the nominal voltage magnitude for area 1 and the base value S_B can be the rated apparent power of one of the transformers in area 1, so that its rated voltage is 1 pu and the rated power is 1 pu. The base values for all other quantities in the entire network are then calculated from these two values (S_B, V_{1B}) so that these base values satisfy:

- Kirchhoff's laws within each area;
- ideal transformer gains across areas;
- three-phase relations.

We derive in Chapter 3.5.1 the base values within area 1 and in Chapter 3.5.2 the base values of other areas connected by transformers to area 1. In Chapter 3.5.3 we describe the normalization of off-nominal transformers. In Chapter 3.5.4 we describe how to calculate base values of three-phase quantities in a balanced three-phase system. In Chapter 3.5.5 we summarize the procedure for per-unit per-phase analysis.

3.5.1 Kirchhoff's and Ohm's laws

Consider a single-phase system or the per-phase equivalent circuit of a three-phase system. Start with area 1 for which we have the power base S_B in VA (or W or var for real and reactive powers respectively) for the entire network, and the voltage base V_{1B} in V. The base values I_{1B}, Z_{1B} of currents and impedances respectively are calculated

as:

$$I_{1B} := \frac{S_B}{V_{1B}} \ A, \qquad Z_{1B} := \frac{V_B^2}{S_B} \ \Omega \tag{3.19}$$

so that the base values satisfy the Kirchhoff's laws:

$$V_{1B} = Z_{1B} I_{1B} \ V, \qquad S_B = V_{1B} I_{1B} \ VA$$

Since

$$\frac{V_1}{V_{1B}} = \frac{Z_1 I_1}{Z_{1B} I_{1B}}, \qquad \frac{S_1}{S_B} = \frac{V_1 I_1^*}{V_{1B} I_{1B}}$$

the per-unit quantities satisfy Kirchhoff's laws as the actual quantities do:

$$V_{1\text{pu}} = Z_{1\text{pu}} I_{1\text{pu}}, \qquad S_{1\text{pu}} = V_{1\text{pu}} I_{1\text{pu}}^*$$

We can therefore perform circuit analysis using the per-unit quantities instead of the actual quantities. We can convert the result of the analysis back to the original quantities by multiplying the per-unit quantities by their base values.

Extensions to other related quantities are straightforward. For example S_B is also the base value for real power in W and reactive power in var so that

$$P_{1\text{pu}} := \frac{P_1}{S_B}, \quad Q_{1\text{pu}} := \frac{Q_1}{S_B}$$

and $S_{1\text{pu}} = P_{1\text{pu}} + jQ_{1\text{pu}}$. Z_B is the base value for resistances and reactances so that

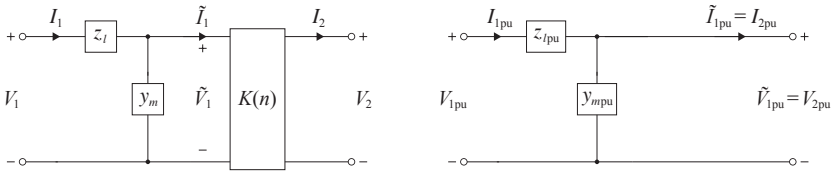
$$R_{1\text{pu}} := \frac{R_1}{Z_{1B}}, \quad X_{1\text{pu}} := \frac{X_1}{Z_{1B}}$$

and $Z_{1\text{pu}} = R_{1\text{pu}} + jX_{1\text{pu}}$. Similarly $Y_{1B} := 1/Z_{1B}$ in Ω^{-1} is the base value for admittances $Y_1 := 1/Z_1 = G - jB$ in Ω^{-1} as well as conductances G and susceptances B also in Ω^{-1}.

3.5.2 Across ideal transformer

Consider now a neighboring area, say, area 2 that is connected to area 1 through a transformer. We choose the bases for different sides of the transformer in a way that respects the transformer gains. Consider the circuit in Figure 3.29(a) where areas 1 and 2 are connected through a transformer with a voltage gain $K(n)$. If it is a single-phase system then $K(n) = n$, the reciprocal of the turns ratio. If it is the per-phase equivalent of a balanced three-phase system then $K(n)$ may be complex if the transformer is not in YY or $\Delta\Delta$ configuration. Given the bases $(S_B, V_{1B}, I_{1B}, Z_{1B})$ for area 1 calculated in Chapter 3.5.1, the bases for the other side of the transformer are calculated according to:

$$V_{2B} := |K(n)| V_{1B} \ V, \qquad I_{2B} := \frac{I_{1B}}{|K(n)|} \ A, \qquad Z_{2B} := |K(n)|^2 Z_{1B} \ \Omega \tag{3.20}$$



(a) In standard unit (b) In per unit

Figure 3.29 Per-phase equivalent circuit of balanced three-phase transformers with gain $K(n)$.

The base power value remains $S_B = V_{1B} I_{1B} = V_{2B} I_{2B}$ for all areas since the power gain across an ideal transformer is 1. Even though $K(n)$ may be complex all base values remain real positive numbers.

Referring to Figure 3.29(a), the per-unit quantities $(\tilde{V}_{1\text{pu}}, \tilde{I}_{1\text{pu}})$ at the input and the per-unit quantities $(V_{2\text{pu}}, I_{2\text{pu}})$ at the output of the *ideal* transformer satisfy $(a := 1/n)$

$$\tilde{V}_{1\text{pu}} = \frac{\tilde{V}_1}{V_{1B}} = \frac{V_2}{K(n)} \frac{|K(n)|}{V_{2B}} = V_{2\text{pu}}\, e^{-j\angle K(n)}$$
$$\tilde{I}_{1\text{pu}} = \frac{\tilde{I}_1}{\tilde{I}_{1B}} = \frac{K^*(n) I_2}{|K(n)| I_{2B}} = I_{2\text{pu}}\, e^{-j\angle K(n)}$$

This also implies that the per-unit power $\tilde{S}_{1\text{pu}} := \tilde{V}_{1\text{pu}} \tilde{I}^*_{1\text{pu}} = V_{2\text{pu}} I^*_{2\text{pu}} = S_{2\text{pu}}$. If $\angle K(n)$ can be taken as zero then on the input side of the transformer, $(\tilde{V}_{1\text{pu}}, \tilde{I}_{1\text{pu}}, \tilde{S}_{1\text{pu}})$ can be replaced by $(V_{2\text{pu}}, I_{2\text{pu}}, S_{2\text{pu}})$, i.e., the voltages, currents, and power remain the same, in per unit, when crossing an *ideal* transformer. Within each side of the ideal transformer the per-unit quantities $(S_{i\text{pu}}, V_{i\text{pu}}, I_{i\text{pu}}, Z_{i\text{pu}})$ satisfy the Kirchhoff's laws as explained in Chapter 3.5.1. Hence the per-phase equivalent circuit can be simplified into that in Figure 3.29(b) where the ideal transformer has disappeared. The voltage gain angle $\angle K(n) = 0$ if (i) the system is single phased, or (ii) it is balanced three phased with transformers in YY or $\Delta\Delta$ configuration, or (iii) it is a normal system where the connection induced phase shift $\angle K(n)$ can be ignored for external behavior. Hence ideal transformers and connection-induced phase shifts can be omitted in a normal per-phase system if we use the simplified per-phase diagram and the per-unit normalization. This simplified per-phase per-unit diagram is called an *impedance diagram*. Otherwise the per-unit circuit will contain a phase-shifting transformer with voltage gain $e^{j\angle K(n)}$; see Example 3.10.

We proceed in a similar manner to calculate the base values $(S_B, V_{iB}, I_{iB}, Z_{iB})$ in each neighboring area i, until all connected areas are covered. It can be easily checked that the per-unit quantities in each area satisfy the Kirchhoff's laws, as long as the per-unit quantities in area 1 satisfy the Kirchhoff's laws and those in other areas respect transformer gains. This is where system normality is important: on each parallel path in its per-phase equivalent circuit, (i) the product of ideal transformer gain magnitudes is the same, and (ii) the sum of ideal transformer phase shifts is the

same. As discussed above these properties prevent loop flows between transformers, as illustrated in Example 3.8. Note that in Figure 3.28(b) of that example, the secondary-side voltages of the two *ideal* transformers are the same but their primary-side voltages are different when $K_2 = K_1 e^{j\theta}$ with $\theta \neq 0$. The first property also ensures that the calculation (3.20) of base values across areas is consistent, i.e., does not depend on the order in which the areas are chosen for calculation; see Exercise 3.13.

Example 3.9 (Single-phase system)**.** Consider the single-phase system in Figure 3.30 where the voltage source has a nameplate rated voltage magnitude of v V and a nameplate rated power of s VA. Calculate the base values for the system.

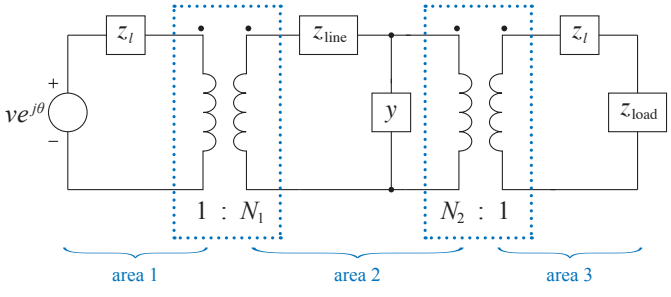


Figure 3.30 Single-phase system for Example 3.9 with a rated voltage magnitude of v in V and a rated apparent power of s in VA.

Solution. Let the base value for power be $S_B := s$ in VA for the entire system and the base value for voltage in area 1 (where the voltage source is) be $V_{1B} := v$ in V. Then the base values for currents and impedances in area 1 are respectively:

$$I_{1B} := \frac{s}{v} \ A \quad \text{and} \quad Z_{1B} := \frac{v^2}{s} \ \Omega$$

The base values in area 2 connected by the first transformer with a voltage gain n_1 are:

$$\begin{aligned} V_{2B} &:= n_1 V_{1B} = n_1 v \ V \\ I_{2B} &:= \frac{I_{1B}}{n_1} = \frac{s}{n_1 v} \ A \\ Z_{2B} &:= n_1^2 Z_{1B} = \frac{(v_1 v)^2}{s} \ \Omega, \qquad Y_{2B} := \frac{1}{Z_{2B}} = \frac{s}{(v_1 v)^2} \ \Omega^{-1} \end{aligned}$$

The base values in area 3 connected by the second transformer are:

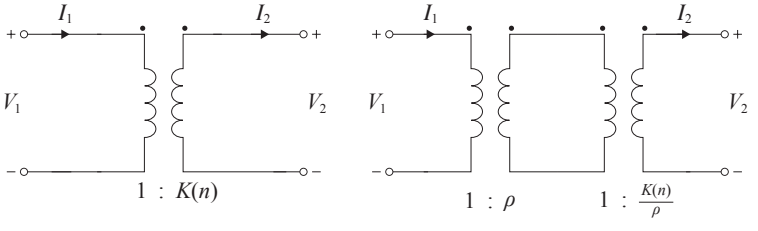
$$\begin{aligned} V_{3B} &:= \frac{V_{2B}}{n_2} = \frac{n_1}{n_2} v \ V \\ I_{3B} &:= n_2 I_{2B} = \frac{n_2}{n_1} \frac{s}{v} \ A \\ Z_{3B} &:= \frac{1}{n_2^2} Z_{2B} = \frac{n_1^2}{n_2^2} \frac{v^2}{s} \ \Omega, \qquad Y_{3B} := \frac{1}{Z_{3B}} = \frac{n_2^2}{n_1^2} \frac{s}{v^2} \ \Omega^{-1} \end{aligned}$$

□

3.5.3 Off-nominal transformer

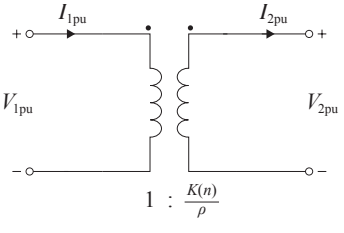
Power systems employ two types of regulating transformers. The first type regulates voltage magnitudes, e.g., through variable taps on some of its windings that control the number of turns and hence the voltage gain. Such a transformer is usually connected at the end of a line to regulate the voltage magnitude at a node. Its turns ratio may be variable and different from the ratio of the voltage bases in its primary and secondary areas. The second type regulates phase angle displacement between two nodes. Their voltage gains may be complex $K(n) = \rho\angle\phi$ where ϕ may be variable and cannot be omitted in normalization. These transformers are said to be *off-nominal*. They will not disappear under per-unit normalization but will appear as a transformer with a different (normalized) voltage gain, as we now explain.

Consider an ideal transformer with a possibly complex voltage gain $\frac{V_2}{V_1} =: K(n)$ as shown in Figure 3.31(a). Suppose the ratio of the voltage base in area 2 to that in area



(a) Off-nominal transformer (standard unit)

(b) Equivalent transformer in series (standard unit)



(c) Per-unit equivalent circuit

Figure 3.31 Normalization of an off-nominal transformer.

1 is $\frac{V_{2B}}{V_{1B}} =: \rho$. Since

$$V_2 = K(n)\,V_1 \;=\; \frac{K(n)}{\rho}\cdot\rho V_1$$

the transformer is equivalent to two ideal transformers in series with voltage gains ρ and $K(n)/\rho$ respectively as shown in Figure 3.31(b). Since the first transformer has an voltage gain of ρ, it disappears in per-unit normalization and hence the per-unit equivalent circuit of the original transformer has a gain reduced by ρ as shown in Figure

3.31(c). For instance for a phase shifting transformer with voltage gain $K(n) = \rho\angle\phi$ its voltage gain in the per-unit circuit will be $1\angle\phi$.

Example 3.10 (Normalization with connection-induced phase shifts)**.** Consider a balanced three-phase ideal transformer in ΔY or $Y\Delta$ configuration with a complex voltage gain $K(n)$. Let the bases for one side of the transformer be $(S_B, V_{1B}, I_{1B}, Z_{1B})$. Choose the bases for the other side according to (3.20). Suppose we cannot ignore the connection-induced phase shift. Then the per-unit equivalent circuit of the ideal transformer will be an off-nominal phase shifting transformer with a gain $\frac{K(n)}{|K(n)|} = \angle K(n)$ as shown in Figure 3.32. □

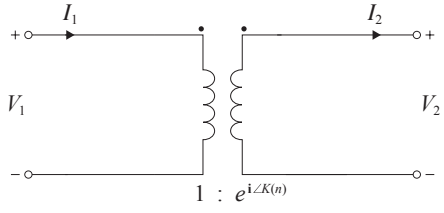


Figure 3.32 Normalization when connection-induced phase shifts cannot be ignored.

As we will see in Chapter 4.2 a *nonideal* transformer, whether in standard unit or per unit, can be represented by a phase impedance matrix for power flow analysis.

3.5.4 Three-phase quantities

In Chapters 3.5.1–3.5.3 we explain how to choose bases for a single-phase system. They are also applicable to the per-phase equivalent of a three-phase system where the voltages and currents are line-to-neutral voltages and line-to-neutral currents. Suppose the base values $(S_B^{1\phi}, V_B^{1\phi}, I_B^{1\phi}, Z_B^{1\phi})$ for a single-phase system are given. When single-phase devices (sources, loads, impedances, transformers) are connected to form a balanced three-phase system, three-phase quantities are created for which base values need to be defined. For instance the ratings of a three-phase transformer are always specified in terms of three-phase power and line-to-line voltages. In this subsection we will derive these base values, in terms of $(S_B^{1\phi}, V_B^{1\phi}, I_B^{1\phi}, Z_B^{1\phi})$, in a way that respects three-phase relations. The main issue is to define the meaning of these base values and the relation they intend to capture in Y and in Δ configurations.

Let $(S^{1\phi}, V^{1\phi}, I^{1\phi}, Z^{1\phi})$ denote respectively the power generated or consumed by a single-phase device, the voltage across and current through the device, and the impedance of the device. We are interested in the following three-phase quantities. The three-phase power $S^{3\phi}$ is defined to be the sum of power generated or consumed by each device in either Y or Δ configuration. The line-to-line voltages V^{ll} and terminal

(line) currents $I^{3\phi}$ are external quantities. In an Y configured three-phase device, a line-to-neutral voltage $V^{\ln}$ and a three-phase impedance $Z^{3\phi}$ are equal to the voltage $V^{1\phi}$ and impedance $Z^{1\phi}$ respectively associated with each single-phase device. For a Δ configured three-phase device $V^{\ln}$ and $Z^{3\phi}$ are defined to be the line-to-neutral voltage and the impedance respectively in its Y *equivalent circuit*. As explained in Chapter 1 these quantities are related to the corresponding single-phase quantities according to:[3]

$$S^{3\phi} = 3S^{1\phi}, \qquad V^{\ll} = \sqrt{3}e^{\mathbf{i}\pi/6}V^{\ln} \tag{3.21a}$$

$$I^{3\phi} = \begin{cases} I_{an} = I^{1\phi} & \text{for } Y \text{ configuration} \\ I_{ab} - I_{ca} = \sqrt{3}e^{-j\pi/6}I^{1\phi} & \text{for } \Delta \text{ configuration} \end{cases} \tag{3.21b}$$

$$V^{\ln} = \begin{cases} V^{1\phi} & \text{for } Y \text{ configuration} \\ \left(\sqrt{3}e^{j\pi/6}\right)^{-1} V^{1\phi} & \text{for } \Delta \text{ configuration} \end{cases} \tag{3.21c}$$

$$Z^{3\phi} = \begin{cases} Z^{1\phi} & \text{for } Y \text{ configuration} \\ Z^{1\phi}/3 & \text{for } \Delta \text{ configuration} \end{cases} \tag{3.21d}$$

Motivated by the three-phase relations (3.21) we define the base values $(S_B^{3\phi}, V_B^{\ll}, I_B^{3\phi}, V_B^{\ln}, Z_B^{3\phi})$ for the three-phase quantities $(S^{3\phi}, V^{\ll}, I^{3\phi}, V^{\ln}, Z^{3\phi})$ in terms of the single-phase base values $(S_B^{1\phi}, V_B^{1\phi}, I_B^{1\phi}, Z_B^{1\phi})$ as follows:

$$S_B^{3\phi} := 3S_B^{1\phi}, \qquad V_B^{\ll} := \sqrt{3}V_B^{\ln} \tag{3.22a}$$

$$I_B^{3\phi} := \begin{cases} I_B^{1\phi} & \text{for } Y \text{ configuration} \\ \sqrt{3}I_B^{1\phi} & \text{for } \Delta \text{ configuration} \end{cases} \tag{3.22b}$$

$$V_B^{\ln} := \begin{cases} V_B^{1\phi} & \text{for } Y \text{ configuration} \\ \left(\sqrt{3}\right)^{-1} V_B^{1\phi} & \text{for } \Delta \text{ configuration} \end{cases} \tag{3.22c}$$

$$Z_B^{3\phi} := \begin{cases} Z_B^{1\phi} & \text{for } Y \text{ configuration} \\ Z_B^{1\phi}/3 & \text{for } \Delta \text{ configuration} \end{cases} \tag{3.22d}$$

In light of (3.19) we could also have defined the base values $I_B^{3\phi}$ and $Z_B^{3\phi}$ in terms of $S_B^{3\phi}$ and $V_B^{\ll}$ as (see Exercise 3.14):

$$I_B^{3\phi} := \frac{S_B^{3\phi}}{\sqrt{3}V_B^{\ll}}, \qquad Z_B^{3\phi} := \frac{\left(V_B^{\ll}\right)^2}{S_B^{3\phi}} \tag{3.22e}$$

These definitions replace (3.22b) and (3.22d) and are applicable for both Y and Δ configurations (note that $V_B^{\ll}$ are different functions of $V_B^{1\phi}$ for Y and Δ configurations).

With these base values the per-unit quantities satisfy the following relations (see

Exercise 3.15):

$$S_{\text{pu}}^{3\phi} = S_{\text{pu}}^{1\phi}, \qquad V_{\text{pu}}^{\text{ll}} = V_{\text{pu}}^{\text{ln}}, \qquad Z_{\text{pu}}^{3\phi} = Z_{\text{pu}}^{1\phi} \tag{3.23a}$$

$$\left|I_{\text{pu}}^{3\phi}\right| = \left|I_{\text{pu}}^{1\phi}\right|, \qquad \left|V_{\text{pu}}^{\text{ln}}\right| = \left|V_{\text{pu}}^{1\phi}\right| \tag{3.23b}$$

Therefore in per unit, the three-phase power, voltage, current and impedance equal their per-phase quantities (at least in magnitude). In particular when one says that the voltage magnitude is 1 pu, it means that the line-to-line voltage magnitude is 1 pu (i.e., equal to its base value V_B^{ll} which is $\sqrt{3}V_B^{1\phi}$ for Y configuration and $V_B^{1\phi}$ for Δ configuration), and the phase voltage magnitude is 1 pu (i.e., equal to its base value V_B^{ln} which is $V_B^{1\phi}$ for Y configuration and $\left(\sqrt{3}\right)^{-1} V_B^{1\phi}$). We sometimes need not specify whether a per-unit voltage is line-to-line or line-to-neutral, or whether a per-unit power is single-phase or three-phase. In Δ configuration the line-to-neutral voltage $V_{\text{pu}}^{\text{ln}}$ is related to single-phase voltage $V_{\text{pu}}^{1\phi}$ according to

$$V_{\text{pu}}^{\text{ln}} := \frac{V^{\text{ln}}}{V_B^{\text{ln}}} = \frac{\left(\sqrt{3}e^{i\pi/6}\right)^{-1} V^{1\phi}}{\left(\sqrt{3}\right)^{-1} V_B^{1\phi}} = e^{-i\pi/6} V_{\text{pu}}^{1\phi}$$

Similarly for line currents $I_{\text{pu}}^{3\phi}$ and $I_{\text{pu}}^{1\phi}$.

The next example illustrates the calculation of three-phase bases from single-phase bases. It shows in particular that impedances, including transformer parameters, will have the same per-unit values in single-phase or three-phase circuits and regardless of Y or Δ configuration.

Example 3.11 (Three-phase system)**.** Consider a single-phase distribution transformer with nameplate ratings of

- Power rating (1ϕ): 50 kVA;
- Voltage ratio: 408 V – 120 V;
- Transformer parameter: $X_l = 0.1$ pu, $X_m = 100$ pu (referred to the primary).

They are used to build three-phase transformer banks in YY, $\Delta\Delta$, ΔY or $Y\Delta$ configurations. Find the per-unit normalization "induced" by the nameplate ratings and the impedance diagram of the per-phase circuit in per unit.

Solution. The nameplate-induced base for the *single-phase* transformer is such that the power rating is 1pu and voltage rating is 1pu. Hence

$$S_B^{1\phi} := 50\,\text{kVA}, \qquad V_{1B}^{1\phi} := 408\,\text{V}, \qquad V_{2B}^{1\phi} := 120\,\text{V}$$

Therefore the current bases are

$$I_{1B}^{1\phi} := \frac{S_B^{1\phi}}{V_{1B}^{1\phi}} = \frac{50\,\text{kVA}}{408\,\text{V}} = 122.55\,\text{A}, \qquad I_{2B}^{1\phi} := \frac{S_B^{1\phi}}{V_{2B}^{1\phi}} = \frac{50\,\text{kVA}}{120\,\text{V}} = 416.67\,\text{A}$$

Since $S = |V|^2/Z$, the impedance base for the single-phase transformer induced by the nameplate ratings is:

$$Z_{1B}^{1\phi} = \frac{\left(V_{1B}^{1\phi}\right)^2}{S_B^{1\phi}} = \frac{(408\,\mathrm{V})^2}{50\,\mathrm{kVA}} = 3.33\,\Omega, \qquad Z_{2B}^{1\phi} = \frac{\left(V_{2B}^{1\phi}\right)^2}{S_B^{1\phi}} = \frac{(120\,\mathrm{V})^2}{50\,\mathrm{kVA}} = 0.29\,\Omega$$

Hence the actual transformer reactances X_l and X_m in Ω in the single-phase system are:

$$X_l = (0.1)\, Z_{1B}^{1\phi} = 0.333\,\Omega, \qquad X_m = (100)\, Z_{1B}^{1\phi} = 333\,\Omega$$

Consider now a three-phase transformer bank obtained from connecting three of these single-phase transformers. We consider first the base values for the primary side; the base values for the secondary side can be similarly chosen. What we will find is that if we choose our bases $(S_B^{3\phi}, V_B^{\mathrm{ll}}, I_B^{3\phi}, Z_B^{3\phi})$ according to (3.22), then the impedance diagram of the per-phase equivalent circuit is independent of Y or Δ configuration.

Case 1: primary side in Y configuration. From (3.22), the base values of the three-phase power and line-to-line voltage induced by the nameplate ratings are

$$S_B^{3\phi} := 3\, S_B^{1\phi} = 3(50) = 150\,\mathrm{kVA}$$
$$V_{1B}^{\mathrm{ll}} := \sqrt{3}\, V_B^{1\phi} = \sqrt{3}\,(408) = 706.68\,\mathrm{V}$$

These three-phase quantities are used as the power and voltage ratings on the three-phase transformer nameplate. Hence a line voltage of 1 pu corresponds to the rated primary voltage (706.68 V) on the nameplate. The base values for the terminal currents and impedances are:

$$I_{1B}^{3\phi Y} := I_{1B}^{1\phi} = 122.55\ A, \qquad Z_{1B}^{3\phi Y} := Z_{1B}^{1\phi} = 3.33\,\Omega$$

It can be checked that $(S_B^{3\phi}, V_B^{\mathrm{ll}}, I_B^{3\phi}, Z_B^{3\phi})$ as defined indeed satisfy three-phase relations:

$$I_{1B}^{3\phi Y} = \frac{S_B^{3\phi}}{\sqrt{3}\, V_{1B}^{\mathrm{ll}}}, \qquad Z_{1B}^{3\phi Y} = \frac{\left(V_B^{\mathrm{ll}}\right)^2}{S_B^{3\phi}}$$

Since $Z_{1B}^{3\phi Y} = Z_{1B}^{1\phi}$, $X_l = 0.1$pu and $X_m = 100$pu as before for the three-phase transformer.

Case 2: primary side in Δ configuration. From (3.22), the base values of the three-phase power and line-to-line voltage induced by the nameplate ratings are

$$S_B^{3\phi} := 3\, S_B^{1\phi} = 3(50) = 150\,\mathrm{kVA}, \quad V_{1B}^{\mathrm{ll}} := V_B^{1\phi} = 408\,\mathrm{V}$$

The terminal current and the impedance bases are:

$$I_{1B}^{3\phi Y} := \sqrt{3}\, I_{1B}^{1\phi} = \sqrt{3}(122.55) = 212.26\ A, \qquad Z_{1B}^{3\phi\Delta} = \frac{Z_B^{1\phi}}{3} = \frac{3.33}{3} = 1.11\,\Omega$$

To convert the transformer circuit model in Δ configuration to its equivalent Y configuration, the transformer reactances are reduced by a factor of 3, i.e., $X_l^Y = X_l/3$ and $X_m^Y = X_m/3$. Hence the transformer reactances in pu are:

$$X_{l\text{pu}}^Y := \frac{X_l^Y}{Z_{1B}^{3\phi}} = \frac{X_l/3}{Z_{1B}^{1\phi}/3} = \frac{X_l}{Z_{1B}^{1\phi}} = 0.1 \text{ pu}$$

$$X_{m\text{pu}}^Y := \frac{X_m^Y}{Z_{1B}^{3\phi}} = \frac{X_m/3}{Z_{1B}^{1\phi}/3} = \frac{X_m}{Z_{1B}^{1\phi}} = 100 \text{ pu}$$

as expected.

In summary, with the three-phase base values defined in (3.22), the transformer reactances X_l and X_m remain the same in pu regardless of how the single-phase transformers are connected into a three-phase transformer bank. The impedance diagram of its per-phase circuit is shown in Figure 3.33. □

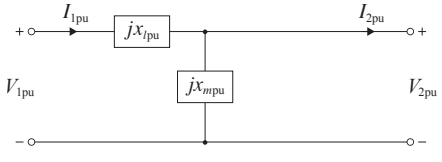


Figure 3.33 Impedance diagram of a three-phase transformer bank.

3.5.5 Per-unit per-phase analysis

Consider a balanced three-phase normal system. Recall that the nameplate ratings of three-phase transformers are specified in terms of their three-phase power and line-to-line voltages. The procedure for per-unit per-phase analysis is summarized as follows:

1 For a single-phase system, pick a power base $S_B^{1\phi}$ for the *entire* system and a voltage base V_{1B}^{ln} in *one* of the areas, e.g., induced by the nameplate ratings of one of the single-phase transformers.
2 For a balanced three-phase system, pick a three-phase power base $S_B^{3\phi}$ and line-to-line voltage base V_{1B}^{ll} induced by the nameplate ratings of one of the three-phase transformers in area 1 (choose either the primary or secondary circuit as area 1). Then choose the power and voltage bases for the per-phase equivalent circuit of the balanced three-phase system according to (3.22a):

$$S_B^{1\phi} := \frac{S_B^{3\phi}}{3} \quad \text{and} \quad V_{1B}^{1\phi} := \frac{V_{1B}^{\text{ll}}}{\sqrt{3}}$$

$S_B^{1\phi}$ will be the power base for the entire per-phase circuit.

3 Calculate the current and impedance bases in that area by:

$$I_{1B} := \frac{S_B^{1\phi}}{V_{1B}^{1\phi}} \text{ and } Z_{1B} := \frac{\left(V_{1B}^{1\phi}\right)^2}{S_B^{1\phi}}$$

4 Calculate the base values for voltages, currents, and impedances in areas i connected to area 1 by the magnitudes n_i of the transformer gains (assuming area 1 is the primary side of the transformers):

$$V_{iB}^{1\phi} := n_i V_{1B}^{1\phi}, \quad V_{iB}^{ll} := n_i V_{1B}^{ll} \quad I_{iB} := \frac{1}{n_i} I_{1B}, \quad Z_{iB} := n_i^2 Z_{1B}$$

Continue this process to calculate the voltage, current, and impedance base values for all areas.

5 For real, reactive, apparent power in the entire system, use $S_B^{1\phi}$ as the base value. For resistances and reactances, use Z_{iB} as the base value in area i. For admittances, conductances, and susceptances, use $Y_{iB} := 1/Z_{iB}$ as the base value in area i.

6 Draw the impedance diagram of the entire system, and solve for the desired per unit quantities.

7 Convert back to actual quantities if desired.

3.6 Bibliographical notes

There are many excellent textbooks on basic power system concepts and many materials in this chapter follow [1]. Some of the materials on per-unit normalization, e.g., off-nominal regulating transformer in Chapter 3.5.3, follow [2]. [10] describes a rigorous approach that treats per-unit normalization as a similarity transformation of a dynamical system in the time domain. The per-unit normalization presented in this chapter represents the steady-state of the per-unit dynamical system of [10].

3.7 Problems

Chapter 3.1.

Exercise 3.1 (T model of transformer)**.** For the T equivalent circuit of transformer in Figure 3.34, show that the transmission matrix is given in (3.5). If $y_m = 0$ then

$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = \begin{bmatrix} a & n\left(z_p + a^2 z_s\right) \\ 0 & n \end{bmatrix} \begin{bmatrix} V_2 \\ I_2 \end{bmatrix}$$

which is the same as the transmission matrix in (3.7a).

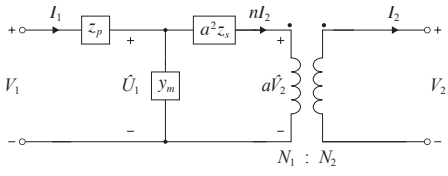


Figure 3.34 Exercise 3.1: T equivalent circuit of transformer with $n := N_2/N_1$ and $a := N_1/N_2$.

Exercise 3.2 (T model of transformer)**.** Given the primary voltages and primary currents (V_{sc}, I_{sc}) and (V_{oc}, I_{oc}) of a short-circuit and open-circuit tests respectively, derive (3.6), reproduced here:

$$V_{sc} = \left(z_p + \left(y_m + \frac{1}{a^2 z_s}\right)^{-1}\right) I_{sc}, \qquad V_{oc} = \left(z_p + \frac{1}{y_m}\right) I_{oc} \tag{3.24}$$

from (3.4), reproduced here:

$$\text{Nonideal elements:} \quad V_1 = z_p I_1 + \hat{V}_1, \quad \hat{I}_m = y_m \hat{V}_1, \qquad \hat{V}_2 = z_s I_2 + V_2 \tag{3.25a}$$

$$\text{Ideal transformer:} \quad \hat{V}_2 = \frac{N_2}{N_1}\hat{V}_1, \qquad I_2 = \frac{N_1}{N_2}\left(I_1 - \hat{I}_m\right) \tag{3.25b}$$

where the series impedances

Exercise 3.3 (Simplified model)**.** Consider the transformer model in Figure 3.5 and its transmission matrix $\hat{M}$ in (3.7a). This question shows that when the shunt admittance matrix y_m is small compared with the series admittances z_s, $\hat{M}$ is a good approximation the transmission matrix M in (3.5). Let $\epsilon := a^2 z_s y_m$.

1 Show that their difference is $\hat{M} - M = \epsilon \begin{bmatrix} a & -nz_p \\ 0 & -n \end{bmatrix}$.

2 Suppose $z_p = \eta z_s = \eta(r_s + \mathbf{i}x_s)$ for some real number $\eta > 0$ with $r_s > 0$ and $x_s > 0$, $y_m = -\mathbf{i}b_m$ with $b_m > 0$, and $|\epsilon| \ll 1$. Show that $\frac{\|\hat{M} - M\|}{\|M\|} < |\epsilon| \ll 1$, where $\|A\|$ denotes the sum norm $\|A\| := \sum_{i,j} |A_{ij}|$.

Exercise 3.4 (Unitary voltage network)**.** Show that the T equivalent circuit described by (3.5) is equivalent to the transformer model $I = (MY_{\text{uvn}}M)V$ given by (3.11).

Exercise 3.5 (Unitary voltage network)**.** Show that, instead of the numbers N_1, N_2 of turns of the primary and secondary windings respectively, the admittance matrix

$MY_{\text{uvn}}M$ in (3.11) can equivalently be written in terms of the turns ratio $a := N_1/N_2$:

$$MY_{\text{uvn}}M = \frac{y_p y_s}{a^2 y_m + a^2 y_p + y_s} \begin{bmatrix} 1 + a^2 y_m / y_s & -a \\ -a & a^2(1 + y_m/y_p) \end{bmatrix}$$

Chapter 3.2.

Exercise 3.6 (ΔY and $Y\Delta$ configurations)**.** Consider ideal balanced three-phase transformers in ΔY and $Y\Delta$ configurations shown in Figure 3.14(b). Show that an $Y\Delta$ transformer with *single-phase* voltage gains $1/n$ is equivalent to a ΔY transformer with single-phase voltage gains n with its primary and secondary sides switched.

Exercise 3.7 (Nonideal ΔY transformer)**.** Consider a balanced three-phase transformers in ΔY configuration and its per-phase equivalent circuit shown in Figure 3.17. Show that the transmission matrix of the per-phase equivalent circuit is given by:

$$\begin{bmatrix} V_1^{an} \\ I_1^a \end{bmatrix} = \begin{bmatrix} K_{\Delta Y}^{-1}(n)\,(1 + z_l y_m) & \bar{K}_{\Delta Y}(n)(z_l/3) \\ K_{\Delta Y}^{-1}(n)\,(3y_m) & \bar{K}_{\Delta Y}(n) \end{bmatrix} \begin{bmatrix} V_2^{an} \\ I_2^a \end{bmatrix}$$

where $K_{\Delta Y}(n) := \sqrt{3}\, n\, e^{\mathbf{i}\pi/6}$.

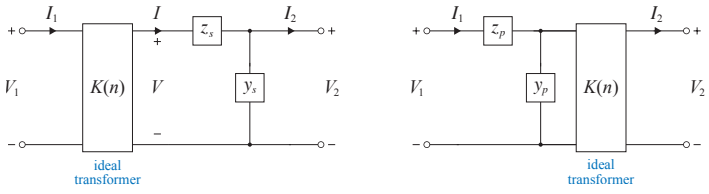
Exercise 3.8 (Referring shunt admittance in one side to the other)**.** Show that the transmission matrix for the circuit in Figure 3.20(a) is the same as that in Figure 3.20(b) provided that the relation (3.14b) between shunt admittances y_p and y_s holds.

Exercise 3.9 (Transmission matrix)**.** Consider a balanced three-phase ideal transformer with a complex gain $K(n)$ connected to a balanced three-phase series impedance z_s and a balanced three-phase shunt admittance y_s on the secondary side. The per-phase equivalent circuit is shown in Figure 3.35(a). Show directly that transmission matrix of the circuit in Figure 3.35(a) is the same as that in Figure 3.35(b) provided the relation (3.14) between impedances/admittances (z_p, y_p) and (z_s, y_s) holds.

.

Exercise 3.10 (Driving-point impedance)**.** Refer to Figure 3.23.

1 Show that the driving-point impedance V_1/I_1 on the primary side is the same in both circuits in Figure 3.23(a).



(a) (z_s, y_s) in the secondary (b) (z_p, y_p) in the primary

Figure 3.35 Referring (z_s, y_s) on the secondary to the primary for an ideal transformer with a complex gain $K(n)$.

2 Show that the driving-point impedance V_2/I_2 on the secondary side is the same in both circuits in Figure 3.23(b).

Exercise 3.11 (Driving-point impedance on primary side)**.** Suppose the secondary sides of the (equivalent) circuits in Figure 3.35 are connected to an identical load Z_{load} so that $V_2 = Z_{\text{load}} I_2$ in both circuits.

1 Show that the driving-point impedances on the primary side of the circuit in Figure 3.35(a) is:

$$\frac{V_1}{I_1} = \frac{1}{|K(n)|^2}\left(Z_s + \frac{1}{Y_s + Z_{\text{load}}^{-1}}\right) \tag{3.26a}$$

The term in the bracket is the Thévain equivalent impedance in the secondary circuit, seen from the output of the ideal transformer.

2 Show that the driving-point impedances on the primary side of the circuit in Figure 3.35(b) is:

$$\frac{V_1}{I_1} = Z_p + \frac{1}{Y_p + |K(n)|^2 Z_{\text{load}}^{-1}} \tag{3.26b}$$

3 Show that (3.26a) and (3.26b) are equivalent provided that (Z_p, Y_p) and (Z_s, Y_s) satisfy (3.14).

Exercise 3.12. Consider the balanced three phase system in Figure 3.36 where the line-to-line voltage of the three-phase generator in Δ configuration is V_{gen}. The 3ϕ transformer consists of single-phase transformers in ΔY configuration. Each single-phase transformer is modeled by a series impedance Z_l (and negligible shunt admittance) on the primary side followed by an ideal transformer with turn ratio n. The transmission line is modeled by a Π-model with a series impedance Z_s and a shunt admittance $Y_m/2$ at each end of the line. The transmission line is connected to a balanced 3ϕ impedance load in Y configuration with an impedance Z_{load} in each phase.

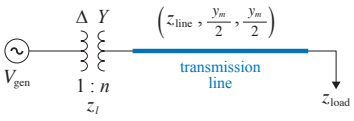


Figure 3.36 A three-phase generator in Δ configuration connected to a three-phase ΔY transformer and then to a three-phase load in Y configuration through a three-phase AC transmission line.

1 Draw the equivalent per-phase circuit.
2 Derive the complex power delivered to the load Z_{load} in each phase.

Exercise 3.13 (Bases across transformers)**.** For a normal system, on each parallel path in its per-phase equivalent circuit, the product of ideal transformer gain magnitudes is the same. Show that this property allows us to consistently define base values between two neighboring areas using (3.20). (Hint: Show that around any loop, (3.20) holds only if the product of voltage gain magnitudes around the loop is 1.)

Exercise 3.14 (Terminal current and three-phase impedance bases)**.** Show the definition (3.22b) (3.22d) for base values $I_B^{3\phi}$ and $Z_B^{3\phi}$ respectively are equivalent to definition (3.22e).

Exercise 3.15 (Per unit properties)**.** Prove the per-unit properties (3.23).

Exercise 3.16 (Caltech ACN: transformers)**.** Figure 3.37 shows the layout of the Adaptive Charging Network (ACN) for electric vehicles (EVs) in a Caltech garage. The

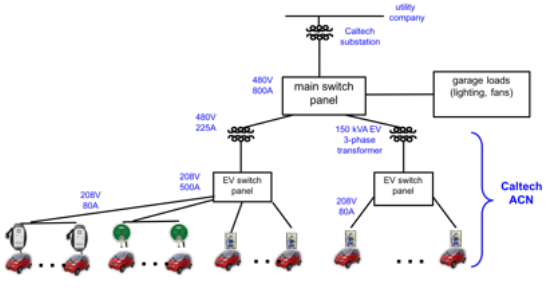
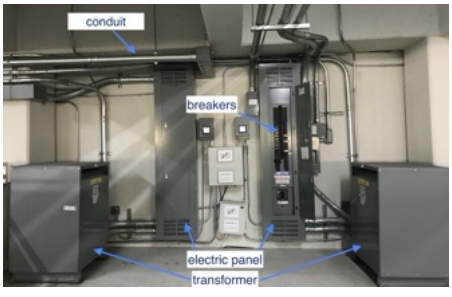


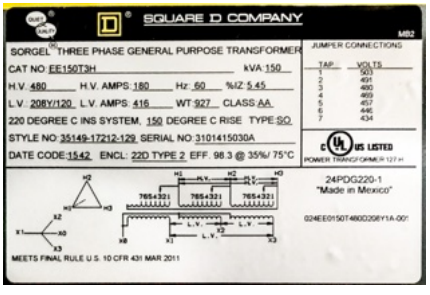
Figure 3.37 Caltech Adaptive Charging Network (ACN) layout. (March 9, 2025: Remove this figure? Poor resolution from Naomi Choppra.)

Caltech ACN consists of two three-phase stepdown transformers in ΔY configuration

with Δ on the primary side. Each of these transformers is connected to an electric panel, to which charging stations and subpanels are connected. Figure 3.38(a) shows the two three-phase transformers and the two electric panels. Figure 3.38(b) shows the



(a) Transformers and panels



(b) Transformer ratings

Figure 3.38 (a) The two 150 kVA transformers and two electric panels in Caltech ACN to which charging stations and electric subpanels are connected. (b) The transformer ratings.

ratings of each of the three-phase transformers:

- Power rating 150 kVA (three-phase).
- Primary (high voltage) side: 480V in Delta configuration with rated line current of 180A.

- Secondary (low voltage) side: 208Y/120V in Wye configuration with rated line current of 416A.
- Impedance voltage (percentage impedance): $\beta = 5.45\%$ on the primary side (the shunt admittance is negligible).

The impedance voltage is the voltage drop across the series impedance Z_l on the primary side of the transformer in a short-circuit test, as a percentage of the rated primary voltage. In a short-circuit test the secondary side is short-circuited. The β specification means that the voltage needed on the primary side to produce a rated primary current is β times the rated primary voltage.

Verify that the rated line currents on the primary and secondary sides are consistent with the power rating and voltage ratings. Determine the magnitude $|Z_l|$ of the series impedance of the transformer and draw the circuit model of the three-phase transformer.

Exercise 3.17 (Caltech ACN: estimating distribution line impedances)**.** Suppose the transformer in Exercise 3.16 is connected to a three-phase voltage source with a line voltage of $|V_{\text{line}}| = 480V$ on the primary side through a three-phase distribution line modeled by a series impedance $Z_{\text{line},1}$, and to a three-phase load on the secondary side through another three-phase distribution line modeled by a series impedance $Z_{\text{line},2}$,

as shown in Figure 3.39. Suppose the system is balanced. The load is a three-phase

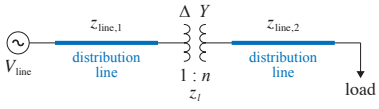


Figure 3.39 The three-phase transformer is connected to a three-phase voltage source and a three-phase load through two three-phase lines.

constant-current load in Δ configuration with a known current I_{load} from phase a to phase b. The voltage is measured to be V_2 across the load between phase a and phase b. The phase a voltage on the secondary side of the transformer (before the distribution line) is measured to be V_{an}.

Determine the distribution line impedances $Z_{\text{line},1}$ and $Z_{\text{line},2}$ in terms of the line voltage $|V_{\text{line}}|$, the series impedance Z_l of the transformer, and the complex gain $K(n)$ of the ideal ΔY transformer, as well as the measured voltages V_2, V_{an} and current I_{load}. Assume without loss of generality that the voltage source has $V_{ab} = |V_{\text{line}}|\angle 0°$ and the sources are in positive sequence.

Exercise 3.18 (Caltech ACN: network design)**.** This problem considers the deployment costs of different network designs for ACN. Referring to Figure 3.38(a), the output (secondary side) of each of the 150 KVA transformers is connected to the input of one of the two electric panels. A wire connects a circuit breaker in the panel to an electric vehicle (EV) charger or a subpanel and these wires are housed in conduits. We consider the network that connects all the EV chargers to one of the two panels in Figure 3.38(a). In this network, the main components are wires, conduits, and subpanels and the types and sizes of these hardware determine the deployment costs, both parts and labor. The types and sizes depend on the current limit (ampacity) of each wire segment required to carry the current to chargers it supplies and the distance of that wire segment. Consider an *idealized* layout in Figure 3.40 where the network connects a total of nk EV chargers to the electric panel. These chargers are clustered into n groups. Each group i is associated with a junction $i = 1, \ldots, n$ as shown in the figure. Every group consists of k identical chargers labeled by $EV_1, \ldots, EV_k$. Each charger can draw a maximum current of I (in A).

Design 1. The first design runs a wire from the electric panel at junction 0 directly to each charger following the path labeled in black in Figure 3.40(a). Let (D, A_i) denote the distance and the cross-sectional area of the wire between each junction $i-1$ to i. Let (d, a) denote the distance and the cross-sectional area of the wire from a junction to every EV in its group. The cross-sectional area of a wire depends on the maximum current it needs to supply. We assume the maximum current that can be drawn by any charger is the same, and therefore the wires from a junction to any EV in its group all

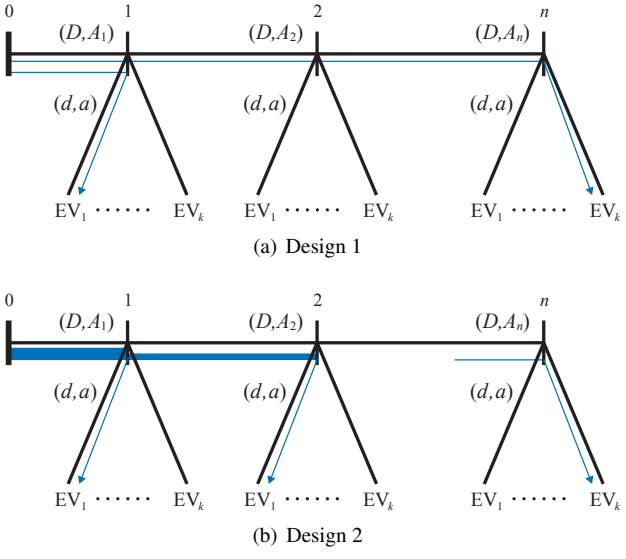


Figure 3.40 Caltech ACN network design.

have the same size a. The wire size A_i between junctions $i-1$ and i depends on the layout. In design 1, $A_i = a$ for all i. This will be different in design 2 (see below).

For example, the wire connecting EV_1 in group 1 goes from junction 0 (electric panel) to junction 1 to the charger, as shown in blue, and has a total length of $D+d$ and size a. The wire connecting EV_k in group n goes from junction 0 to junctions $1,\ldots,n$, to the charger, and has a total length of $nD+d$ and size a.

Design 2. In this design a single wire of length D and size A_1 connects the electric panel at junction 0 to an electric subpanel at junction 1; see Figure 3.40(b). Then k wires each of length d and size a connects the k chargers in group 1 to the subpanel. A single wire of length D and size $A_2 < A_1$ connects the subpanel at junction 1 to a subpanel at junction 2, and k wires each of (d,a) then connects the k chargers in group 2, and so on.

For both design 1 and design 2, the cross-sectional area of the wire used for any segment of the layout depends on the maximum current (called the *ampacity* of the wire in ampere) that it needs to carry. That is, the wire sizes a, A_i above are functions $\alpha(x)$ where x is the ampacity. See below for an example of $\alpha(x)$.

Deployment costs. The total deployment cost (parts and labor) involve mainly three types of hardware.

1 *Wire*. The cost of deploying a wire of length λ and cross-sectional area α is denoted by the function $C_w(\lambda,\alpha)$.

2 *Conduit.* The cost $C_c(\lambda,\alpha)$ of deploying a conduit of length λ that carries wires with a *total* cross-sectional areas α has two components:

$$C_c(\lambda,\alpha) := C_{c1}(\lambda,\alpha) + C_{c2}(\alpha)$$

The first component $C_{c1}(\lambda,\alpha)$ depends on the length λ and total wire size α, the longer and larger the conduit, the higher the cost. The second component $C_{c2}(\alpha)$ depends only on the total wire size α and is usually a step function: when the total wire size exceeds a threshold, a special machine is needed to deploy the conduit at an extra cost. In Design 1, all wires that share the same segment (say) between junctions $i-1$ to i will be housed in the same conduit. For example, the conduit between junction 1 and junction 2 will carry $(n-1)k$ wires. We assume that if a conduit carries wires of areas $\alpha_1,\ldots,\alpha_m$, then the total wire size is simply its sum $\alpha := \sum_{i=1}^m \alpha_i$.

3 *Subpanel.* For simplicity we assume every subpanel (in design 2) has the same cost c_s.

Assumptions on cost functions. Assume the cost functions take the following form:

$$C_w(\lambda,\alpha) := c_w\,\lambda\,\alpha, \qquad C_{c1}(\lambda,\alpha) := c_c\,\lambda\,\alpha, \qquad C_{c2}(\alpha) = \beta\mathbf{1}(\alpha \geq \alpha_0) \tag{3.27a}$$

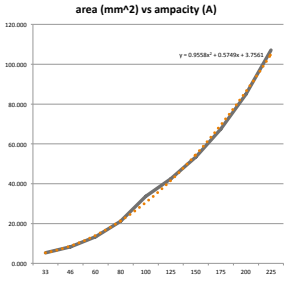
Figure 3.41(a) shows the wire size dependence $\alpha(x)$ on ampacity x from (a version of) the American Wire Gauge (AWG) standard. Based on the data, Figure 3.41(b) shows that $\alpha(x)$ can be well approximated by a quadratic function

$$\alpha(x) := x^2 + 0.6x + 4 \tag{3.27b}$$

with x in A and $\alpha(x)$ in mm^2. The quadratic term represents the fact that the thermal

AWG #	area (mm^2)	ampacity (enclosed, A)
10	5.269	33
8	8.347	46
6	13.332	60
4	21.156	80
2	33.593	100
1	42.429	125
0	53.456	150
00	67.491	175
000	84.949	200
0000	107.146	225

(a) AWG table



(b) AWG plot

Figure 3.41 (a) American Wire Gauge (AWG) standard: dependence of wire cross-sectional area $\alpha(x)$ on ampacity x. (b) The data for $\alpha(x)$ in the table can be approximated by the quadratic function in (3.27b). The black solid line is the plot of the data and the orange dashed line is the quadratic fit.

power loss due to a current I_0 through a wire with resistance r is roughly rI_0^2. Doubling the current means that the resistance must be scaled down by a factor of 4 in order

to maintain the same heat loss. Since r is inversely proportional to the cross-sectional area of the wire, this requires a wire with 4 times the area.

1 Evaluate the total cost of network design 1 and design 2.
2 Prove that design 1 is always less expensive than design 2 as long as the maximum current I that can be drawn by a charger is at least 2A.[4]

Exercise 3.19 (Caltech ACN: network design)**.** This problem generalizes problem 3.18 to show that design 2 is more expensive even for very general cost functions and wire size dependency. Suppose the cost functions $C_w(\lambda,\alpha), C_{c1}(\lambda,\alpha), C_{c2}(\alpha)$ and the dependency of wire size $\alpha(x)$ on its ampacity satisfy the following conditions:

C1: For any fixed α, $C_w(\lambda,\alpha)$ is linear in λ. For any fixed λ, $C_w(\lambda,\alpha)$ linear and increasing in α.
C2: $C_{c1}(\lambda,\alpha)$ is increasing in α for any fixed λ. $C_{c2}(\alpha)$ is increasing in α.
C3: There is an ampacity set X such that for all $x \in X$, $\alpha(ix) \geq i\alpha(x)$ for any integer $i \geq 1$.

Prove that design 2 is more expensive for any ampacity $x \in X$.

It can be easily verified that the cost functions and $\alpha(x)$ in (3.27) satisfy these conditions. In particular the ampacity set X in condition C3 is $X = \{x \geq 2\text{A}\}$. Therefore the conditions C1–C3 allow a much larger set of cost functions and $\alpha(x)$ than (3.27).

We now interpret these conditions to illustrate that they are realistic. Condition C1 says that the total deployment cost (parts and labor) grows linearly in wire length λ and in wire size α. If either one doubles, the cost exactly doubles. Condition C2 says that regardless of its length, both the first and second cost components of the conduit increase as the cross-sectional area of the conduit increases. Finally condition C3 implies in particular that, for any ampacity x in X, doubling the ampacity more than doubles the cost. As explained immediately after (3.27b), since thermal loss is quadratic in ampacity, the required wire size satisfies this condition. The proof reveals that this is the key condition that makes design 2 more expensive than design 1, i.e., it is always cheaper to use more and longer small wires because the wire size *grows faster than linearly in ampacity*.

Exercise 3.20 (Caltech ACN: network design)**.** Problem 3.19 shows that, under very general and realistic conditions, design 2 is always more expensive than design 1. This assumes that, in design 2, the ampacity of the wire between junction $i-1$ and i must be the sum of the ampacities of all the downstream wires supplying groups $i, i+1, \ldots, n$. In

[4] Currently a level-2 EV charger typically has a current limit of 32A or higher.

practice however it is unlikely all the EV chargers in these groups will draw maximum currents simultaneously and therefore it is reasonable to install a smaller ampacity between junction $i-1$ and i, i.e., each subpanel can be over-subscribed. Discuss over-subscription conditions under which design 2 is less expensive than design 1 (open-ended problem).

4 Bus injection models

In previous chapters we introduce mathematical models of basic power system components. In this and the next chapter we use these component models to describe a power network consisting of an interconnection of components such as generators, loads, transmission and distribution lines, and transformers. In Chapter 4.1 we summarize the component models from previous chapters. In Chapter 4.2 we explain how to model a power network by a matrix that linearly relates nodal current injections to nodal voltages of the network. In Chapter 4.3 we present power flow equations that relate nodal power injections and nodal voltages. In Chapter 4.4 we discuss classical solution methods. In Chapter 4.6 we study a linearized model, called the DC power flow model, that is widely used in power systems applications such as electricity markets.

4.1 Component models

sV

The component models summarized in this section will be used to construct network models in Chapters 4.2 and 4.3.

4.1.1 Single-phase sources and impedance

In Chapters 1.1.2 and 1.3.1 we describe circuit models of single-phase single-terminal devices. They are also per-phase models of balanced three-phase devices. Associated with each device j is its terminal voltage, current, and power $(V_j, I_j, s_j) \in \mathbb{C}^3$. There is an arbitrary *reference point* with respect to which all voltages are defined. If the common reference point is taken to be the ground then voltage V_j is the voltage drop between terminal j and the ground. The current from terminal j flows from the terminal to the reference point (see Figure 4.1). Such a single-terminal device is characterized by relations between the terminal variables (V_j, I_j, s_j).

1 *Voltage source* (E_j, z_j). This is a device with a constant internal voltage E_j in

series with an impedance z_j as shown in Figure 1.3(a). Its external model is the relation $V_j = E_j - z_j I_j$ between its terminal voltage and current (V_j, I_j). This yields a relation $s_j = V_j I_j^{\mathsf{H}} = V_j \left(E_j - V_j\right)^{\mathsf{H}} / z_j^{\mathsf{H}}$ between the terminal variables (V_j, s_j).

2 *Current source* (J_j, y_j). This is a device with a constant internal current J_j in parallel with an admittance y_j as shown in Figure 1.3(b). Its external model is the relation $I_j = J_j - y_j V_j$ between its terminal voltage and current (V_j, I_j). This yields a relation $s_j = V_j I_j^{\mathsf{H}} = V_j \left(J_j - y_j V_j\right)^{\mathsf{H}}$ between the terminal variables (V_j, s_j).

3 *Power source* (σ_j, z_j). This is a device with a constant internal power σ_j in series with an impedance z_j. Its external model is the relation $\sigma_j = \left(V_j - z_j I_j\right) I_j^{\mathsf{H}}$ between (V_j, I_j). Its terminal power is $s_j = V_j I_j^{\mathsf{H}} = \sigma_j + z_j |I_j|^2$.

4 *Impedance* z_j. The external (and internal) model is $V_j = z_j I_j$ and $s_j = \left|V_j\right|^2 / z_j^{\mathsf{H}}$.

We often assume the voltage, current, or power sources are ideal in which case z_j and y_j are zero.

4.1.2 Single-phase line

In Chapter 2.2.2 we describe the Π circuit model of a single-phase transmission or distribution line. It is also a per-phase model of balanced three-phase lines. A line has two terminals (j,k) and is specified by a three-tuple $(y_{jk}^s, y_{jk}^m, y_{kj}^m) \in \mathbb{C}^3$ where $y_{jk}^s = y_{kj}^s$ is the series admittance of the line, y_{jk}^m is the shunt admittance of the line at terminal j, and y_{kj}^m is the shunt admittance of the line at terminal k; see Figure 4.1. Recall that (y_{jk}^m, y_{kj}^m) models the line capacitance, called *line charging* or *shunt*

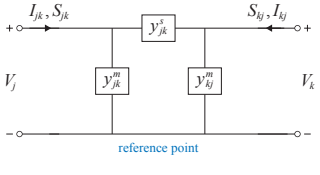


Figure 4.1 Π circuit model of a single-phase line.

admittances of line (j,k), and the currents through these shunt admittances model the current supplied to the line capacitance called the charging current.

Associated with terminal j is the terminal voltage V_j, and the sending-end line current I_{jk} and power S_{jk} from j to k. Similarly, associated with terminal k is $(V_k, I_{kj}, S_{kj}) \in \mathbb{C}^3$. Unlike in Chapter 2.2.2 we have defined here I_{kj} to be the current injected from the right terminal into the line. A line is characterized by the relation between the terminal voltages (V_j, V_k) and line currents $\left(I_{jk}, I_{kj}\right)$ or that between (V_j, V_k) and line powers $\left(S_{jk}, S_{kj}\right)$, which we now explain.

IV relation.

The terminal voltages with respect to, and the sending-end currents flowing from the terminals to, the reference point are related by

$$I_{jk} = y^s_{jk}(V_j - V_k) + y^m_{jk} V_j, \qquad I_{kj} = y^s_{kj}(V_k - V_j) + y^m_{kj} V_k \tag{4.1a}$$

This defines a matrix Y_{line} for a line that maps terminal voltages to sending-end currents:

$$\begin{bmatrix} I_{jk} \\ I_{kj} \end{bmatrix} = \underbrace{\begin{bmatrix} y^s_{jk} + y^m_{jk} & -y^s_{jk} \\ -y^s_{jk} & y^s_{jk} + y^m_{kj} \end{bmatrix}}_{Y_{\text{line}}} \begin{bmatrix} V_j \\ V_k \end{bmatrix} \tag{4.1b}$$

where we have used the fact that $y^s_{jk} = y^s_{kj}$ to obtain a symmetric Y_{line}. The off-diagonal entries of Y_{line} are the negatives of the series admittances while the diagonal entries are the sum of series and shunt admittances. As we will see this structure holds for general networks.

In general the sending-end currents (I_{jk}, I_{kj}) are not negative of each other when the shunt admittances are nonzero. Since $y^s_{jk} = y^s_{kj}$, their sum represents the total current loss along the line due to shunt admittances:

$$I_{jk} + I_{kj} = y^m_{jk} V_j + y^m_{kj} V_k \neq 0$$

Thermal limits on branch current flows should be imposed on both $|I_{jk}|$ and $|I_{kj}|$:

$$|I_{jk}| = \left| y^s_{jk}(V_j - V_k) + y^m_{jk} V_j \right| \leq I^{\max}_{jk}$$
$$|I_{kj}| = \left| y^s_{kj}(V_k - V_j) + y^m_{kj} V_k \right| \leq I^{\max}_{kj}$$

not just on $\left| y^s_{jk}(V_j - V_k) \right|$ unless the shunt admittances are zero.

sV relation.

The sending-end line power flows from terminals j to k and that from terminals k to j are respectively (using (4.1a)):

$$S_{jk} := V_j I^{\mathsf{H}}_{jk} = \left(y^s_{jk}\right)^{\mathsf{H}} \left(|V_j|^2 - V_j V^{\mathsf{H}}_k\right) + \left(y^m_{jk}\right)^{\mathsf{H}} |V_j|^2 \tag{4.2a}$$

$$S_{kj} := V_k I^{\mathsf{H}}_{kj} = \left(y^s_{kj}\right)^{\mathsf{H}} \left(|V_k|^2 - V_k V^{\mathsf{H}}_j\right) + \left(y^m_{kj}\right)^{\mathsf{H}} |V_k|^2 \tag{4.2b}$$

They are not negatives of each other because of power loss along the line. Since $y^s_{jk} = y^s_{kj}$, the total complex power loss is:

$$S_{jk} + S_{kj} = \left(y^s_{jk}\right)^{\mathsf{H}} |V_j - V_k|^2 + \left(y^m_{jk}\right)^{\mathsf{H}} |V_j|^2 + \left(y^m_{kj}\right)^{\mathsf{H}} |V_k|^2 \tag{4.3}$$

The first term on the right-hand side is loss due to series impedance and the last two terms are losses due to shunt admittances of the line. Thermal limits on branch power

flows should be imposed on both $|S_{jk}|$ and $|S_{kj}|$:

$$|S_{jk}| = \left|\left(y^s_{jk}\right)^{\mathsf{H}}\left(|V_j|^2 - V_jV_k^{\mathsf{H}}\right) + \left(y^m_{jk}\right)^{\mathsf{H}}|V_j|^2\right| \le S^{\max}_{jk}$$
$$|S_{kj}| = \left|\left(y^s_{kj}\right)^{\mathsf{H}}\left(|V_k|^2 - V_kV_j^{\mathsf{H}}\right) + \left(y^m_{kj}\right)^{\mathsf{H}}|V_k|^2\right| \le S^{\max}_{kj}$$

not just on $\left|\left(y^s_{jk}\right)^{\mathsf{H}}\left(|V_j|^2 - V_jV_k^{\mathsf{H}}\right)\right|$ and $\left|\left(y^s_{kj}\right)^{\mathsf{H}}\left(|V_k|^2 - V_kV_j^{\mathsf{H}}\right)\right|$ unless the shunt admittances are zero.

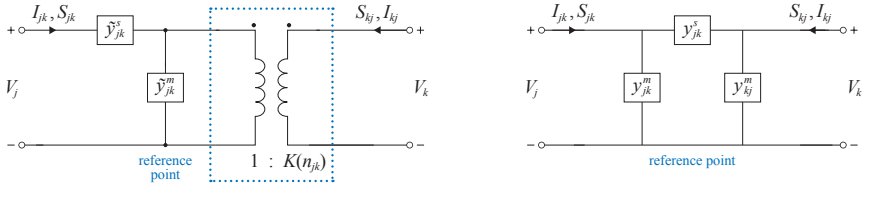
If the shunt admittances y^m_{jk} and y^m_{kj} of the line are zero then the power loss has a simple relation with line currents. Setting $y^m_{jk} = y^m_{kj} = 0$ in (4.3) and (4.1a) and using $y^s_{jk} = y^s_{kj}$, we have

$$S_{jk} + S_{kj} = z^s_{jk} \cdot \left|y^s_{jk}\right|^2 \left|V_j - V_k\right|^2 = z^s_{jk}\left|I_{jk}\right|^2$$

because $I_{jk} = y^s_{jk}(V_j - V_k) = -I_{kj}$ when the shunt elements are zero and $y^s_{jk} = y^s_{kj}$. This is not the case otherwise.

4.1.3 Single-phase transformer

In Chapters 3.1 and 3.2 we describe circuit models of a single-phase transformer. They are also per-phase models of balanced three-phase transformers. A transformer has two terminals (j,k) and is specified by its voltage gain n_{jk} which is the reciprocal of the turns ratio $a_{jk} := 1/n_{jk}$. If the single-phase transformer is the per-phase model of a balanced three-phase transformer, then the voltage gain $K(n_{jk})$ can be complex, e.g., $K(n_{jk}) = \sqrt{3}n_{jk}\,e^{\mathbf{i}\pi/6}$ for ΔY configuration. In addition to the voltage gain n_{jk}, a single-phase transformer also has series resistance and leakage inductance and shunt admittance due to the primary and secondary magnetizing currents. These effects can be modeled by a series admittance $\tilde{y}^s_{jk}$ and shunt admittance $\tilde{y}^m_{jk}$ in the primary circuit, as shown in Figure 4.2(a).



(a) Non-ideal transformer (b) Π circuit model (real n_{jk})

Figure 4.2 Single-phase transformer.

As for a line model, associated with terminals j and k are the terminal voltages,

sending-end line currents and sending-end line power flows $(V_j, I_{jk}, S_{jk}) \in \mathbb{C}^3$ and $(V_k, I_{kj}, S_{kj}) \in \mathbb{C}^3$ respectively. Notice that the direction of I_{kj} at terminal k is opposite to that in Chapter 3. The behavior of the transformer in Figure 4.2 is characterized by the relation between the terminal voltages (V_j, V_k) and line currents (I_{jk}, I_{kj}) or that between (V_j, V_k) and line powers (S_{jk}, S_{kj}), which we now summarize.

Real voltage gain $K(n_{jk}) = n_{jk}$.

Using Kirchhoff's and Ohm's laws and transformer gains we have

$$I_{jk} = \tilde{y}^s_{jk}\left(V_j - a_{jk}V_k\right), \qquad I_{jk} = \tilde{y}^m_{jk}\, a_{jk}V_k + n_{jk}(-I_{kj}) \tag{4.4a}$$

where $a_{jk} := 1/n_{jk}$. This defines a matrix $Y_{\text{transformer}}$ for the single-phase transformer:

$$\begin{bmatrix} I_{jk} \\ I_{kj} \end{bmatrix} = \underbrace{\begin{bmatrix} \tilde{y}^s_{jk} & -a_{jk}\,\tilde{y}^s_{jk} \\ -a_{jk}\,\tilde{y}^s_{jk} & a^2_{jk}\left(\tilde{y}^s_{jk} + \tilde{y}^m_{jk}\right) \end{bmatrix}}_{Y_{\text{transformer}}} \begin{bmatrix} V_j \\ V_k \end{bmatrix} \tag{4.4b}$$

Since transformer gains n_{jk} are real, $Y_{\text{transformer}}$ is symmetric and their terminal behavior can be modeled by a Π circuit, the same way a transmission line is. Specifically $Y_{\text{transformer}}$ can be rewritten in terms of admittances $\left(y^s_{jk}, y^m_{jk}, y^m_{kj}\right)$ of a Π circuit:

$$Y_{\text{transformer}} := \begin{bmatrix} y^s_{jk} + y^m_{jk} & -y^s_{jk} \\ -y^s_{jk} & y^s_{jk} + y^m_{kj} \end{bmatrix} \tag{4.5a}$$

where

$$y^s_{jk} := a_{jk}\,\tilde{y}^s_{jk}, \quad y^m_{jk} := (1 - a_{jk})\tilde{y}^s_{jk}, \quad y^m_{kj} := a_{jk}(a_{jk} - 1)\tilde{y}^s_{jk} + a^2_{jk}\,\tilde{y}^m_{jk} \tag{4.5b}$$

as illustrated in Figure 4.2(b). In particular the shunt admittances y^m_{jk} and y^m_{kj} of the Π circuit model are different unless $(1 - a^2)\tilde{y}^s_{jk} = a^2\tilde{y}^m_{jk}$. Moreover (y^m_{jk}, y^m_{kj}) are generally nonzero even if the transformer shunt admittance $\tilde{y}^m_{jk} = 0$.

Complex voltage gain $K(n)$.

A physical transformer always has a real voltage gain n. The per-phase model of three-phase transformer in a balanced setting however can have a complex voltage gain $K(n)$ as we have seen in Chapter 3.2. In that case $-nI_{kj}$ in the above derivation should be replaced by $K^{\mathsf{H}}_{jk}(n)I_{kj}$, leading to:

$$\begin{bmatrix} I_{jk} \\ I_{kj} \end{bmatrix} = \underbrace{\begin{bmatrix} \tilde{y}^s_{jk} & -\tilde{y}^s_{jk}/K_{jk}(n) \\ -\tilde{y}^s_{jk}/\bar{K}_{jk}(n) & \left(\tilde{y}^s_{jk} + \tilde{y}^m_{jk}\right)/|K_{jk}(n)|^2 \end{bmatrix}}_{Y_{\text{transformer}}} \begin{bmatrix} V_j \\ V_k \end{bmatrix}$$

In this case the matrix $Y_{\text{transformer}}$ is *not* symmetric. This means that the terminal behavior of the transformer does not have an equivalent Π circuit model and we have to use the admittance matrix $Y_{\text{transformer}}$ for power flow analysis. In this case the transformer is characterized by two pairs of admittances, (y^s_{jk}, y^m_{jk}) from j to k and (y^s_{kj}, y^m_{kj}) in the opposite direction, defined by transformer parameters $\left(K(n), \tilde{y}^s_{jk}, \tilde{y}^m_{jk}\right)$. Equivalently, the admittance matrix $Y_{\text{transformer}}$ is not symmetric and takes the form:

$$Y_{\text{transformer}} := \begin{bmatrix} y^s_{jk} + y^m_{jk} & -y^s_{jk} \\ -y^s_{kj} & y^s_{kj} + y^m_{kj} \end{bmatrix} \tag{4.6a}$$

where

$$y^s_{jk} := \frac{\tilde{y}^s_{jk}}{K_{jk}(n)}, \qquad y^m_{jk} := \left(1 - \frac{1}{K_{jk}(n)}\right)\tilde{y}^s_{jk} \tag{4.6b}$$

$$y^s_{kj} := \frac{\tilde{y}^s_{jk}}{\bar{K}_{jk}(n)}, \qquad y^m_{kj} := \frac{1 - K_{jk}(n)}{|K_{jk}(n)|^2}\tilde{y}^s_{jk} + \frac{1}{|K_{jk}(n)|^2}\tilde{y}^m_{jk} \tag{4.6c}$$

The admittance matrix reduces to (4.5) when $K(n) = n$. The relation between powers (S_{jk}, S_{kj}) and voltages (V_j, V_k) is the same as for transmission and distribution lines, even though y^s_{jk} and y^s_{kj} may not be equal:

$$\begin{aligned} S_{jk} := V_j I^{\mathsf{H}}_{jk} &= \left(y^s_{jk}\right)^{\mathsf{H}} \left(|V_j|^2 - V_j V^{\mathsf{H}}_k\right) + \left(y^m_{jk}\right)^{\mathsf{H}} |V_j|^2 \\ S_{kj} := V_k I^{\mathsf{H}}_{kj} &= \left(y^s_{kj}\right)^{\mathsf{H}} \left(|V_k|^2 - V_k V^{\mathsf{H}}_j\right) + \left(y^m_{kj}\right)^{\mathsf{H}} |V_k|^2 \end{aligned}$$

where the admittances are given in (4.6). If $y^s_{jk} \neq y^s_{kj}$ then the line loss is not given by (4.3).

4.2 Network model: IV relation

In this section we explain how to use the component models of Chapter 4.1 to model a single-phase network consisting of generators and loads connected by a network of transmission or distribution lines and transformers. We will construct an equivalent circuit consisting of *ideal* voltage and current sources connected by a network of series and shunt admittances. The nodal current injections I are linearly related to nodal voltages V through a matrix Y called an admittance matrix, $I = YV$. This relation represents the Kirchhoff's laws and the Ohm's law. In this section we derive the admittance matrix Y and study its properties.

We start in Chapter 4.2.1 with a few examples and present in Chapter 4.2.2 our abstract line model. In Chapter 4.2.3 we define the admittance matrix Y for a general network and study sufficient conditions for the invertibility of Y. In Chapter 4.2.4.1 we explain Kron reduction of an admittance matrix Y and study the invertibility of a Kron-reduced admittance matrix. In Chapter 4.2.5 we present a common method for

solving $I = YV$ numerically. When the network graph is a tree, called a radial network, a reduced admittance matrix is always invertible and we derive explicitly its inverse in Chapter 4.2.6.

4.2.1 Examples

In this subsection we derive the admittance matrix Y of a single-phase network shown in Figure 4.3 where:

1 The generator on the left end is modeled as a current source with parameters (I_1, y_1).
2 The non-ideal single-phase transformer has a real voltage gain n, a series admittance $\tilde{y}^s$ and shunt admittance $\tilde{y}^m$ in the primary circuit.
3 The transmission line is modeled as a series admittance y (and zero shunt admittances).
4 The motor load on the right end is modeled as another current source (I_2, y_2).

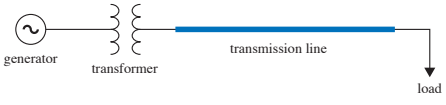


Figure 4.3 One-line diagram of a generator supplying a load through a transformer and a transmission line.

We will derive the admittance matrix Y for the overall system in two steps.

Example 4.1 (Non-ideal transformer and transmission line)**.** Figure 4.4 shows the circuit model of the non-ideal transformer in series with the transmission line. To

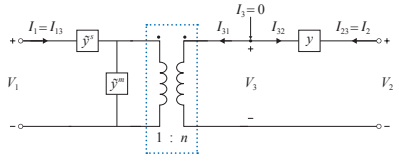


Figure 4.4 A non-ideal transformer in series with a transmission line.

determine the admittance matrix that relates (I_1, I_2) to (V_1, V_2), we introduce an additional network node 3 between the transformer and the transmission line y with an auxiliary voltage V_3 and an auxiliary injection current I_3 at node 3, as shown in the figure.

Since the voltage gain n is real, use the transformer model (4.4b) and the line model

(4.1) to get

$$\begin{bmatrix} I_{13} \\ I_{31} \end{bmatrix} = \begin{bmatrix} \tilde{y}^s & -a\,\tilde{y}^s \\ -a\,\tilde{y}^s & a^2\,(\tilde{y}^s + \tilde{y}^m) \end{bmatrix} \begin{bmatrix} V_1 \\ V_3 \end{bmatrix}, \qquad \begin{bmatrix} I_{32} \\ I_{23} \end{bmatrix} = \begin{bmatrix} y & -y \\ -y & y \end{bmatrix} \begin{bmatrix} V_3 \\ V_2 \end{bmatrix}$$

Kirchhoff's current law at each node gives:

$$I_1 = I_{13}, \qquad 0 = I_3 = I_{31} + I_{32}, \qquad I_2 = I_{23}$$

Eliminating branch currents relates nodal currents (I_1, I_2, I_3) to nodal voltages (V_1, V_2, V_3) through matrix Y_1:

$$\begin{bmatrix} I_1 \\ I_2 \\ I_3 \end{bmatrix} = \underbrace{\begin{bmatrix} \tilde{y}^s & 0 & -a\tilde{y}^s \\ 0 & y & -y \\ -a\tilde{y}^s & -y & y + a^2\,(\tilde{y}^s + \tilde{y}^m) \end{bmatrix}}_{Y_1} \begin{bmatrix} V_1 \\ V_2 \\ V_3 \end{bmatrix} \tag{4.7}$$

The matrix Y_1 is complex symmetric and is therefore an admittance matrix that can be represented as a Π circuit as shown in Figure 4.5 where $y^s_{13} := a\tilde{y}^s$, $y^m_{13} := (1-a)\tilde{y}^s$

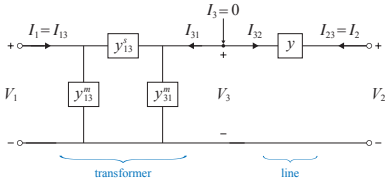


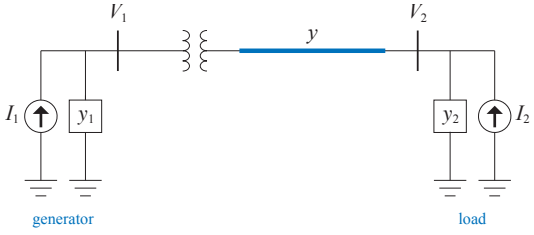
Figure 4.5 Π circuit model of the system in Figure 4.4.

and $y^m_{31} := a(a-1)\tilde{y}^s + a^2\tilde{y}^m$. □

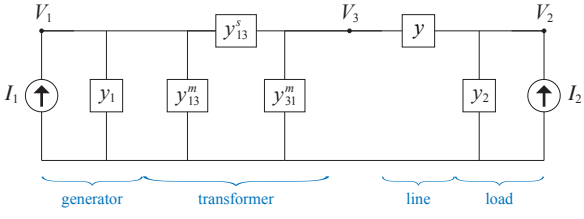
Example 4.2 (Overall system)**.** Finally the circuit model of the overall system that includes the two current sources that model the generator and the load is shown in Figure 4.6(a). The only changes to the admittance matrix, compared with the admittance matrix Y_1 in (4.7), are the additional shunt admittances y_1, y_2 at nodes 1 and 2 respectively. They should be added to the first two diagonal entries of Y_1. The overall network can therefore be modeled by an admittance matrix Y that relates nodal current injections and nodal voltages (setting $I_3 = 0$):

$$\begin{bmatrix} I_1 \\ I_2 \\ 0 \end{bmatrix} = \begin{bmatrix} \tilde{y}^s + y_1 & 0 & -a\tilde{y}^s \\ 0 & y + y_2 & -y \\ -a\tilde{y}^s & -y & y + a^2\,(\tilde{y}^s + \tilde{y}^m) \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \\ V_3 \end{bmatrix}$$

The external behavior can be modeled by a 2×2 admittance matrix that relates (I_1, I_2) and (V_1, V_2) which can be obtained from Y through Kron reduction making use of the fact that the internal injection $I_3 = 0$; see Chapter 4.2.4.1. □



(a) One-line diagram



(b) Equivalent circuit model

Figure 4.6 Generator, transformer, transmission line and load.

4.2.2 Line model

In general we model a power network by a connected undirected graph $G = (\overline{N}, E)$ of $N+1$ nodes and M lines, where $\overline{N} := \{0\} \cup N$, $N := \{1, 2, \ldots, N\}$ and $E \subseteq \overline{N} \times \overline{N}$. Each node j in $\overline{N}$ may represent a bus and each edge (j,k) in E may represent a transmission or distribution line or transformer. We also write $j \sim k$ instead of $(j,k) \in E$. We use "bus, node, terminal" interchangeably and "line, branch, link, edge" interchangeably.

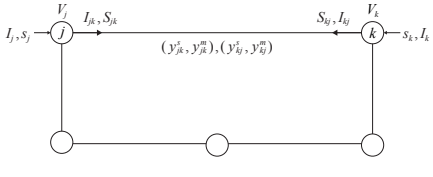


Figure 4.7 Network graph and notations.

For each line $(j,k) \in E$ let (V_j, V_k) denote the terminal (or nodal) voltages at each end of the line. Let I_{jk} denote the sending-end line current from j to k and I_{kj} the sending-end line current from k to j. Each line $(j,k) \in E$ is characterized by four admittances $\left(y^s_{jk}, y^m_{jk}\right) \in \mathbb{C}^2$ from j to k and $\left(y^s_{kj}, y^m_{kj}\right) \in \mathbb{C}^2$ from k to j; see Figure 4.7. We call $\left(y^s_{jk}, y^s_{jk}\right)$ the *series admittances* and $\left(y^m_{jk}, y^m_{jk}\right)$ the *shunt admittances* of

line (j,k). They define the relation between (V_j, V_k) and (I_{jk}, I_{kj}):

$$I_{jk} = y^s_{jk}(V_j - V_k) + y^m_{jk}V_j, \qquad I_{kj} = y^s_{kj}(V_k - V_j) + y^m_{kj}V_k \tag{4.8a}$$

or in matrix form:

$$\begin{bmatrix} I_{jk} \\ I_{kj} \end{bmatrix} = \underbrace{\begin{bmatrix} y^s_{jk} + y^m_{jk} & -y^s_{jk} \\ -y^s_{kj} & y^s_{kj} + y^m_{kj} \end{bmatrix}}_{Y_{jk}} \begin{bmatrix} V_j \\ V_k \end{bmatrix} \tag{4.8b}$$

We emphasize that the series admittances y^s_{jk} and y^s_{kj} may be different and therefore this general model may not have a Π circuit representation. It can model a single-phase transmission or distribution line, a single-phase transformer, or the per-phase model of a balanced three-phase transformer with a real or complex voltage gain, as summarized in Chapters 4.1.2 and 4.1.3. Specifically when (j,k) models a transmission or distribution line, the line parameters $(y^s_{jk} = y^s_{kj}, y^m_{jk}, y^m_{kj})$ are the series and shunt admittances of the transmission or distribution line. When (j,k) models a transformer, the line parameters (y^s_{jk}, y^m_{jk}) and (y^s_{kj}, y^m_{kj}) are given by (4.6) in terms of transformer voltage gain and leakage and shunt admittances $(K(n), \tilde{y}^s_{jk}, \tilde{y}^m_{jk})$. Note that y^s_{kj} and y^s_{jk} may be different, and (y^m_{jk}, y^m_{kj}) are generally different and nonzero even if the transformer shunt admittance $\tilde{y}^m_{jk} = 0$. When the voltage gain $K(n) = n$ is real, (4.6) reduces to (4.5) with $y^s_{kj} = y^s_{jk}$. As we have seen in Example 4.2, a line (j,k) in the graph G, the matrix Y_{jk} may also contain generator and load impedances.

We will often restrict ourselves to the special case where the following assumption holds:

C4.1: The series admittances $y^s_{jk} = y^s_{kj}$ for every line $(j,k) \in E$.

In this case (4.8) reduces to

$$I_{jk} = y^s_{jk}(V_j - V_k) + y^m_{jk}V_j, \qquad I_{kj} = y^s_{jk}(V_k - V_j) + y^m_{kj}V_k \tag{4.9a}$$

or in terms of Y_{jk}:

$$\begin{bmatrix} I_{jk} \\ I_{kj} \end{bmatrix} = \underbrace{\begin{bmatrix} y^s_{jk} + y^m_{jk} & -y^s_{jk} \\ -y^s_{jk} & y^s_{jk} + y^m_{kj} \end{bmatrix}}_{Y_{jk}} \begin{bmatrix} V_j \\ V_k \end{bmatrix} \tag{4.9b}$$

Since Y_{jk} is symmetric, it has a Π circuit representation and behaves like a transmission or distribution line (though with generally different y^m_{jk} and y^m_{kj}). We characterize such a line by three admittances $\left(y^s_{jk}, y^m_{jk}, y^m_{kj}\right)$. As noted above, this model cannot be used as the per-phase model of a balanced three-phase transformer in ΔY or $Y\Delta$ configuration that has a complex voltage gain $K(n)$. It is however widely applicable, e.g., when the network does not contain transformers with complex voltage gains or when used in per unit systems where (nominal) transformers disappear. We therefore often adopt this model and will explicitly state it as assumption C4.1 when we use it.

4.2.3 Admittance matrix Y and its properties

In bus injection models we are interested in nodal variables $(V_j, I_j, s_j) \in \mathbb{C}^3$, $j \in \overline{N}$, where V_j is the complex voltage at bus j with respect to an arbitrary but fixed common reference point, e.g., the ground. Here I_j and s_j are the complex nodal current and power injections respectively into the network at bus j. These nodal variables are related by $s_j = V_j I_j^{\mathsf{H}}$ for each bus $j \in \overline{N}$. As mentioned above the current and power injections can be interpreted as flowing from terminal j to the common reference point in the circuit model. In this section we construct the admittance matrix Y that linearly relates nodal voltages V to nodal current injections I and study its properties.

4.2.3.1 Admittance matrix Y

The nodal current injections $I := (I_j, j \in \overline{N})$ and voltages $V := (V_j, j \in \overline{N})$ are linearly related. The admittance matrix Y relates, not the line currents, but the *net* nodal current injections I to nodal voltages V. Applying (4.8) to KCL $I_j = \sum_{k:j\sim k} I_{jk}$ at each node j, we have[1]

$$I_j = \sum_{k:j\sim k} I_{jk} \;=\; \left(\sum_{k:j\sim k} y^s_{jk} \,+\, y^m_{jj}\right) V_j \;-\; \sum_{k:j\sim k} y^s_{jk} V_k, \qquad j \in \overline{N} \tag{4.10a}$$

where y^m_{jj} denotes the total shunt admittance of the lines connected to bus j:

$$y^m_{jj} := \sum_{k:j\sim k} y^m_{jk} \tag{4.10b}$$

In vector form, this is $I = YV$ where the matrix Y is given by:

$$Y_{jk} = \begin{cases} -y^s_{jk}, & j \sim k \;\; (j \neq k) \\ \sum_{l:j\sim l} y^s_{jl} \,+\, y^m_{jj}, & j = k \\ 0 & \text{otherwise} \end{cases} \tag{4.10c}$$

We refer to Y that maps nodal voltages to nodal current injections as an *admittance matrix*, or a network admittance matrix or bus admittance matrix. Equation (4.10c) prescribes a way to write down the admittance matrix Y by inspection of the network connectivity and line admittances: its off-diagonal entries are the negatives of series admittances $\left(y^s_{jk}, y^s_{kj}\right)$ in each direction on line (j,k) while its diagonal entries are the sum of the series and shunt admittances incident on the corresponding buses. Note that Y_{jk} and Y_{kj} may not be equal if (j,k) models a transformer. If we restrict ourselves to the special where $y^s_{jk} = y^s_{kj}$ for all $(j,k) \in E$ (assumption C4.1) then each line (j,k) has a Π circuit representation and the admittance matrix Y is complex symmetric. It is not Hermitian unless Y is a real matrix.

Example 4.3. Consider the three-bus network shown in Figure 4.8. Under condition

[1] If there is a load attached to bus j with shunt admittance y^{sh}_j, then the *net* injection becomes $I_j - y^{\mathrm{sh}}_j V_j = \sum_{k:j\sim k} I_{jk}$ instead of I_j on the left-hand side of (4.10a).

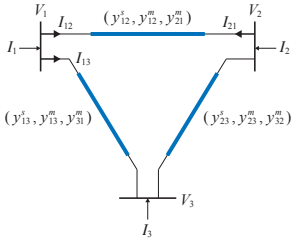


Figure 4.8 Three-bus network of Example 4.3.

C4.1, each line (j,k) is modeled by a Π circuit with series and shunt admittances $\left(y^s_{jk}, y^m_{jk}, y^m_{kj}\right)$. The sending-end branch current from bus j to bus k is I_{jk} and that from bus k to bus j is I_{kj}. Applying Kirchhoff's current law and Ohm's law at bus 1 gives

$$\begin{aligned} I_{12} &= y^s_{12}(V_1 - V_2) + y^m_{12} V_1 \\ I_{13} &= y^s_{13}(V_1 - V_3) + y^m_{13} V_1 \\ \therefore \quad I_1 = I_{12} + I_{13} &= \left(y^s_{12} + y^s_{13} + y^m_{12} + y^m_{13}\right) V_1 - y^s_{12} V_2 - y^s_{13} V_3 \end{aligned}$$

Similarly applying KCL and Ohm's law at buses 2 and 3 we obtain

$$\begin{bmatrix} I_1 \\ I_2 \\ I_3 \end{bmatrix} = \underbrace{\begin{bmatrix} y^s_{12} + y^s_{13} + y^m_{11} & -y^s_{12} & -y^s_{13} \\ -y^s_{12} & y^s_{12} + y^s_{23} + y^m_{22} & -y^s_{23} \\ -y^s_{13} & -y^s_{23} & y^s_{13} + y^s_{23} + y^m_{33} \end{bmatrix}}_{Y} \begin{bmatrix} V_1 \\ V_2 \\ V_3 \end{bmatrix}$$

where

$$y^s_{jk} = y^s_{kj}, \qquad y^m_{jj} := \sum_{k:j\sim k} y^m_{jk}$$

Again the off-diagonal entries of the admittance matrix Y are given by the series admittances on the lines:

$$Y_{jk} := \begin{cases} -y^s_{jk} & \text{if } j \sim k \ \ (j \neq k) \\ 0 & \text{otherwise} \end{cases}$$

and the diagonal entries of Y by the sum of series and shunt admittances incident on buses j:

$$Y_{jj} := \sum_{k:j\sim k} y^s_{jk} + y^m_{jj}$$

□

Under Assumption C4.1, the admittance matrix Y given in (4.10) can also be expressed in terms of more elementary matrices. Fix an arbitrary orientation for the

graph $G := (\overline{N}, E)$ so that a line $l = j \to k \in E$ is now considered pointing from bus j to bus k. Let $C \in \{-1, 0, 1\}^{|\overline{N}| \times |E|}$ be the bus-by-line incidence matrix defined by:

$$C_{jl} = \begin{cases} 1 & \text{if } l = j \to k \text{ for some bus } k \\ -1 & \text{if } l = i \to j \text{ for some bus } i \\ 0 & \text{otherwise} \end{cases} \tag{4.11}$$

Let $D_y^s := \operatorname{diag}\left(y_l^s, l \in E\right)$ be the $|E| \times |E|$ diagonal matrix with the series admittances y_l^s as its diagonal entries. Let $D_y^m := \operatorname{diag}\left(y_{jj}^m, j \in \overline{N}\right)$ be the $|\overline{N}| \times |\overline{N}|$ diagonal matrix with the total shunt admittances y_{jj}^m in (4.10b) as its diagonal entries. Then the admittance matrix in (4.10c) is, when $y_{jk}^s = y_{kj}^s$,

$$Y = C\, D_y^s C^\mathsf{T} + D_y^m \tag{4.12}$$

Clearly the matrix $C D_y^s C^\mathsf{T}$ has zero row and column sums. It verifies that Y is symmetric but not Hermitian unless D_y^s and D_y^m are real matrices. This representation can be used to study the inverse of Y; see Exercise 4.7.

Bus 0 is often called the *slack bus*. Its voltage is fixed and we sometimes assume that $V_0 = 1\angle 0°$ per unit (pu), i.e., the voltage drop between bus 0 and the reference point is $1\angle 0°$. A bus $j \in \overline{N}$ can have a generator, a load, both or neither and (I_j, s_j) are the net current and power injections (generation minus load) at bus j, as the next remark shows.

Remark 4.1 (Nodal devices)**.** Our notation for current injection I_j suggests that there is a single device at each bus j. This simplifies notation and loses no generality. If there are multiple devices connected to bus j, e.g., a non-ideal voltage source $\left(E_j, z_j^v\right)$, a non-ideal current source $\left(J_j, y_j^c\right)$, and a bus shunt admittance y_j^a or equivalently its impedance $z_j^i = \left(y_j^a\right)^{-1}$, as shown in Figure 4.9. then I_j is the net current injection

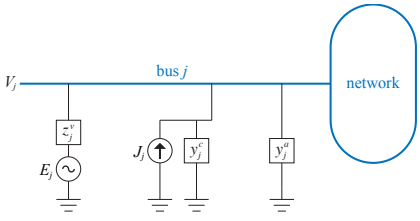


Figure 4.9 Multiple devices connected to the same bus.

from bus j to the rest of the network:

$$I_j = \underbrace{\frac{E_j - V_j}{z_j^v}}_{\text{voltage source}} + \underbrace{\left(J_j - y_j^c V_j\right)}_{\text{current source}} - \underbrace{y_j^a V_j}_{\text{shunt admittance}}$$

This assumes all voltages are defined with respect to the ground and if a single-phase device is the per-phase model of a three-phase Y configured device, then its neutral is grounded directly. Then (4.10a) becomes

$$\frac{E_j - V_j}{z_j^v} + \left(J_j - y_j^c V_j\right) - y_j^a V_j = \left(\sum_{k:j\sim k} y_{jk}^s + y_{jj}^m\right) V_j - \sum_{k:j\sim k} y_{jk}^s V_k, \qquad j \in \overline{N}$$

□

In the rest of this subsection we collect some analytical properties of the admittance matrix Y, particularly on their invertibility. Invertibility is of interests because given $I \in \mathbb{C}^{N+1}$ we may be interested in inverting Y to obtain $V \in \mathbb{C}^{N+1}$ from $I = YV$ as discussed in Chapter 4.2.5. The inverse $Z_{\text{bus}} := Y^{-1}$ is called a *bus impedance matrix* or an *impedance matrix* and is useful for fault analysis (which we will not cover in this book). The admittance matrix Y can be constructed easily by inspection of a network graph or its one-line diagram as specified by (4.10c). It inherits the sparsity structure of the network graph. The impedance matrix Z on the other hand cannot be easily inferred from the one-line diagram and is usually dense even for a sparse network. LU decomposition can be used for both computing Z and solving V from $I = YV$ (see Chapter 4.2.5).

We first consider the case where the shunt admittances of lines are negligible, i.e., $y_{jj}^m = 0$ for all $j \in \overline{N}$, so that all row sums of Y are zero. In this case Y is not invertible and we present its pseudo-inverse. We then discuss sufficient conditions under which Y with nonzero shunt admittances is invertible. We often assume C4.1 holds in this section and will explicitly state it where it is needed.

4.2.3.2 Pseudo-inverse and Takagi decomposition

Suppose $y_{jj}^m = 0$ for all $j \in \overline{N}$ so that Y has zero row (and hence column) sums.[2] Then Y is not invertible. Its pseudo-inverse always exists and can be obtained through singular value decomposition (see Chapter A.6 for singular value decomposition and Chapter A.7 for pseudo-inverse). Let $\bar{Y}$ denote the componentwise complex conjugate of Y, i.e., $[\bar{Y}]_{jk} = \left(Y_{jk}\right)^{\mathsf{H}}$. Then $Y = Y^{\mathsf{T}} = \left(\bar{Y}\right)^{\mathsf{H}}$. Let the singular value decomposition of Y be

$$Y = U\Sigma W^{\mathsf{H}}$$

where $\Sigma := \text{diag}\,(\sigma_0, \cdots, \sigma_N)$ is a $(N+1)\times(N+1)$ real nonnegative diagonal matrix whose diagonal entries $\sigma_j \geq 0$, called the singular values of Y, are the nonnegative square roots of the eigenvalues of $Y\bar{Y}$, and $U, W \in \mathbb{C}^{(N+1)\times(N+1)}$ are unitary matrices (see discussion after Theorem A.11 in Chapter A.6 for their derivation). The pseudo-inverse of Y is then

$$Y^{\dagger} := W\Sigma^{\dagger}U^{\mathsf{H}}$$

[2] If Y were real symmetric with zero row sums, then its rank is N and its null space is $\text{span}(\mathbf{1})$ when the network is connected. This property may not hold when Y is complex symmetric; see Exercise 4.2 for a sufficient condition for this property.

where $\Sigma^\dagger$ is the real nonnegative diagonal matrix obtained from Σ by replacing the nonzero singular values σ_j by $1/\sigma_j$.

If null(Y) =span$(\mathbf{1})$ then, for each current vector I with $\mathbf{1}^\mathsf{T} I = 0$, there is a subspace of solutions to $I = YV$ given by

$$V = Y^\dagger I \; + \; \gamma \mathbf{1}, \qquad \gamma \in \mathbb{C}$$

parametrized by γ. Hence V is unique up to an arbitrary reference voltage. For example the solution $V = Y^\dagger I$ corresponds to a solution with $\gamma = 0$. Alternatively γ can be chosen so that $V_0 = 1\angle 0°$ at bus 0. If null$(Y) \supset$span(1) then I needs to be orthogonal to all vectors in null(Y) for $I = YV$ to have a solution for V.

Under assumption C4.1, Y is symmetric. Since it is generally not Hermitian, it may not be *unitarily* diagonalizable. A matrix is unitarily diagonalizable if and only if it is normal (Theorem A.13 in Appendix A.6). Y may or may not be normal. See Exercise 4.3 for sufficient conditions under which Y is normal and hence unitarily diagonalizable. Even when Y is not normal, it can still be diagonalized but the unitary matrix U may consist of neither the singular vectors nor the eigenvectors of Y, according to Theorem A.17 in Appendix A.6.

Theorem 4.1 (Takagi decomposition of Y)**.** Suppose $y^m_{jj} = 0$ for all $j \in \overline{N}$ and condition C4.1 holds. There exists a unitary matrix $U \in \mathbb{C}^{(N+1)\times(N+1)}$ and a real nonnegative diagonal matrix $\Sigma :=$diag$(\sigma_1, \ldots, \sigma_{N+1})$ such that $Y = U\Sigma U^\mathsf{T}$ where the diagonal entries $\sigma_j \geq 0$ of Σ are the singular values of Y. □

Since $U^\mathsf{T} \neq U^\mathsf{H}$ in general, the Takagi decomposition is generally different from the singular decomposition of Y and therefore $Y^\dagger$ is generally not equal to $U\Sigma^\dagger U^\mathsf{T}$.

4.2.3.3 Inverse of Y

In this subsection we derive the inverse of Y, assuming it is invertible, in terms of its real and imaginary parts when either is invertible. Using the result in this subsection we will study conditions under which Y is invertible in Chapter 4.2.3.4.

Let $Y =: G + \mathbf{i}B$ with $G, B \in \mathbb{R}^{(N+1)\times(N+1)}$. Let $Z := R + \mathbf{i}X$ with $R, X \in \mathbb{R}^{(N+1)\times(N+1)}$. By definition Y^{-1} exists and is equal to Z if and only if there exist unique (R, X) such that $ZY = YZ = I$, the identity matrix. Consider

$$YZ \; = \; (G+\mathbf{i}B)(R+\mathbf{i}X) \; = \; (GR - BX) + \mathbf{i}(BR + GX) \; = \; I$$

or

$$\underbrace{\begin{bmatrix} G & -B \\ B & G \end{bmatrix}}_{M} \begin{bmatrix} R \\ X \end{bmatrix} \; = \; \begin{bmatrix} I \\ 0 \end{bmatrix} \tag{4.13a}$$

Therefore Y^{-1} exists if and only if the matrix $M := \begin{bmatrix} G & -B \\ B & G \end{bmatrix}$ is nonsingular. Suppose G is nonsingular. According to Theorem A.4 in Appendix A.3.1, M is nonsingular if and only if the Schur complement $M/G := G + BG^{-1}B$ of G is nonsingular (given that G is nonsingular). Moreover the inverse of M is

$$M^{-1} = \begin{bmatrix} (M/G)^{-1} & (M/G)^{-1}BG^{-1} \\ -G^{-1}B(M/G)^{-1} & G^{-1} - G^{-1}B(M/G)^{-1}BG^{-1} \end{bmatrix}$$

Hence if both G and M/G are nonsingular, then Y is nonsingular and, from (4.13a), its inverse $Z := R + \mathbf{i}X$ is given by

$$\begin{bmatrix} R \\ X \end{bmatrix} = \begin{bmatrix} (M/G)^{-1} \\ -G^{-1}B(M/G)^{-1} \end{bmatrix} = \begin{bmatrix} (G + BG^{-1}B)^{-1} \\ -G^{-1}B(G + BG^{-1}B)^{-1} \end{bmatrix} \tag{4.13b}$$

Suppose B is nonsingular. Then (4.13a) can be written equivalently as

$$\underbrace{\begin{bmatrix} B & G \\ G & -B \end{bmatrix}}_{M'} \begin{bmatrix} R \\ X \end{bmatrix} = \begin{bmatrix} 0 \\ I \end{bmatrix} \tag{4.14a}$$

Applying again Theorem A.4 in Appendix A.3.1, M' is nonsingular if and only if the Schur complement $M'/B := -B - GB^{-1}G$ of B is nonsingular (given that B is nonsingular). Moreover the inverse of M' is

$$M'^{-1} = \begin{bmatrix} B^{-1} + B^{-1}G(M'/B)^{-1}GB^{-1} & -B^{-1}G(M'/B)^{-1} \\ -(M'/B)^{-1}GB^{-1} & (M'/B)^{-1} \end{bmatrix}$$

Hence if both B and M'/B are nonsingular, then Y is nonsingular and, from (4.14a), its inverse $Z := R + \mathbf{i}X$ is given by

$$\begin{bmatrix} R \\ X \end{bmatrix} = \begin{bmatrix} -B^{-1}G(M'/B)^{-1} \\ (M'/B)^{-1} \end{bmatrix} = \begin{bmatrix} B^{-1}G(B + GB^{-1}G)^{-1} \\ -(B + GB^{-1}G)^{-1} \end{bmatrix} \tag{4.14b}$$

To recap, Y is invertible when both G and M/G are invertible or when both B and M'/B are invertible. When neither G nor B is invertible, $Y = G + \mathbf{i}B$ may still be invertible though its inverse $Z := R + \mathbf{i}X$ is not given by (4.13b) or (4.14b) (Exercise 4.4).

4.2.3.4 Properties of Y

We now use (4.13)(4.14) to study the invertibility of Y. Nonzero shunt admittances do not guarantee the invertibility of Y. A strictly diagonally dominant matrix is invertible (Theorem A.8 in Appendix A.3). Shunt admittances however does not guarantee strict diagonal dominance, i.e., $|Y_{ii}| > \sum_{j:j\neq i} |Y_{ij}|$ may not hold for some i. This can be the case for a transmission line since the susceptances of line charging admittances and those of series admittances are typically of different signs. Strict diagonal dominance

is however only sufficient for invertibility and a network of transmission lines typically has an invertible Y (see Remark 4.3). We now discuss two sufficient conditions for Y to be invertible.

The first sufficient condition builds on (4.13) and (4.14). It ensures both G and M/G are nonsingular, or both B and M'/B are nonsingular. Recall that a real matrix A is positive definite, denoted $A \succ 0$, if A is symmetric and $v^{\mathsf{T}}Av > 0$ for all real vectors v (See Remark A.1 in Appendix A.5). A positive definite matrix is nonsingular since all its eigenvalues are strictly positive. A real matrix A is negative definite, denoted $A \prec 0$, if $-A \succ 0$.

Theorem 4.2. Consider a complex symmetric matrix $Y = G + \mathbf{i}B$ (i.e., Y satisfies condition C4.1).

1 If $\mathrm{Re}(Y) \succ 0$ then Y^{-1} exists, is symmetric, and $\mathrm{Re}(Y^{-1}) \succ 0$.
2 If $\mathrm{Im}(Y) \prec 0$ then Y^{-1} exists, is symmetric, and $\mathrm{Im}(Y^{-1}) \succ 0$.

Proof For part 1, suppose $\mathrm{Re}(Y) = G \succ 0$. The Schur complement M/G of G is, from (4.13a), $M/G := G + BG^{-1}B$. Since $B = B^{\mathsf{T}}$ and G, G^{-1} are positive definite, $M/G := G + BG^{-1}B \succ 0$. Therefore both G and M/G are nonsingular, and hence Y is nonsingular according to Theorem A.4 in Appendix A.3.1. It also implies that $\mathrm{Re}(Y^{-1}) \succ 0$ since, from (4.13b), $\mathrm{Re}(Y^{-1}) = (M/G)^{-1}$ which is positive definite since M/G is.

Finally if $Z := Y^{-1}$ then Z is the unique matrix such that $YZ = ZY = I$ where I is the identity matrix. Then

$$Z^{\mathsf{T}}Y^{\mathsf{T}} = Y^{\mathsf{T}}Z^{\mathsf{T}} = Z^{\mathsf{T}}Y = YZ^{\mathsf{T}} = I$$

Hence $Z^{\mathsf{T}} = Y^{-1}$. Since inverse is unique, $Z^{\mathsf{T}} = Z$, i.e., Y^{-1} is (complex) symmetric.

Part 2 follows the same argument and is left as Exercise 4.5. (Also see Exercise 4.6 for an alternative proof of the nonsingularity of Y.) □

Remark 4.2 (Generalization)**.** Theorem 4.2 holds with small modifications as long as either $\mathrm{Re}(Y)$ or $\mathrm{Im}(Y)$ is not indefinite. Specifically if Y is complex symmetric then

1 Y^{-1} exists and is symmetric if (a) $\mathrm{Re}(Y) \succ 0$; or (b) $\mathrm{Re}(Y) \prec 0$; or (c) $\mathrm{Im}(Y) \succ 0$; or (d) $\mathrm{Im}(Y) \prec 0$.
2 (a) If $\mathrm{Re}(Y) \succ 0$ then $\mathrm{Re}(Y^{-1}) \succ 0$; and (b) if $\mathrm{Re}(Y) \prec 0$ then $\mathrm{Re}(Y^{-1}) \prec 0$.
3 (a) If $\mathrm{Im}(Y) \succ 0$ then $\mathrm{Im}(Y^{-1}) \prec 0$; and (b) if $\mathrm{Im}(Y) \prec 0$ then $\mathrm{Im}(Y^{-1}) \succ 0$.

□

The second set of sufficient conditions for the invertibility of Y is in terms of the series admittances y^s_{jk} and shunt admittances y^m_{jk}. These conditions ensure either

$\mathrm{Re}(Y)$ or $\mathrm{Im}(Y)$ is either positive or negative definite, and hence Y is nonsingular by Theorem 4.2 and Remark 4.2.

Let $Y = G + \mathbf{i}B$, i.e., for all $(j,k) \in E$,

$$y^s_{jk} =: g^s_{jk} + \mathbf{i}b^s_{jk}, \qquad y^m_{jk} =: g^m_{jk} + \mathbf{i}b^m_{jk}, \qquad y^m_{kj} =: g^m_{kj} + \mathbf{i}b^m_{kj}$$

Recall $y^m_{jj} := \sum_{k:j\sim k} y^m_{jk}$ and let $g^m_{jj} := \sum_{k:j\sim k} g^m_{jk}$, $b^m_{jj} := \sum_{k:j\sim k} b^m_{jk}$. Previous discussion implies that, for Y to be invertible, it is necessary to have at least one nonzero shunt element. Additional conditions on $\left(g^s_{jk}, g^m_{jk}, g^m_{kj}\right)$ are needed to guarantee invertibility, as follows.

C4.2: For all lines $(j,k) \in E$, $g^s_{jk}, g^m_{jk}, g^m_{kj}$ are nonnegative.
C4.3a: For all buses $j \in \overline{N}$, $g^m_{jj} := \sum_{k:k\sim j} g^m_{jk} \neq 0$, i.e., for all j, there exists a line $(j,k) \in E$ such that $g^m_{jk} \neq 0$.
C4.3b: For all lines $(j,k) \in E$, $g^s_{jk} \neq 0$. Furthermore there exists a line $(j',k') \in E$ such that $g^m_{j'k'} \neq 0$.

Condition C4.2 can be replaced by: for all lines $(j,k) \in E$, all nonzero $g^s_{jk}, g^m_{jk}, g^m_{kj}$ have the same sign, and the invertibility conditions below will still hold with obvious modifications. Indeed if $g^s_{jk}, g^m_{jk}, g^m_{kj}$ are all nonpositive then the proof below shows that $\mathrm{Re}(Y) \prec 0$ (see Remark 4.2).

Theorem 4.3. Suppose the network is connected and the admittance matrix Y satisfies condition C4.1. If C4.2 and one of C4.3a and C4.3b hold, then

1 $\mathrm{Re}(Y) \succ 0$.
2 Y^{-1} exists, is symmetric, and $\mathrm{Re}(Y^{-1}) \succ 0$.

Proof Recall that $\mathrm{Re}(Y) =: G \in \mathbb{R}^{(N+1)\times(N+1)}$ is given by $G_{jk} = -g^s_{jk}$ if $j \sim k$, $\sum_{i:j\sim i}(g^s_{ji} + g^m_{ji})$ if $j = k$, and 0 otherwise. Hence for any nonzero vector $\rho \in \mathbb{R}^{N+1}$ we have

$$\begin{aligned}
\rho^{\mathsf{T}} G \rho &= \sum_j \sum_k \rho_j \rho_k G_{jk} = \sum_j \left(\sum_{k:j\sim k} -\rho_j \rho_k g^s_{jk} + \rho_j^2 \sum_{i:j\sim i} (g^s_{ji} + g^m_{ji}) \right) \\
&= \sum_{(j,k)\in E} \left(\rho_j^2 - 2\rho_j\rho_k + \rho_k^2\right) g^s_{jk} + \sum_{j\in\overline{N}} \rho_j^2 g^m_{jj} \\
&= \sum_{(j,k)\in E} (\rho_j - \rho_k)^2 g^s_{jk} + \sum_{j\in\overline{N}} \rho_j^2 g^m_{jj}
\end{aligned}$$

Every summand is nonnegative by C4.2. Moreover if C4.3a holds then the second summation is strictly positive since $\rho \neq 0$. If C4.3b holds then for the first summation to be zero, $\rho_j = \rho_k$. Since the network is connected this implies $\rho_j = \rho_1$ for all j. Then the second summation becomes $\sum_j \rho_j^2 g^m_{jj} \geq \rho_1^2 g^m_{j'k'} > 0$ since $\rho \neq 0$. Therefore $\mathrm{Re}(Y) = G \succ 0$. Theorem 4.2 then completes the proof. □

Instead of $\left(g^s_{jk}, g^m_{jk}, g^m_{kj}\right)$ conditions on $\left(b^s_{jk}, b^m_{jk}, b^m_{kj}\right)$ can also ensure the invertibility of Y.

C4.4: For all lines $(j,k) \in E$, $b^s_{jk}, b^m_{jk}, b^m_{kj}$ are nonpositive.
C4.5a: For all buses $j \in \overline{N}$, $b^m_{jj} := \sum_{k:k\sim j} b^m_{jk} \neq 0$, i.e., for all j, there exists a line $(j,k) \in E$ such that $b^m_{jk} \neq 0$.
C4.5b: For all lines $(j,k) \in E$, $b^s_{jk} \neq 0$. Furthermore there exists a line $(j',k') \in E$ such that $b^m_{j'k'} \neq 0$.

As before C4.2 can be replaced by: for all lines $(j,k) \in E$, all nonzero $b^s_{jk}, b^m_{jk}, b^m_{kj}$ have the same sign, and the invertibility conditions below will still hold with obvious modifications.

Theorem 4.4. Suppose the network is connected and the admittance matrix Y satisfies condition C4.1. If C4.4 and one of C4.5a and C4.5b hold, then

1 $\mathrm{Im}(Y) \prec 0$.
2 Y^{-1} exists, is symmetric, and $\mathrm{Im}(Y^{-1}) \succ 0$.

Proof The proof is similar to that for Theorem 4.3. For $\mathrm{Im}(Y) =: B$, for any nonzero real vector ρ, the same calculation yields

$$\rho^{\mathsf{T}} B \rho \;=\; \sum_{(j,k)\in E} \left(\rho_j - \rho_k\right)^2 b^s_{jk} \;+\; \sum_{j\in\overline{N}} \rho_j^2 b^m_{jj}$$

Every summand is nonpositive by C4.4. Moreover if C4.5a holds then the second summation is strictly negative since $\rho \neq 0$. If C4.5b holds then for the first summation to be zero, $\rho_j = \rho_1$ for all j since the network is connected. Then the second summation becomes $\sum_j \rho_j^2 b^m_{jj} \leq \rho_1^2 b^m_{j'k'} < 0$ since $\rho \neq 0$. Therefore $\mathrm{Im}(Y) = B \prec 0$. Theorem 4.2 then completes the proof. □

Remark 4.3 (Transmission line)**.** A transmission line (j,k) typically has nonnegative series conductance $g^s_{jk} \geq 0$ and negative series susceptance $b^s_{jk} < 0$ (inductive line). Its shunt conductances $g^m_{jk} \geq 0$ are usually nonnegative, but shunt susceptances $b^m_{jk} \geq 0$ are usually nonnegative (capacitive).

1 Hence the conditions in Theorem 4.3 are usually satisfied for transmission lines (but not for transformers; see Example 4.4).
2 Since $b^s_{jk} < 0$ but $b^m_{jk} \geq 0$ for a typical transmission line, condition C4.4 in Theorem 4.4 is usually not satisfied.

□

Remark 4.4 (Distribution feeder test systems)**.** 1 The validity of $\mathrm{Re}(Y^{-1}) \succ 0$ has been checked on a set of test distribution feeders in [11, Section VI] ....

The conditions in Theorems 4.3 and 4.4 are sufficient but not necessary. The next example shows that, even though Condition C4.2 in Theorem 4.3 is usually not satisfied for a transformer, the admittance matrix may nonetheless be nonsingular.

Example 4.4 (Sufficiency only)**.** Consider Example 4.1. An alternative solution approach is to introduce an internal node 3 on the primary side of the ideal transformer, not the secondary side as in Example 4.1. [3] Then the parameters of lines $(1,3)$ and $(2,3)$ are

$$\begin{aligned}(y^s_{13}, y^m_{13}, y^m_{31}) &:= (\tilde{y}^s, 0, \tilde{y}^m)\\ (y^s_{23}, y^m_{23}, y^m_{32}) &:= (ny, (1-n)y, n(n-1)y)\end{aligned}$$

where n is the voltage gain of the transformer, y is the series admittance of the line and $(\tilde{y}^s, \tilde{y}^m)$ are the series and shunt admittances of the transformer. The admittance matrix is therefore

$$Y = \begin{bmatrix} \tilde{y}^s & 0 & -\tilde{y}^s \\ 0 & y & -ny \\ -\tilde{y}^s & -ny & \tilde{y}^s + \tilde{y}^m + n^2 y \end{bmatrix}$$

Let the admittances be of the form:

$$y =: g^s + \mathbf{i}b^s, \qquad \tilde{y}^s =: \tilde{g}^s + \mathbf{i}\tilde{b}^s, \qquad \tilde{y}^m =: \mathbf{i}\tilde{b}^m$$

and suppose $g^s, \tilde{g}^s > 0$, $b^s, \tilde{b}^s \le 0$, and $\tilde{b}^m \ge 0$. We now show that the admittance matrix Y does not satisfy condition C4.2 in Theorem 4.3, but Y is invertible if and only if $\tilde{b}^m > 0$.

We have

$$\begin{aligned} y^s_{13} = y^s_{31} &= \tilde{g}^s + \mathbf{i}\tilde{b}^s, \qquad y^s_{23} = y^s_{32} = ng^s + \mathbf{i}nb^s \\ y^m_{11} = 0, \qquad y^m_{22} &= (1-n)g^s + \mathbf{i}(1-n)b^s, \qquad y^m_{33} = n(n-1)g^s + \mathbf{i}\left(n(n-1)b^s + \tilde{b}^m\right) \end{aligned}$$

Hence condition C4.1 is satisfied but C4.2 is not since $g^m_{23} := (1-n)g^s$ and $g^m_{32} := n(n-1)g^s$ have opposite signs unless $n = 1$. For any complex symmetric matrix $\hat{Y}$ with line parameters $(\hat{y}^s_{jk}, \hat{y}^m_{jk}, \hat{y}^m_{kj})$, for any nonzero vector α^{H}, one can show (Exercise 4.9)

$$\alpha^{\mathsf{H}}\hat{Y}\alpha = \left(\sum_{(j,k)\in E} \hat{g}^s_{jk}\left|\alpha_j - \alpha_k\right|^2 + \sum_{j\in\overline{N}} \hat{g}^m_{jj}\left|\alpha_j\right|^2\right) + \mathbf{i}\left(\sum_{(j,k)\in E} \hat{b}^s_{jk}\left|\alpha_j - \alpha_k\right|^2 + \sum_{j\in\overline{N}} \hat{b}^m_{jj}\left|\alpha_j\right|^2\right)$$

Hence

$$\begin{aligned}\mathrm{Re}\left(\alpha^{\mathsf{H}}Y\alpha\right) &= \left(\tilde{g}^s|\alpha_1 - \alpha_3|^2 + ng^s|\alpha_2 - \alpha_3|^2\right) + \left((1-n)g^s|\alpha_2|^2 + n(n-1)g^s|\alpha_3|^2\right) \\ &= \tilde{g}^s|\alpha_1 - \alpha_3|^2 + g^s|\alpha_2 - n\alpha_3|^2\end{aligned}$$

Therefore

$$\mathrm{Re}\left(\alpha^{\mathsf{H}}Y\alpha\right) = 0 \text{ if and only if } \alpha_1 = \alpha_3 = \frac{\alpha_2}{n} \tag{4.15}$$

[3]

On the other hand

$$\begin{aligned}\mathrm{Im}\left(\alpha^{\mathsf{H}} Y \alpha\right) &= \left(\tilde{b}^s|\alpha_1-\alpha_3|^2+nb^s|\alpha_2-\alpha_3|^2\right) \;+\; \left((1-n)b^s|\alpha_2|^2+\left(n(n-1)b^s+\tilde{b}^m\right)|\alpha_3|^2\right)\\ &= b^l|\alpha_1-\alpha_3|^2 \;+\; b^s|\alpha_2-n\alpha_3|^2 \;+\; \tilde{b}^m|\alpha_3|^2\end{aligned}$$

In light of (4.15), if $\tilde{b}^m > 0$ then $\alpha^{\mathsf{H}} Y \alpha = 0$ if and only if $\alpha_1 = \alpha_2 = \alpha_3 = 0$. Hence if $\tilde{b}^m > 0$ then Y is invertible.

Conversely if $\tilde{b}^m = 0$ then there exists nonzero $\alpha \in \mathbb{C}^3$ with $\alpha^{\mathsf{H}} Y \alpha = 0$. Exercise 4.8 says that, since Y is complex symmetric (but not Hermitian), this does not necessarily imply $Y\alpha = 0$ and hence may not imply that Y is singular. Using the admittance matrix given above, however, it can be verified that, when $y^m = \mathbf{i}\tilde{b}^m = 0$, $\alpha := \begin{bmatrix}1 & n & 1\end{bmatrix}^{\mathsf{T}}$ is indeed an eigenvector of Y corresponding to zero eigenvalue. Hence Y is singular if the (only) shunt element $\tilde{b}^m$ in the model is zero, even when y^m_{22} and y^m_{33}, which originate from the effect of an ideal transformer, are nonzero. □

4.2.4 Kron reduction Y/Y_{22} and its properties

In many applications we are interested in the relation between the current injections and voltages at only a subset $N_{\mathrm{red}} \subset \overline{N}$ of the buses. For example we are interested in the external behavior of a system defined by the relationship between currents and voltages of the end devices. In this subsection we define Kron reduction that describes the relation between the nodal voltages and current injections at buses in N_{red} and study its properties.

4.2.4.1 Kron reduction Y/Y_{22}

Denote the number of buses in N_{red} also by N_{red}. Without loss of generality we can partition the buses such that $I_1 \in \mathbb{C}^{N_{\mathrm{red}}}$ denotes the first N_{red} current injections and I_2 the remaining $N+1-N_{\mathrm{red}}$ current injections. Similarly partition the voltages into (V_1, V_2) with $V_1 \in \mathbb{C}^{N_{\mathrm{red}}}$, $V_2 \in \mathbb{C}^{N+1-N_{\mathrm{red}}}$. Partition the admittance matrix Y so that

$$\begin{bmatrix}I_1\\ I_2\end{bmatrix} = \underbrace{\begin{bmatrix}Y_{11} & Y_{12}\\ Y_{21} & Y_{22}\end{bmatrix}}_{Y}\begin{bmatrix}V_1\\ V_2\end{bmatrix}$$

If Y_{22} is invertible then we can eliminate V_2 by substituting $V_2 = -Y_{22}^{-1}Y_{21}V_1 + Y_{22}^{-1}I_2$ to obtain

$$\left(Y_{11} - Y_{12}Y_{22}^{-1}Y_{21}\right)V_1 = I_1 \;-\; Y_{12}Y_{22}^{-1}I_2 \tag{4.16}$$

The $N_{\mathrm{red}} \times N_{\mathrm{red}}$ matrix $Y/Y_{22} := Y_{11} - Y_{12}Y_{22}^{-1}Y_{21}$ is the Schur complement of Y_{22} of matrix Y (see Appendix A.3 for its properties). It can be interpreted as the admittance matrix of the reduced network consisting only of buses in N_{red} and describes the effective connectivity and line admittances of the reduced network. The quantity $I_1 -$

$Y_{12}Y_{22}^{-1}I_2$ describes the effective current injections at these buses. This is called a *Kron reduction* of network G. If Y is complex symmetric, its Kron reduced admittance matrix Y/Y_{22} is also complex symmetric and hence satisfies Assumption C4.1 (Exercise 4.10). Two buses j and k are adjacent in the Kron-reduced network, i.e., $[Y/Y_{22}]_{jk} \neq 0$, if and only if j and k are adjacent in the original graph (i.e., $Y_{jk} \neq 0$) or if there is a path in the original graph that connects j and k.

Example 4.5 (Kron reduction)**.** Consider the network shown in Figure 4.10(a). Under

(a) Original network (b) Kron reduced network

Figure 4.10 Kron reduction: $N_{\text{red}} := \{1,2,3\}$ with internal bus 4. While the original network is a tree, the Kron reduced network is fully connected.

condition C4.1 its admittance matrix Y is (0 and symmetric entries are omitted for simplicity)

$$Y := \begin{bmatrix} y^s_{14}+y^m_{11} & & & -y^s_{14} \\ & y^s_{24}+y^m_{22} & & -y^s_{24} \\ & & y^s_{34}+y^m_{33} & -y^s_{34} \\ & & & \sum_j y^s_{j4}+y^m_{44} \end{bmatrix}$$

with $Y_{22} := \sum_j y^s_{j4} + y^m_{44}$. The Schur complement Y/Y_{22} of Y_{22} is

$$\begin{aligned} & Y_{11} - Y_{12}Y_{22}^{-1}Y_{21} \\ &= \begin{bmatrix} y^s_{14}+y^m_{11} & & \\ & y^s_{24}+y^m_{22} & \\ & & y^s_{34}+y^m_{33} \end{bmatrix} - \frac{1}{Y_{22}} \begin{bmatrix} -y^s_{14} \\ -y^s_{24} \\ -y^s_{34} \end{bmatrix} \begin{bmatrix} -y^s_{14} & -y^s_{24} & -y^s_{34} \end{bmatrix} \\ &= \begin{bmatrix} \frac{y^s_{14}}{Y_{22}}\left(y^s_{24}+y^s_{34}\right)+\left(y^m_{11}+\gamma\, y^s_{14}\right) & \frac{-y^s_{14}y^s_{24}}{Y_{22}} & \frac{-y^s_{14}y^s_{34}}{Y_{22}} \\ \frac{y^s_{24}}{Y_{22}}\left(y^s_{14}+y^s_{34}\right)+\left(y^m_{22}+\gamma\, y^s_{24}\right) & \frac{-y^s_{24}y^s_{34}}{Y_{22}} & \\ \frac{y^s_{34}}{Y_{22}}\left(y^s_{14}+y^s_{24}\right)+\left(y^m_{33}+\gamma\, y^s_{34}\right) & & \end{bmatrix} \end{aligned}$$

where $\gamma := y^m_{44}/Y_{22} = y^m_{44}/\left(\sum_j y^s_{j4}+y^m_{44}\right)$. The Kron reduced network corresponding to Y/Y_{22} is fully connected as shown in Figure 4.10(b).

The effective current injections in the Kron reduced network are

$$\begin{bmatrix} I_1 \\ I_2 \\ I_3 \end{bmatrix} - Y_{12}Y_{22}^{-1}I_3 = \begin{bmatrix} I_1 \\ I_2 \\ I_3 \end{bmatrix} + \begin{bmatrix} y^s_{14} \\ y^s_{24} \\ y^s_{34} \end{bmatrix} \frac{I_3}{Y_{22}}$$

□

An admittance matrix Y has zero row, and hence column, sums if and only if all line charging admittances are zero, $y^m_{jk} = y^m_{kj} = 0$ for $(j,k) \in E$. In that case the Kron-reduced admittance matrix Y/Y_{22} also has zero, and hence column, sums (Exercise 4.10). The converse may not hold.

Given current injections $I = (I_1, I_2)$, we can obtain V_1 in terms of the Schur complement Y/Y_{22} and the effective current injections:

$$V_1 = \left(Y_{11} - Y_{12}Y_{22}^{-1}Y_{21}\right)^{-1} \left(I_1 \; - \; Y_{12}Y_{22}^{-1}I_2\right)$$

In many applications current injections $I_2 = 0$. For example the buses in $\overline{N} \setminus N_{\text{red}}$ represent internal buses without generators or loads (see Example 4.1). Then (4.16) reduces to:

$$I_1 = \underbrace{\left(Y_{11} - Y_{12}Y_{22}^{-1}Y_{21}\right)}_{Y/Y_{22}} V_1$$

and the reduced network is described by the Schur complement Y/Y_{22} that directly relates V_1 and I_1.

4.2.4.2 Properties of Y_{22}

We now study sufficient conditions for the existence of Kron reduction, i.e., of Y_{22}. The principal submatrix Y_{22} may not be strictly diagonal dominant nor invertible.[4] The situation is similar to the invertibility of Y and Theorems 4.3 and 4.4 to Y_{22} and their proofs extend directly to its submatrix Y_{22}.

Let $A \subsetneq \overline{N}$ denote the set of buses corresponding to Y_{22} and assume A is a strict subset of $\overline{N}$. For the rest of this subsection denote the (j,k) entry of a matrix M by $M[j,k]$, e.g., $Y[j,k], Y_{22}[j,k]$. Note that the indices j,k of Y_{22} take values in A, e.g., if Y_{22} corresponds to the last n buses, they run from $N-n+2,\ldots,N+1$, not $1,\ldots,n$. The argument is similar to that for the invertibility of Y. By definition Y_{22} is singular if and only if zero is an eigenvalue of Y_{22}. [5] If λ is an eigenvalue and $\alpha \in \mathbb{C}^n$ is a corresponding eigenvector then

$$\alpha^{\mathsf{H}} Y_{22} \alpha = \sum_{j \in A} \sum_{k \in A} Y[j,k] \alpha_j^{\mathsf{H}} \alpha_k \;\; = \;\; \lambda ||\alpha||_2^2 \tag{4.17}$$

where $||\cdot||_2$ denotes the Euclidean norm. Hence for Y_{22} to be invertible it is sufficient, but not necessary, that $\alpha^{\mathsf{H}} Y_{22} \alpha \neq 0$ for all nonzero vectors $\alpha \in \mathbb{C}^n$ (see Exercise 4.8).

[4] This is in contrast to the Laplacian matrix $Y = L$ in the DC power flow model for which a strict principal submatrix is always strictly diagonally dominant and hence invertible. See Chapter 4.6.2.

[5]

We have from (4.10c)

$$Y_{22}[j,j] = \sum_{k\notin A:(j,k)\in E} y^s_{jk} + \sum_{k\in A:(j,k)\in E} y^s_{jk} + y^m_{jj}, \qquad j \in A$$

Substituting this and $Y[j,k] = -y^s_{jk}$ for $j \sim k$ into (4.17) we have

$$\begin{aligned}
\alpha^{\mathsf{H}} Y_{22}\alpha &= \sum_{j\in A}\left(\left(\sum_{k\notin A:(j,k)\in E} y^s_{jk} + \sum_{k\in A:(j,k)\in E} y^s_{jk} + y^m_{jj}\right)|\alpha_j|^2 - \sum_{k\in A:(j,k)\in E} y^s_{jk}\,\alpha_j^{\mathsf{H}}\alpha_k\right) \\
&= \sum_{j,k\in A:(j,k)\in E}\left(y^s_{jk}|\alpha_j|^2 - y^s_{jk}\alpha_j^{\mathsf{H}}\alpha_k - y^s_{kj}\alpha_k^{\mathsf{H}}\alpha_j + y^s_{kj}|\alpha_k|^2\right) + \sum_{j\in A}\left(\sum_{k\notin A:(j,k)\in E} y^s_{jk} + y^m_{jj}\right) \\
&= \sum_{j,k\in A:(j,k)\in E} y^s_{jk}\left|\alpha_j - \alpha_k\right|^2 + \sum_{j\in A}\left(\sum_{k\notin A:(j,k)\in E} y^s_{jk} + y^m_{jj}\right)|\alpha_j|^2
\end{aligned}$$

where the third equality uses from $y^s_{jk} = y^s_{kj}$ when condition C4.1 holds. The first term sums over links in the subgraph induced by A. The second term sums over links between the subgraph induced by A and that by $\overline{N}\setminus A$. Recall $y^s_{jk} =: g^s_{jk} + \mathbf{i}b^s_{jk}$ and $y^m_{jj} =: g^m_{jj} + \mathbf{i}b^m_{jj}$. Then

$$\mathrm{Re}\left(\alpha^{\mathsf{H}} Y_{22}\alpha\right) = \sum_{j,k\in A:(j,k)\in E} g^s_{jk}\left|\alpha_j - \alpha_k\right|^2 + \sum_{j\in A}\left(\sum_{k\notin A:(j,k)\in E} g^s_{jk} + g^m_{jj}\right)|\alpha_j|^2 \tag{4.18a}$$

$$\mathrm{Im}\left(\alpha^{\mathsf{H}} Y_{22}\alpha\right) = \sum_{j,k\in A:(j,k)\in E} b^s_{jk}\left|\alpha_j - \alpha_k\right|^2 + \sum_{j\in A}\left(\sum_{k\notin A:(j,k)\in E} b^s_{jk} + b^m_{jj}\right)|\alpha_j|^2 \tag{4.18b}$$

The subgraph corresponding to Y_{22} may consist of multiple connected components $C_i \subseteq A$. Each connected component C_i is a disjoint set of buses that are connected to each other and to no buses outside C_i such that $\cup_i C_i = A$. Let

$$G_j := \sum_{k\notin A:(j,k)\in E} g^s_{jk} + g^m_{jj}, \qquad B_j := \sum_{k\notin A:(j,k)\in E} b^s_{jk} + b^m_{jj}, \qquad j \in A \tag{4.19a}$$

Then we can rewrite (4.18) in terms of the connected components C_i and G_j, B_j:

$$\mathrm{Re}\left(\alpha^{\mathsf{H}} Y_{22}\alpha\right) = \sum_i\left(\sum_{j,k\in C_i:(j,k)\in E} g^s_{jk}\left|\alpha_j - \alpha_k\right|^2 + \sum_{j\in C_i} G_j|\alpha_j|^2\right) \tag{4.19b}$$

$$\mathrm{Im}\left(\alpha^{\mathsf{H}} Y_{22}\alpha\right) = \sum_i\left(\sum_{j,k\in C_i:(j,k)\in E} b^s_{jk}\left|\alpha_j - \alpha_k\right|^2 + \sum_{j\in C_i} B_j|\alpha_j|^2\right) \tag{4.19c}$$

These expressions are similar to $\rho^{\mathsf{T}} G\rho$ and $\rho^{\mathsf{T}} B\rho$ in the proofs of Theorems 4.3 and 4.4 respectively. Hence Theorems 4.3 and 4.4 extend directly to Y_{22} as stated in the next two results.

Consider the following conditions on the conductances g^s_{jk} and G_j:

C4.6: For all lines $(j,k) \in E$, $g^s_{jk} \geq 0$ and for all buses $j \in \overline{N}$, $G_j \geq 0$.
C4.7a: For all buses $j \in \overline{N}$, $G_j \neq 0$,
C4.7b: For all lines $(j,k) \in E$, $g^s_{jk} \neq 0$. Furthermore on each connected component C_i there exists a bus $j_i \in C_i$ such that $G_{j_i} \neq 0$.

Conditions C4.6 can be changed to g^s_{jk}, G_j having the same sign.

Theorem 4.5. Suppose the admittance matrix Y satisfies condition C4.1. If C4.6 and one of C4.7a and C4.7b hold, then the strict principal submatrix Y_{22} satisfies

1 $\mathrm{Re}(Y_{22}) \succ 0$.
2 Y_{22}^{-1} exists, is symmetric, and $\mathrm{Re}\big(Y_{22}^{-1}\big) \succ 0$.

Proof The proof is similar to that for Theorem 4.3. Condition C4.6 implies that every summand in (4.19b) is nonnegative. Moreover if C4.7a holds then the second summation is strictly positive if $\alpha \neq 0$. If C4.7b holds then for the first summation to be zero, $\alpha_j = \alpha_k$ for all j,k in each connected component C_i. Then the second summation becomes, on each C_i, $\sum_{j \in C_i} G_j |\alpha_j|^2 \geq G_{j_i} |\alpha_{j_i}|^2 > 0$ unless $\alpha_j = \alpha_{j_i} = 0$ for all $j \in C_i$. Therefore $\mathrm{Re}\big(\alpha^{\mathsf{H}} Y_{22} \alpha\big) > 0$ if $\rho \neq 0$, i.e., $\mathrm{Re}(Y_{22}) \succ 0$. Since Y_{22} is symmetric Theorem 4.2 then completes the proof. □

Consider the following conditions on the susceptances b^s_{jk} and B_j:

C4.8: $b^s_{jk} \leq 0$ for all lines $(j,k) \in E$ and $B_j \leq 0$ for all buses $j \in \overline{N}$.
C4.9a: For all buses $j \in \overline{N}$, $B_j \neq 0$,
C4.9b: For all lines $(j,k) \in E$, $b^s_{jk} \neq 0$. Furthermore on each connected component C_i there exists a bus $j_i \in C_i$ such that $B_{j_i} \neq 0$.

Conditions C4.8 can be changed to b^s_{jk}, B_j having the same sign respectively.

Theorem 4.6. Suppose the admittance matrix Y satisfies condition C4.1. If C4.8 and one of C4.9a and C4.9b hold, then the strict principal submatrix Y_{22} satisfies

1 $\mathrm{Im}(Y_{22}) \prec 0$.
2 Y_{22}^{-1} exists, is symmetric, and $\mathrm{Im}\big(Y_{22}^{-1}\big) \succ 0$.

The invertibility conditions in Theorems 4.5 and 4.6 for the submatrix Y_{22} are less restrictive than those in Theorems 4.3 and 4.4 for Y, as we explain in Remark 4.5. Therefore if conditions of Theorem 4.3 or 4.4 are satisfied then Y^{-1}, Y_{22}^{-1} and Y/Y_{22} all exist.

Remark 4.5 (Transmission line)**.** As discussed in Remark 4.3, for a transmission line, we usually have $g^s_{jk} \geq 0$, $b^s_{jk} < 0$, $g^m_{jj} \geq 0$ and $b^m_{jj} \geq 0$.

1 If all lines (j,k) have strictly positive conductances, then conditions C4.6 and C4.7b are satisfied. This is the case even with zero shunt admittances $y^m_{jk} = y^m_{kj} = 0$ in which case Y has zero row sums and is singular.
2 For C4.8, even though b^s_{jk} and b^m_{jj} have opposite signs, the shunt susceptances b^m_{jk} are typically much smaller than the series susceptances b^s_{jk} such that usually B_j in (4.19a) has the same sign as b^s_{jk}. Hence both C4.8 and C4.9a are likely to be satisfied since b^s_{jk} are usually nonzero for transmission lines.

When shunt admittances $y^m_{jk} = y^m_{kj} = 0$.

When $y^m_{jk} = y^m_{kj} = 0$ for all lines $(j,k) \in E$ a symmetric admittance matrix Y has zero row and column sums and is hence singular. In this case G_j and B_j in (4.19a) becomes

$$G_j := \sum_{k \notin A:(j,k)\in E} g^s_{jk}, \qquad B_j := \sum_{k \notin A:(j,k)\in E} b^s_{jk}, \qquad j \in A$$

Hence Theorems 4.5 and 4.6 imply the following simple conditions for the invertibility of a strict principal submtirx Y_{22} of Y.

Corollary 4.7. Suppose the admittance matrix Y satisfies condition C4.1 and $y^m_{jk} = y^m_{kj} = 0$ for all lines $(j,k) \in E$. Consider the strict principal submatrix Y_{22}.

1 If $g^s_{jk} > 0$ for all lines $(j,k) \in E$ then Y_{22}^{-1} exists and is symmetric. Moreover both $\mathrm{Re}(Y_{22}) \succ 0$ and $\mathrm{Re}\big(Y_{22}^{-1}\big) \succ 0$.
2 If $b^s_{jk} < 0$ for all lines $(j,k) \in E$ then then Y_{22}^{-1} exists and is symmetric. Moreover $\mathrm{Im}(Y_{22}) \prec 0$ but $\mathrm{Im}\big(Y_{22}^{-1}\big) \succ 0$.

For a real symmetric Laplacian matrix L with zero row and column sums (which is the admittance matrix of the DC power flow model studied in Chapter 4.6), Theorem 4.13 shows that any strict principal submatrix L_{22} is nonsingular. See Remark 4.9 for connection of the invertibility conditions of Corollary 4.8 for complex symmetric matrices Y to that in Theorem 4.13 for real symmetric Laplacian matrix L.

When not all g^s_{jk} are strictly positive and not all b^s_{jk} are strictly negative, then neither $\mathrm{Re}(Y_{22}) \succ 0$ nor $\mathrm{Im}(Y_{22}) \prec 0$ may hold. It turns out however that $\mathrm{Re}(Y_{22}) - \mathrm{Im}(Y_{22}) \succ 0$ as long as $g^s_{jk} \geq 0$ and $b^s_{jk} \leq 0$ because they cannot be zero simultaneously, i.e., $z^s_{jk} \neq 0$ if $(j,k) \in E$. This implies the nonsingularity of Y_{22}, as the following result from [12] shows.

Theorem 4.8. Suppose the admittance matrix Y satisfies condition C4.1 and $y^m_{jk} = y^m_{kj} = 0$ for all lines $(j,k) \in E$. If $g^s_{jk} \geq 0$ and $b^s_{jk} \leq 0$ for all lines $(j,k) \in E$ then the strict principal submatrix Y_{22} satisfies

1 $\mathrm{Re}(Y_{22}) \succeq 0$, $\mathrm{Im}(Y_{22}) \preceq 0$, but $\mathrm{Re}(Y_{22}) - \mathrm{Im}(Y_{22}) \succ 0$.
2 Y_{22}^{-1} exists and is symmetric.

Proof Write $Y =: G + \mathbf{i}B$ and $Y_{22} =: G_{22} + \mathbf{i}B_{22}$. Denote the (j,k) element of a matrix M by $M[j,k]$, e.g., $Y[j,k]$, $G_{22}[j,k]$, etc. Since $y^m_{jk} = y^m_{kj} = 0$ for all lines $(j,k) \in E$ and hence Y has zero row (and column) sums, each row of G_{22} and B_{22} are diagonally dominant:

$$|G_{22}[j,j]| = \left| \sum_{k \notin A:(j,k)\in E} g^s_{jk} + \sum_{k\in A:(j,k)\in E} g^s_{jk} \right| \geq \sum_{k\in A:(j,k)\in E} g^s_{jk} = \sum_{k\in A:k\neq j} |G_{22}[j,k]|, \quad j \in$$

$$|B_{22}[j,j]| = \left| \sum_{k \notin A:(j,k)\in E} b^s_{jk} + \sum_{k\in A:(j,k)\in E} b^s_{jk} \right| \geq \sum_{k\in A:(j,k)\in E} -b^s_{jk} = \sum_{k\in A:k\neq j} |B_{22}[j,k]|, \quad j \in$$

Since G_{22} and B_{22} are real and symmetric their eigenvalues are all real. The Geršgorin disc theorem states that all eigenvalues of a real matrix $M \in \mathbb{R}^{n\times n}$ lie in the union of n discs

$$\cup_{i=1}^n \left\{ z \in \mathbb{C}^n \ : \ |z - M_{ii}| \leq \sum_{j:j\neq i} |M_{ij}| \right\}$$

Therefore all eigvenvalues of the G_{22} are nonnegative and those of B_{22} are nonpositive, i.e., $G_{22} \geq 0$ and $B_{22} \leq 0$, since G_{22} and B_{22} are real symmetric. This implies that $G_{22} - B_{22} \geq 0$.

We now show that, indeed, $G_{22} - B_{22} > 0$ because the network is connected and $A \subset \overline{N}$ is a strict subset. Since $G_{22} - B_{22}$ is real symmetric, consider, for any nonzero real vector ρ,

$$\begin{aligned}
\rho^{\mathsf{T}}(G_{22} - B_{22})\rho &= \sum_{j\in A}\sum_{k\in A} \rho_j (G_{22}[j,k] - B_{22}[j,k])\rho_k \\
&= \sum_{j\in A}\sum_{k\in A:(j,k)\in E} \rho_j(-g^s_{jk} + b^s_{jk})\rho_k + \sum_{j\in A} \rho_j^2 \left(\sum_{k\in A:(j,k)\in E} (g^s_{jk} - b^s_{jk}) + \sum_{k\notin A:(j,k)} \right. \\
&= \sum_{j,k\in A:(j,k)\in E} (\rho_j - \rho_k)^2 (g^s_{jk} - b^s_{jk}) + \sum_{j\in A} \rho_j^2 (G_j - B_j)
\end{aligned}$$

where the third equality uses $g^s_{jk} = g^s_{kj}$ and $b^s_{jk} = b^s_{kj}$ from C4.1. Here $G_j - B_j = \sum_{k\notin A:(j,k)\in E}(g^s_{jk} - b^s_{jk})$ for $j \in A$ and the summation is not vacuous because the network is connected and $A \subsetneq \overline{N}$. For every line $(j,k) \in E$, $y^s_{jk} \neq 0$ and hence $g^s_{jk} - b^s_{jk} > 0$ since $g^s_{jk} \geq$ and $b^s_{jk} \geq 0$. This implies $G_j - B_j > 0$ as well for all $j \in A$. Therefore for $\rho^{\mathsf{T}}(G_{22} - B_{22})\rho > 0$ for any real vector $\rho \neq 0$, i.e., $G_{22} - B_{22} > 0$.

Finally we use $G_{22} - B_{22} > 0$ to show that Y_{22} is nonsingular (it is clear that Y_{22}^{-1} is symmetric if it exists). If Y_{22} is singular then it has a nonzero eigenvector $\alpha = \rho + \mathbf{i}\epsilon$ corresponding to the zero eigenvalue and hence

$$0 = Y_{22}\alpha = (G_{22} + \mathbf{i}B_{22})(\rho + \mathbf{i}\epsilon) = (G_{22}\rho - B_{22}\epsilon) + \mathbf{i}(G_{22}\epsilon + B_{22}\rho)$$

Therefore

$$G_{22}\rho - B_{22}\epsilon = 0, \qquad B_{22}\rho + G_{22}\epsilon = 0$$

To solve for (ρ, ϵ), subtract the second equation from the first to get $(G_{22}-B_{22})\rho = (G_{22}+B_{22})\epsilon$. Since $G_{22}-B_{22} \succ 0$ we have $\rho = (G_{22}-B_{22})^{-1}(G_{22}+B_{22})\epsilon$. Substituting into the first equation we have

$$\begin{aligned} 0 &= \left(G_{22}(G_{22}-B_{22})^{-1}(G_{22}+B_{22}) - B_{22}\right)\epsilon \\ &= \left(G_{22}(G_{22}-B_{22})^{-1}G_{22} + G_{22}(G_{22}-B_{22})^{-1}B_{22} - B_{22}\right)\epsilon \end{aligned}$$

But $G_{22}(G_{22}-B_{22})^{-1}B_{22} - B_{22} = (G_{22} - (G_{22}-B_{22}))\,(G_{22}-B_{22})^{-1}B_{22} = B_{22}(G_{22}-B_{22})^{-1}B_{22}$ and hence

$$0 = \left(G_{22}(G_{22}-B_{22})^{-1}G_{22} + B_{22}(G_{22}-B_{22})^{-1}B_{22}\right)\epsilon$$

Multiplying on the left by ϵ^{T} we have

$$0 = \epsilon^{\mathsf{T}}\left(G_{22}(G_{22}-B_{22})^{-1}G_{22} + B_{22}(G_{22}-B_{22})^{-1}B_{22}\right)\epsilon$$

which implies $\epsilon = 0$ since $(G_{22}-B_{22})^{-1} \succ 0$. But then $\rho = (G_{22}-B_{22})^{-1}(G_{22}+B_{22})\epsilon = 0$ and therefore $\alpha = \rho + \mathbf{i}\epsilon = 0$, contradicting that the eigenvector α is nonzero. Hence Y_{22} is nonsingular. □

4.2.4.3 Properties of Y/Y_{22}

Theorem 4.2 extends directly to the Schur complement $Y/Y_{22} := Y_{11} - Y_{12}Y_{22}^{-1}Y_{12}^{\mathsf{T}}$.

Theorem 4.9. Consider a complex symmetric matrix $Y =: \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{12}^{\mathsf{T}} & Y_{22} \end{bmatrix}$ (i.e., Y satisfies condition C4.1). Suppose Y_{22} is nonsingular.

1. If $\mathrm{Re}(Y) \succ 0$, then $(Y/Y_{22})^{-1}$ exists and is symmetric. Moreover $\mathrm{Re}(Y/Y_{22}) \succ 0$ and $\mathrm{Re}\big((Y/Y_{22})^{-1}\big) \succ 0$.
2. If $\mathrm{Im}(Y) \prec 0$, then $(Y/Y_{22})^{-1}$ exists and is symmetric. Moreover $\mathrm{Im}(Y/Y_{22}) \prec 0$ but $\mathrm{Im}\big((Y/Y_{22})^{-1}\big) \succ 0$.

Proof Since Y is symmetric, Y_{22}^{-1} and Y/Y_{22} are symmetric as well (Exercise 4.10). From Theorem A.4 in Appendix A.3.1, Y is nonsingular if and only if Y/Y_{22} is nonsingular, given that Y_{22} is nonsingular. If $\mathrm{Re}(Y) \succ 0$ or $\mathrm{Im}(Y) \prec 0$, Theorem 4.2 implies that Y^{-1} exists and $\mathrm{Re}(Y^{-1}) \succ 0$ or $\mathrm{Im}(Y^{-1}) \prec 0$ respectively. Hence Y/Y_{22} is nonsingular if $\mathrm{Re}(Y) \succ 0$ or $\mathrm{Im}(Y) \prec 0$.

Write Y^{-1} in terms of the Schur complement Y/Y_{22} (from Theorem A.4):

$$Y^{-1} = \begin{bmatrix} (Y/Y_{22})^{-1} & -(Y/Y_{22})^{-1}Y_{12}Y_{22}^{-1} \\ -Y_{22}^{-1}Y_{12}^{\mathsf{T}}(Y/Y_{22})^{-1} & A \end{bmatrix}$$

where $A := Y_{22}^{-1} + Y_{22}^{-1}Y_{12}^{\mathsf{T}}(Y/Y_{22})^{-1}Y_{12}Y_{22}^{-1}$. If $\mathrm{Re}(Y) \succ 0$ then Theorem 4.2 implies that $\mathrm{Re}(Y^{-1}) \succ 0$. Hence all the principal submatrices of $\mathrm{Re}(Y^{-1})$ are (symmetric and)

positive definite. In particular $\mathrm{Re}\big((Y/Y_{22})^{-1}\big) \succ 0$. But $(Y/Y_{22})^{-1}$ is symmetric and therefore Theorem 4.2 implies that $\mathrm{Re}(Y/Y_{22}) \succ 0$.

If on the other hand $\mathrm{Im}(Y) \prec 0$, then Theorem 4.2 implies that $\mathrm{Im}(Y^{-1}) \succ 0$. Hence its principal submatrix $\mathrm{Im}\big((Y/Y_{22})^{-1}\big) \succ 0$. But $(Y/Y_{22})^{-1}$ is symmetric and therefore Remark 4.2 implies that $\mathrm{Im}(Y/Y_{22}) \prec 0$. □

4.2.5 Solving $I = YV$

Suppose we are given $I \in \mathbb{C}^{N+1}$ and want to determine $V \in \mathbb{C}^{N+1}$ from $I = YV$. In Chapter 4.2.3.4 we study sufficient conditions under which Y is invertible. For large networks taking the inverse of Y can be difficult computationally even when it exists. In this section we present a common method for solving $I = YV$ using LU factorization of Y, i.e., factorize Y into $Y = LU$ where L is a lower triangular matrix with all diagonal entries being 1 and U an upper triangular matrix. Any square matrix $A \in \mathbb{C}^{n\times n}$ has an LU factorization after possibly an appropriate re-ordering of the rows, i.e., there exists a permutation matrix P such that $PA = LU$ for some L, U. If A is invertible then it admits an LU factorization without permutation (i.e., $A = LU$ for some L, U) if and only if all its leading principal minors are nonzero.[6] In that case, the LU factorization is unique. For a singular A, necessary and sufficient conditions for the existence and uniqueness of LU factorization are known but are more involved.

Possibly after an appropriate permutation of Y (such that e.g. $Y_{11} \neq 0$), we can compute the entries of L and U recursively. From

$$\begin{bmatrix} Y_{00} & Y_{01} & Y_{02} & \cdots & Y_{0N} \\ Y_{10} & Y_{11} & Y_{12} & \cdots & Y_{1N} \\ Y_{20} & Y_{21} & Y_{22} & \cdots & Y_{2N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{N0} & Y_{N1} & Y_{N2} & \cdots & Y_{NN} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ L_{10} & 1 & 0 & \cdots & 0 \\ L_{20} & L_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ L_{N0} & L_{N1} & L_{N2} & \cdots & 1 \end{bmatrix} \begin{bmatrix} U_{00} & U_{01} & U_{02} & \cdots & U_{0N} \\ 0 & U_{11} & U_{12} & \cdots & U_{1N} \\ 0 & 0 & U_{22} & \cdots & U_{2N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & Y_{NN} \end{bmatrix}$$

we proceed as follows:

1 The 0th row of U is set to the 0th row of Y since $L_{00} = 1$:

$$U_{0j} = Y_{0j}, \qquad j = 0, \ldots, N$$

[6] Consider a matrix $A \in \mathbb{C}^{n\times n}$. Let $I := \{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$, $J := \{j_1, \ldots, j_l\} \subseteq \{1, \ldots, n\}$, and A_{IJ} denote the submatrix obtained from deleting rows not in I and columns not in J.

- If $k = l$, i.e., A_{IJ} is square, then the *minor* M_{IJ} of A is the determinant of the submatrix A_{IJ}.
- If $I = J$, then A_{IJ} is called a *principal submatrix* and M_{IJ} a *principal minor* of A.
- If $I = J = \{1, \ldots, k\}$ with $k \leq n$, then A_{IJ} is called a *leading principal submatrix* of order k and M_{IJ} a *leading principal minor* of order k.

2 To compute row-1 entry L_{10} of L, we have

$$Y_{10} = L_{10}U_{00} \Rightarrow L_{10} = \frac{Y_{10}}{U_{00}}$$

To compute row-1 entries U_{1j} of U, we have for columns $j = 1,\ldots,N$,

$$Y_{1j} = L_{10}U_{0j} + U_{1j} \Rightarrow U_{1j} = Y_{1j} - L_{10}U_{0j}$$

3 In general, to compute row-i entries L_{ij} of L $(i = 2,\ldots,N)$, we have for columns $j = 0,\ldots,i-1$,

$$\begin{aligned}
Y_{i0} &= L_{i0}U_{00} && \Rightarrow && L_{i0} = \frac{Y_{i0}}{U_{00}} \\
Y_{i1} &= L_{i0}U_{01} + L_{i1}U_{11} && \Rightarrow && L_{i1} = \frac{1}{U_{11}}(Y_{i1} - L_{i0}U_{01}) \\
&\vdots && && \vdots \\
Y_{i(i-1)} &= \sum_{j=0}^{i-2} L_{ij}U_{j(i-1)} + L_{i(i-1)}U_{(i-1)(i-1)} && \Rightarrow && L_{i(i-1)} = \frac{1}{U_{(i-1)(i-1)}}\left(Y_{i(i-1)} - \sum_{j=0}^{i-2} L_{ij}U_{j(i-1)}\right.
\end{aligned}$$

To compute row-i entries U_{ij} of U $(i = 2,\ldots,N)$, we have for columns $j = i,\ldots,N$,

$$\begin{aligned}
Y_{ii} &= \sum_{j=0}^{i-1} L_{ij}U_{ji} + U_{ii} && \Rightarrow && U_{ii} = Y_{ii} - \sum_{j=0}^{i-1} L_{ij}U_{ji} \\
Y_{i(i+1)} &= \sum_{j=0}^{i-1} L_{ij}U_{j(i+1)} + U_{i(i+1)} && \Rightarrow && U_{i(i+1)} = Y_{i(i+1)} - \sum_{j=0}^{i-1} L_{ij}U_{j(i+1)} \\
&\vdots && && \vdots \\
Y_{iN} &= \sum_{j=0}^{i-1} L_{ij}U_{jN} + U_{iN} && \Rightarrow && U_{iN} = Y_{iN} - \sum_{j=0}^{i-1} L_{ij}U_{jN}
\end{aligned}$$

Once the factorization is obtained we have $I = YV = LUV$. Hence, given I, V can be solved in two steps from:

$$I = L\tilde{V} \tag{4.20a}$$

$$\tilde{V} = UV \tag{4.20b}$$

In step 1, $\tilde{V}$ is solved using (4.20a) by forward substitution (compute $\tilde{V}_1$ then $\tilde{V}_2$ and so on). In step 2, V is solved using (4.20b) by backward substitution (compute V_n then V_{n-1} and so on).

Example 4.6. Suppose

$$Y = \begin{bmatrix} 2(0.5-j)+j0.5 & -0.5+j & -0.5+j \\ -0.5+j & (0.5-j)+j0.1 & 0 \\ -0.5+j & 0 & (0.5-j)+j0.2 \end{bmatrix}$$

Then

$$Y = \begin{bmatrix} 1 & 0 & 0 \\ -0.6154 + j0.0769 & 1 & 0 \\ -0.6154 + j0.0769 & -1.6763 + j0.8960 & 1 \end{bmatrix} \begin{bmatrix} 1 - j1.5 & -0.5 + j & -0.5 + j \\ 0 & 0.2692 - j0.2462 & -0.2308 + j0.65 \\ 0 & 0 & 0.4682 + j1.156 \end{bmatrix}$$

Given I, V can be obtained in two steps: solve for $\tilde{V}$ from:

$$\begin{bmatrix} I_1 \\ I_2 \\ I_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -0.6154 + j0.0769 & 1 & 0 \\ -0.6154 + j0.0769 & -1.6763 + j0.8960 & 1 \end{bmatrix} \begin{bmatrix} \tilde{V}_1 \\ \tilde{V}_2 \\ \tilde{V}_3 \end{bmatrix}$$

and then solve for V from:

$$\begin{bmatrix} \tilde{V}_1 \\ \tilde{V}_2 \\ \tilde{V}_3 \end{bmatrix} = \begin{bmatrix} 1 - j1.5 & -0.5 + j & -0.5 + j \\ 0 & 0.2692 - j0.2462 & -0.2308 + j0.6538 \\ 0 & 0 & 0.4682 + j1.1566 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \\ V_3 \end{bmatrix}$$

4.2.6 Radial network

Suppose

- The network graph G is a (connected) tree.
- Assumption C4.1 holds (i.e., $y^s_{jk} = y^s_{kj}$) and $y^m_{jk} = y^m_{kj} = 0$ for all $(j,k) \in E$.

Distribution systems are mostly radial, i.e., its graph G is a tree. The second assumption is reasonable if all (j,k) model distribution lines (not transformers) where shunt admittances (y^m_{jk}, y^m_{kj}) are often negligible.

Inverses of reduced incidence and admittance matrices $(\hat{C}, \hat{Y})$.

Under these assumption the admittance matrix Y is complex symmetric and has zero row and column sums. Such a matrix is sometimes called a complex Laplacian matrix. From (4.12), we can write

$$Y \;=\; C\, D^s_y C^{\mathsf{T}} \tag{4.21}$$

where the incidence matrix C is defined in (4.11) and the $N \times N$ diagonal matrix $D^s_y := \mathrm{diag}\left(y^s_l, l \in E\right)$ of series admittances y^s_l is nonsingular. Clearly C is singular. The null space $\mathrm{null}(C^{\mathsf{T}}) = \mathrm{span}(\mathbf{1})$ and its $(N+1) \times N$ pseudo-inverse is $(C^{\mathsf{T}})^{\dagger} = C\left(C^{\mathsf{T}}C\right)^{-1}$ (Exercise 5.2). Hence Y is nonsingular with $\mathrm{null}(Y) = \mathrm{span}(\mathbf{1})$. Consider the *reduced incidence matrix* $\hat{C}$ obtained from C by removing its row corresponding to the reference bus 0 and the *reduced admittance matrix* $\hat{Y}$ obtained from Y by removing the row and column corresponding to the reference bus 0. We now show that, for a radial network, both of the $N \times N$ matrices $\hat{C}$ and $\hat{Y}$ are invertible. Moreover the inverse $\hat{Y}^{-1}$ has a very useful structure.

Denote by c_0^T the first row of the incidence matrix C corresponding to bus 0 and by $\hat{C}$ the $N \times N$ submatrix consisting of the remaining rows of C:

$$C =: \begin{bmatrix} c_0^\mathsf{T} \\ \hat{C} \end{bmatrix} \tag{4.22a}$$

The submatrix $\hat{C}$ is called the *reduced incidence matrix*. Then

$$Y = \begin{bmatrix} c_0^\mathsf{T} \\ \hat{C} \end{bmatrix} D_y^s \begin{bmatrix} c_0 & \hat{C}^\mathsf{T} \end{bmatrix} = \begin{bmatrix} c_0^\mathsf{T} D_y^s c_0 & c_0^\mathsf{T} D_y^s \hat{C}^\mathsf{T} \\ \hat{C} D_y^s c_0 & \hat{C} D_y^s \hat{C}^\mathsf{T} \end{bmatrix} =: \begin{bmatrix} Y_{00} & Y_{01} \\ Y_{10} & \hat{Y} \end{bmatrix} \tag{4.22b}$$

Hence the $N \times N$ reduced admittance matrix is $\hat{Y} = \hat{C} D_y^s \hat{C}^\mathsf{T}$. Suppose the lines are directed with an arbitrary orientation. Let T_j denote the subtree rooted at bus j, including j, and P_j denote the unique path from bus 0 to bus j. Buses k in T_j are called descendants of j. If $k \in \mathsf{T}_j$ and they are adjacent, $(j,k) \in E$, then j is called a parent of k. We use "$l \in \mathsf{P}_j$" to mean a directed line l in P_j that points away from bus 0, and "$-l \in \mathsf{P}_j$" to mean a directed line l in P_j that points towards bus 0. The proof of the next theorem is left as Exercise 4.11.

Theorem 4.10 (Radial network: inverses of $\hat{C}$ and $\hat{Y}$)**.** Consider a radial network for which G is a (connected) tree. Suppose assumption C4.1 holds (i.e., $y_{jk}^s = y_{kj}^s$) and $y_{jk}^m = y_{kj}^m = 0$ for all $(j,k) \in E$.

1 The reduced incidence matrix $\hat{C}$ is nonsingular and

$$\left[\hat{C}^{-1}\right]_{lj} = \begin{cases} -1 & l \in \mathsf{P}_j \\ 1 & -l \in \mathsf{P}_j \\ 0 & \text{otherwise} \end{cases} \tag{4.23}$$

Furthermore $\hat{C}^{-\mathsf{T}} c_0 = -\mathbf{1}$ where $\hat{C}^{-\mathsf{T}} := \left(\hat{C}^\mathsf{T}\right)^{-1}$.

2 The reduced admittance matrix $\hat{Y}$ is nonsingular and $\hat{Z} := \hat{Y}^{-1} = \hat{C}^{-\mathsf{T}} D_z^s \hat{C}^{-1}$, i.e.,

$$\hat{Z}_{jk} = \sum_{l \in \mathsf{P}_j \cap \mathsf{P}_k} z_l^s = \sum_{l \in \mathsf{P}_j \cap \mathsf{P}_k} 1/y_l^s \tag{4.24}$$

where $D_z^s := \operatorname{diag}\left(1/y_{jk}^s, (j,k) \in E\right)$. Hence $\hat{Z}_{jk}$ is the sum of impedances on the common segment of the unique paths from the reference bus 0 to buses j and k.

3 Suppose i is a parent of j, i.e., $(i,j) \in E$ and $j \in \mathsf{T}_i$. Then

$$\hat{Z}_{jk} - \hat{Z}_{ik} = \begin{cases} z_{ij}^s & \text{if} \quad k \in \mathsf{T}_j \\ 0 & \text{if} \quad k \notin \mathsf{T}_j \end{cases}$$

Remark 4.6. 1 The nodal voltages and currents $(\hat{V}, \hat{I})$ at non-reference buses are not related by $\hat{I} = \hat{Y}\hat{V}$. From (4.22b) they are related by

$$\hat{I} = \left(\hat{C} D_y^s c_0\right)\left(c_0^\mathsf{T} D_y^s c_0\right)^{-1} I_0 + \underbrace{\left(\hat{Y} - \left(\hat{C} D_y^s c_0\right)\left(c_0^\mathsf{T} D_y^s c_0\right)^{-1}\left(c_0^\mathsf{T} D_y^s \hat{C}^\mathsf{T}\right)\right)}_{Y/Y_{00}} \hat{V}$$

If the current injection $I_0 = 0$ then $\hat{I} = (Y/Y_{00})\hat{V}$ where the $N \times N$ matrix Y/Y_{00} is the Kron reduction of Y studied in Chapter 4.2.4.1.

2 Corollary 4.8 and Theorem 4.8 says roughly that, for a general network, sufficient conditions for a strict leading submatrix Y_{22}, such as $\hat{Y}$, to be nonsingular are $g^s_{jk} > 0$ for all lines $(j,k) \in E$ or $g^s_{jk} \geq 0$, $b^s_{jk} \leq 0$ for all $(j,k) \in E$. In the former case, $\mathrm{Re}(Y_{22}) \succ 0$ whereas in the latter case, $\mathrm{Re}(Y_{22}) - \mathrm{Im}(Y_{22}) \succ 0$. Theorem 4.10 shows that, for a radial network, $\hat{Y}$ is always nonsingular, even though the positive definite properties may not hold.

3 The nonsingularity of $\hat{Y}$ and the simple structure of its inverse $\hat{Z}$ originate from the inverse $\hat{C}^{-1}$ in (4.23) of the reduced incidence matrix $\hat{C}$ of a tree graph, and are independent of whether the "weight matrix" D^s_y is real or complex, positive or not, as long as D^s_y is nonsingular. It therefore applies to the real Laplacian matrix $L := CBC^\mathsf{T}$ of the DC power flow model of Chapter 4.6.2, the linear DistFlow model of Chapter 5.4.2 (see Theorem 5.3), and the linearized polar-form power flow model of Chapter 7.15. The expression (4.24) for $\hat{Z} = \hat{Y}^{-1}$ is particularly useful for various applications in radial networks. We illustrate its application for voltage control in Chapter 7.2 and topology identification in Chapter 7.3. □

Radiality condition.

Many applications can be formulated as a constrained optimization problem, e.g., state estimation, voltage regulation, feeder reconfiguration, or topology identification. Some of these applications involve computing an operational network from a set of possibilities, e.g. feeder reconfiguration and topology identification. A common setup in these applications assumes that a typically meshed infrastructure network is given. Some of the lines contain switches that can be opened or closed. The switches are configured so that at any time the operational network is a spanning tree that connects all nodes. Let there be $N+1$ nodes and $M \geq N+1$ lines in the infrastructure network, and assume without loss of generality that every line has a switch that can be configured. Our goal is to identify/optimally choose the set of switches that are/should be closed. As part of an optimization problem, this can be specified as two constraints:

- The number of switches that are closed should be exactly N.
- The resulting network should be connected.

These two conditions ensure that the resulting graph is a (connected) tree.

A convenient way to specify the second condition is the following linear constraint from [13] on the reduced incidence matrix $\hat{C}$ of the resulting network, defined in (4.22a), among an arbitrary set of $(N+1) \times N$ incidence matrices C. It says that a network is a (connected) tree if and only if there is a power flow solution when all non-reference buses inject a unit of power into the network. This property is used in [13] for joint optimization of feeder reconfiguration and volt/var control on a distribution grid.

Lemma 4.11 (Connectivity)**.** Suppose a network G has $N+1$ buses and N lines with a reduced incidence matrix $\hat{C}$. It is connected (i.e., a tree) if and only if there exists a line flow $P \in \mathbb{R}^N$ such that $\hat{C}P = \mathbf{1}$.

Proof Exercise 4.11 shows that if the network is radial and connected then $\hat{C}$ is invertible, and therefore $P = \hat{C}^{-1}\mathbf{1}$ is well defined. Conversely suppose there exists P that satisfies $\hat{C}P = \mathbf{1}$. Since there are $N+1$ buses and only N lines, the network is connected if and only if it is a tree. Suppose then the network is not connected. Consider a maximal connected component that does not contain the reference bus 0, and let $N_1 \subsetneq \overline{N}$ denote its nodes. Without loss of generality we can partition $\hat{C}$ into a block-diagonal matrix according to nodes in N_1 and those in its complement $N_0 := \overline{N} \setminus N_1$:

$$\hat{C} =: \begin{bmatrix} \hat{C}_0 & 0 \\ 0 & \hat{C}_1 \end{bmatrix}$$

where $\hat{C}_1$ is the (full) incidence matrix of the maximal connected component N_1. Since $C^\mathsf{T}\mathbf{1} = 0$ we have $\hat{C}_1^\mathsf{T}\mathbf{1}_1 = 0$ (whereas $\hat{C}_0^\mathsf{T}\mathbf{1}_0$ may not be the zero vector as $\hat{C}_0$ is the reduced incidence matrix of the subgraph N_0 that contains bus 0). This means that $\mathbf{1}_1$ is in the null space of $\hat{C}_1^\mathsf{T}$ and therefore orthogonal to the range space of $\hat{C}_1$, i.e., there does not exist any P_1 such that $\hat{C}_1 P_1 = \mathbf{1}_1$. This contradicts $\hat{C}P = \mathbf{1}$ for some P. □

4.2.7 Summary

We have explained how to model a power network as a graph with lines parameterized by admittances (y^s_{jk}, y^m_{jk}) and (y^s_{kj}, y^m_{kj}). This can be described by an admittance matrix Y which is complex symmetric if and only if $y^s_{jk} = y^s_{kj}$. The equation $I = YV$ expresses nodal current balance due to KCL. We derive structural properties of Y and its Kron reduction Y/Y_{22}, especially sufficient conditions under which Y is invertible and Y/Y_{22} exists. Finally we have shown that the reduced admittance matrix $\hat{Y}$ of a connected radial network is always invertible, because the reduced incidence matrix $\hat{C}$ is always invertible, and its inverse $\hat{Y}^{-1}$ has a simple structure that we will use in Chapters 7.2 and 7.3 for voltage control and topology identification respectively.

4.3 Network models: sV relation

In Chapter 4.2 we model a power network by its admittance matrix Y that relates linearly the nodal current injections and voltages, $I = YV$. This is simple as it involves linear equations only. Given (V, I) the power injection at each node j can be computed as $s_j = V_j I_j^\mathsf{H}$. All other quantities, such as line power flows or real power loss over a network, can be computed from V (Exercise 4.12). In many applications however loads

and generators are not specified as current or voltage sources. They may be described instead in terms of power injections or removals. For instance, for electric vehicle charging, the travel need is specified in terms of the number of miles required which translates to the amount of energy in kWh required that must be delivered by a deadline. For example it requires roughly 3 kWh for an electric vehicle to travel 10 miles. Hence a charging facility is often characterized by its power requirement to support a certain electric vehicle charging capacity. In this section we present power flow equations that describe the relation between nodal power injections s_j and voltages V_j on the network. As we will see this involves nonlinear equations which are much more difficult to solve.

We often use s_j to denote both the complex number $p_j + \mathbf{i} q_j \in \mathbb{C}$ and the real pair $(p_j, q_j) \in \mathbb{R}^2$ depending on the context.

4.3.1 Complex form

The *bus injection model* (BIM) in its complex form is defined by power balance $s_j = \sum_{k:j\sim k} S_{jk}$ at each node j where S_{jk} are sending-end line powers from j to its neighbors k. Given line admittances (y^s_{jk}, y^m_{jk}) and (y^s_{kj}, y^m_{kj}), the power flows on line $(j,k) \in E$ are

$$S_{jk} := V_j I_{jk}^{\mathsf{H}} = \left(y^s_{jk}\right)^{\mathsf{H}} \left(|V_j|^2 - V_j V_k^{\mathsf{H}}\right) + \left(y^m_{jk}\right)^{\mathsf{H}} |V_j|^2 \tag{4.25a}$$

$$S_{kj} := V_k I_{kj}^{\mathsf{H}} = \left(y^s_{kj}\right)^{\mathsf{H}} \left(|V_k|^2 - V_k V_j^{\mathsf{H}}\right) + \left(y^m_{kj}\right)^{\mathsf{H}} |V_k|^2 \tag{4.25b}$$

This leads to the power flow equations that relate power injections and voltages:

$$s_j = \sum_{k:j\sim k} \left(y^s_{jk}\right)^{\mathsf{H}} \left(|V_j|^2 - V_j V_k^{\mathsf{H}}\right) + \left(y^m_{jj}\right)^{\mathsf{H}} |V_j|^2, \qquad j \in \overline{N} \tag{4.26a}$$

where, from (4.10b), the total shunt admittance $y^m_{jj} := \sum_{k:j\sim k} y^m_{jk}$ associated with bus j is the sum of shunt admittances y^m_{jk} of all lines (j,k) incident on bus j. We can also express (4.26a) in terms of the elements of the admittance matrix Y as

$$s_j = \sum_{k=0}^{N} Y_{jk}^{\mathsf{H}} V_j V_k^{\mathsf{H}}, \qquad j \in \overline{N} \tag{4.26b}$$

where Y is given by:

$$Y_{jk} = \begin{cases} -y^s_{jk}, & j \sim k \quad (j \neq k) \\ \sum_{i:j\sim i} \left(y^s_{ji} + y^m_{ji}\right) & j = k \\ 0 & \text{otherwise} \end{cases} \tag{4.26c}$$

When the total shunt admittance $y^m_{jj} = \sum_{i:j\sim i} y^m_{ji} = 0$, (4.26a) reduces to

$$s_j = \sum_{k:j\sim k} \left(y^s_{jk}\right)^{\mathsf{H}} \left(|V_j|^2 - V_j V_k^{\mathsf{H}}\right), \qquad j \in \overline{N}$$

For convenience we include V_0 in the vector variable $V := (V_j, j \in \overline{N})$ with the understanding that $V_0 := 1\angle 0°$ is fixed. There are $N+1$ equations in (4.26a) in $2(N+1)$ complex variables $(s_j, V_j, j \in \overline{N})$.

This model does not require assumption C4.1.

Remark 4.7 (Nodal devices)**.** If bus j in Remark 4.1 includes, in addition, a power source with a fixed power injection σ_j^p, then s_j is the net bus injection (assuming all neutrals are grounded and all voltages are defined with respect to the ground):

$$s_j = \underbrace{-\left(z_j^{v\mathsf{H}}\right)^{-1}\left(|V_j|^2 - V_j E_j^{\mathsf{H}}\right)}_{\text{voltage source}} + \underbrace{V_j\left(J_j - y_j^c V_j\right)^{\mathsf{H}}}_{\text{current source}} - \underbrace{y_j^{a\mathsf{H}}|V_j|^2}_{\text{shunt admittance}} + \underbrace{\sigma_j^p}_{\text{power source}}$$

and (4.26a) becomes:

$$\begin{aligned} &-\left(z_j^{v\mathsf{H}}\right)^{-1}\left(|V_j|^2 - V_j E_j^{\mathsf{H}}\right) + V_j\left(J_j - y_j^c V_j\right)^{\mathsf{H}} - y_j^{a\mathsf{H}}|V_j|^2 + \sigma_j^p \\ &= \sum_{k:j\sim k}\left(y_{jk}^s\right)^{\mathsf{H}}\left(|V_j|^2 - V_j V_k^{\mathsf{H}}\right) + \left(y_{jj}^m\right)^{\mathsf{H}}|V_j|^2, \qquad j \in \overline{N} \end{aligned}$$

□

4.3.2 Polar form

We may alternatively treat (4.26) as $2(N+1)$ equations in $4(N+1)$ real variables $(p_j, q_j, |V_j|, \theta_j, j \in \overline{N})$ where $s_j := p_j + \mathbf{i}q_j$ are the complex injections and $V_j := |V_j|\, e^{\mathbf{i}\theta_j}$ are the complex voltages. Let $y_{jk}^s =: g_{jk}^s + \mathbf{i}b_{jk}^s$ denote the series admittance and $y_{jk}^m =: g_{jk}^m + \mathbf{i}b_{jk}^m$ the shunt admittance of line (j,k) from j to k, and similarly (y_{kj}^s, y_{kj}^m) in the opposite direction. As discussed in Remark 4.5, if (j,k) models a transmission or distribution line then usually $g_{jk}^s \geq 0$, $b_{jk}^s < 0$ (inductive line), $g_{jk}^m \geq 0$, but $b_{jk}^m \geq 0$ (capacitive shunt). Moreover $b_{jk}^s + b_{jk}^m \leq 0$ typically since $|b_{jk}^m|$ is usually much smaller than $|b_{jk}^s|$.

Substituting all this into (4.26) the admittance matrix is defined by

$$Y_{jk} = \begin{cases} -(g_{jk}^s + \mathbf{i}b_{jk}^s), & j \sim k \quad (j \neq k) \\ \sum_{i:j\sim i}\left(g_{ji}^s + g_{ji}^m\right) + \mathbf{i}\sum_{i:j\sim i}\left(b_{ji}^s + b_{ji}^m\right) & j = k \\ 0 & \text{otherwise} \end{cases}$$

and the power flow equations become:

$$s_j = \sum_{k:k\sim j}\left((g_{jk}^s + g_{jk}^m) - \mathbf{i}(b_{jk}^s + b_{jk}^m)\right)|V_j|^2 - \sum_{k:k\sim j}\left(g_{jk}^s - \mathbf{i}b_{jk}^s\right)|V_j|\,|V_k|\, e^{\mathbf{i}\theta_{jk}} \qquad j \in \overline{N}$$

where $\theta_{jk} := \theta_j - \theta_k$ is the voltage phase angle difference across each line $(j,k) \in E$.

Then we can write (4.26a) in the polar form:

$$p_j = \sum_{k:k\sim j} \left(g^s_{jk} + g^m_{jk}\right)|V_j|^2 - \sum_{k:k\sim j} |V_j||V_k| \left(g^s_{jk}\cos\theta_{jk} + b^s_{jk}\sin\theta_{jk}\right), \quad j \in \overline{N} \tag{4.27a}$$

$$q_j = -\sum_{k:k\sim j} \left(b^s_{jk} + b^m_{jk}\right)|V_j|^2 - \sum_{k:k\sim j} |V_j||V_k| \left(g^s_{jk}\sin\theta_{jk} - b^s_{jk}\cos\theta_{jk}\right), \quad j \in \overline{N} \tag{4.27b}$$

This model does not require assumption C4.1.

4.3.3 Cartesian form

The power flow equations (4.26) or (4.27) can also be reformulated in the real domain by writing V_j in terms of its real and imaginary components (c_j, d_j), i.e., $V_j =: c_j + \mathbf{i}d_j$. Then (4.27) becomes (using $c_j = |V_j|\cos\theta_j$ and $d_j = |V_j|\sin\theta_j$)

$$p_j = \sum_{k:k\sim j} \left(g^s_{jk} + g^m_{jk}\right)\left(c_j^2 + d_j^2\right) - \sum_{k:k\sim j} \left(g^s_{jk}(c_jc_k + d_jd_k) + b^s_{jk}(d_jc_k - c_jd_k)\right), \quad j \in \overline{N} \tag{4.28a}$$

$$q_j = -\sum_{k:k\sim j} \left(b^s_{jk} + b^m_{jk}\right)\left(c_j^2 + d_j^2\right) - \sum_{k:k\sim j} \left(g^s_{jk}(d_jc_k - c_jd_k) - b^s_{jk}(c_jc_k + d_jd_k)\right), \quad j \in \overline{N} \tag{4.28b}$$

These are $2(N+1)$ quadratic equations in $4(N+1)$ variables $(p_j, q_j, c_j, d_j,\ j \in \overline{N})$. This model does not require assumption C4.1.

4.3.4 Types of buses

Each set of power flow equations (4.26)(4.27)(4.28) is a set of $2(N+1)$ nonlinear real equations in $4(N+1)$ real variables $(p_j, q_j, |V_j|, \theta_j, j \in \overline{N})$ Given any $2(N+1)$ of these real variables, these equations can be used to solve for the remaining $2(N+1)$ real variables. There can be zero, unique or multiple solutions. Solving for these solutions is the *power flow* or *load flow* problem (Chapter 4.4).

A popular formulation of the power flow problem uses the polar form where each bus j is classified into one of three types based on which two of the four real variables $(p_j, q_j, |V_j|, \theta_j)$ are specified:

- *PV bus.* This is a bus where the real power injection p_j and the voltage magnitude $|V_j|$ are specified and the reactive power injection q_j and voltage angle θ_j are to be determined. It usually models a bus with a conventional generator.
- *PQ bus.* This is a constant-power bus where the injection (p_j, q_j) is specified and the complex voltage $|V_j|\,e^{j\theta_j}$ is to be determined. It usually models a load but can also model a renewable generator with undispatchable generation.
- *Slack bus.* Bus 0 is taken as a slack bus where $V_0 = |V_0|\,\angle 0^\circ$ is specified and the injection $s_0 = (p_0, q_0)$ is to be determined. This is usually used for mathematical convenience to avoid an ill specified power flow problem that has no solution.

A slack bus (or a set of slack buses) is needed because power needs to be balanced over the network. For example if the resistance of every line is zero then $\sum_j p_j$ must be zero. If all buses are *PV* or *PQ* buses then all active powers p_j are specified; if the specified values do not satisfy power balance then the set of power flow equations will have no solution. This is resolved by taking an arbitrary bus (denoted by bus 0 here) as a slack bus with its power injection s_0 unspecified in order to balance power. For instance a distribution system with a substation at bus 0 and N constant power loads or generations can be modeled by a slack bus and N PQ buses with V_0 and $(p_j, q_j, j \in N)$ specified. The power flow problem solves the power flow equations for the N complex voltages $(V_j, j \in N)$, and the power injection s_0 (see Chapter 4.4).

For optimal power flow problems p_j and $|V_j|$ on generator buses or s_j on load buses can be variables as well. For instance economic dispatch optimizes real power generations p_j at generator buses; demand response optimizes demands s_j at load buses; and volt/var control optimizes reactive powers q_j at capacitor banks, tap changers, or inverters. We will discuss optimal power flow problems in Part II of the book.

4.4 Computation methods

Suppose we are given a set of power flow equations in the bus injection model. Suppose $2(N+1)$ of the $4(N+1)$ real variables are specified and we are interested in solving for the remaining variables. We now present four solution methods. These methods do not require assumption C4.1.

An important application of iterative algorithms for solving a system of equations is in optimization where the system of equations specify an optimality condition (e.g. the KKT condition). We will therefore postpone the convergence analysis of iterative algorithms to Chapter 8.6 after we have introduced a basic theory of and popular algorithms for optimization.

4.4.1 Gauss-Seidel algorithm

Consider the power flow equations (4.26a) in the complex form. To illustrate the basic idea consider first the case with a slack bus and load buses only.

Case 1: Given V_0 and $(s_1, \ldots, s_N)$, determine s_0 and $(V_1, \ldots, V_N)$. The power flow equations are:

$$s_0 = \sum_k Y_{0k}^{\mathsf{H}} V_0 V_k^{\mathsf{H}} \tag{4.29a}$$

$$s_j = \sum_k Y_{jk}^{\mathsf{H}} V_j V_k^{\mathsf{H}}, \qquad j \in N \tag{4.29b}$$

Once we have computed $(V_1,\ldots,V_N)$, s_0 can be evaluated using (4.29a). Hence the main task is to compute $(V_1,\ldots,V_N)$ from (4.29b). We have from (4.29b):

$$\frac{s_j^{\mathsf{H}}}{V_j^{\mathsf{H}}} = Y_{jj}V_j + \sum_{\substack{k=0\\k\neq j}}^{N} Y_{jk}V_k, \qquad j \in N$$

Rearrange to obtain

$$V_j = \frac{1}{Y_{jj}}\left(\frac{s_j^{\mathsf{H}}}{V_j^{\mathsf{H}}} - \sum_{\substack{k=0\\k\neq j}}^{N} Y_{jk}V_k\right) =: f_j(V_1,\ldots,V_N), \qquad j \in N$$

Hence a power flow solution $V := (V_1,\ldots,V_N)$ is a fixed point of $f := (f_1,\ldots,f_N)$ with

$$V = f(V)$$

The Gauss algorithm is the standard fixed point iteration $V(t+1) = f(V(t))$, or

$$\begin{aligned} V_1(t+1) &= f_1(V_1(t),\ldots,V_N(t)) \\ V_2(t+1) &= f_2(V_1(t),\ldots,V_N(t)) \\ &\vdots \\ V_N(t+1) &= f_n(V_1(t),\ldots,V_N(t)) \end{aligned}$$

Starting from an initial vector $V(0)$ (e.g., $V_j(0) = 1\angle 0°$ pu for all j), the Gauss algorithm produces a sequence $V(1), V(2), \ldots$. If the sequence converges to a limit $V^{\lim}$ then $V^{\lim}$ is a fixed point of f and a power flow solution.

When $V_2(t+1)$ is to be computed, $V_1(t+1)$ is already known and can be used in the computation of $V_2(t+1)$, and so on. This is the Gauss-Seidel algorithm where the latest value $V_i(t+1)$ is used to compute $V_{j+1}(t+1)$ for $j > i$:

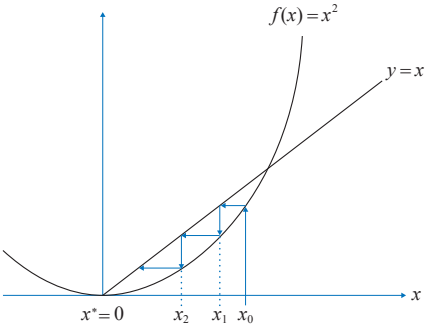
$$\begin{aligned} V_1(t+1) &= f_1(V_1(t),V_2(t),\ldots,V_N(t)) \\ V_2(t+1) &= f_2(V_1(t+1),V_2(t),\ldots,V_N(t)) \\ &\vdots \\ V_N(t+1) &= f_N(V_1(t+1),\ldots,V_{N-1}(t+1),V_N(t)) \end{aligned}$$

Case 2: Given $(V_0,V_1,\ldots,V_m)$ *and* $(s_{m+1},\ldots,s_N)$*, determine* $(s_0,s_1,\ldots,s_m)$ *and* $(V_{m+1},\ldots,V_N)$. In this case, first determine $(V_{m+1},\ldots,V_N)$ from the reduced set of power flow equations (4.29b) for $j = m+1,\ldots,N$, using the same algorithm. Then determine $(s_0,s_1,\ldots,s_m)$ given $(V_0,\ldots,V_N)$.

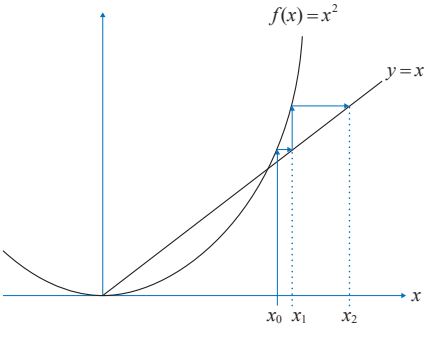
The Gauss-Seidel algorithm is simple and does not require the evaluation of any derivatives. If the function f is a contraction mapping then it has a unique fixed point $V^{\lim}$ and the Gauss or Gauss-Seidel algorithm converges geometrically to $V^{\lim}$. The

formal definition and convergence properties of a contraction mapping are studied in Chapter 8.6.1 (but see Exercise 4.13 for an example). Otherwise there is no guarantee that the algorithms will converge, but if it does, it produces a fixed point which is a power flow solution $V^{\lim}$. Whether it converges can depend on the choice of the initial vector $V(0)$, as the next example shows. The convergence of Gauss-Seidel algorithm is studied in Chapter 8.6.2.

Example 4.7 (Fixed-point iteration)**.** Take for an example $x = f(x) := x^2$ for $x \in \mathbb{R}$ as shown in Figure 4.11. It has two fixed points $x^{\lim} = 0$ or 1. The fixed point iteration



(a) Convergence



(b) Divergence

Figure 4.11 The fixed point iteration $x(t+1) = f(x(t)) := x^2(t)$ is not a contraction mapping and its convergence depends on the initial point $x(0) = x_0$.

$x(t+1) = f(x(t)) = x^2(t)$ converges to $x^{\lim} = 0$ if the initial point $x(0) \in (-1, 1)$ and diverges to positive infinity if $|x(0)| > 1$. The fixed point $x^{\lim} = 0$ is stable in the sense that the iterate $x(t)$ converges back to the origin after a small perturbation. The fixed point $x^{\lim} = 1$ is unstable in the sense that $x(t)$ leaves and will not return after a small perturbation in the positive direction. □

4.4.2 Newton-Raphson algorithm

The Newton-Raphson algorithm is popular for iteratively solving the equation

$$f(x) = 0$$

where $x \in \mathbb{R}^n$ and f is a vector-valued function $f : \mathbb{R}^n \to \mathbb{R}^n$. The iteration is motivated by the Taylor series expansion of f. Suppose we have computed $x(t)$ and wish to determine the next iterate $x(t+1) =: x(t) + \Delta x(t)$. The Taylor series of f around $x(t)$ is

$$f(x(t) + \Delta x(t)) = f(x(t)) \; + \; J(x(t))\Delta x(t) \; + \; \text{higher-order terms}$$

where $J(x(t))$ is the Jacobian of f evaluated at $x(t)$:

$$J(x) := \frac{\partial f}{\partial x}(x) \;\; = \;\; \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \cdots & \frac{\partial f_1}{\partial x_n}(x) \\ \vdots & \vdots & \vdots \\ \frac{\partial f_n}{\partial x_1}(x) & \cdots & \frac{\partial f_n}{\partial x_n}(x) \end{bmatrix}$$

If we ignore the higher-order terms in the Taylor expansion and set $f(x(t+1)) = 0$ then we have

$$J(x(t))\Delta x(t) = -f(x(t)) \tag{4.30}$$

This is illustrated in Figure 4.12. If $J(x(t))$ is invertible then $\Delta x(t) \;\; =$

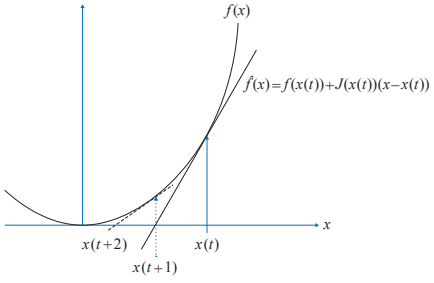


Figure 4.12 Newton-Raphson algorithm: The next iterate $x(t+1)$ is obtained by approximating f by its linear approximation at $x(t)$ and setting the linear approximation $\hat{f}(x) = 0$.

$-J^{-1}(x(t))\, f(x(t))$, yielding the Newton-Raphson iteration:

$$x(t+1) = x(t) \; - \; J^{-1}(x(t))\, f(x(t)) \tag{4.31}$$

In practice we usually do not evaluate the inverse $J^{-1}(x(t))$ except for very small systems. Instead we solve the linear equation (4.30) for $\Delta x(t)$. The next iterate is then $x(t+1) = x(t) + \Delta x(t)$.

We now apply this method to solve the power flow equations in the polar form. To illustrate the idea we consider the case where every bus in the network is either the slack bus (with V_0 specified and s_0 unknown), a PV bus (with $(p_j, |V_j|)$ specified and (q_j, θ_j) unknown), or a PQ bus (with (p_j, q_j) specified and $(\theta_j, |V_j|)$ unknown).

The idea can be extended to more general cases. As mentioned before, (p_j, q_j) can be evaluated directly from the power flow equations once all $(\theta_j, |V_j|)$ are determined. Hence the main task is to solve for those $(\theta_j, |V_j|)$ that are not specified.

Let $N_{pq} \subseteq N$ be the set of PQ buses where $|V_j|$ (as well as θ_j) are unknown. We abuse notation and use N_{pq} to also denote the number $|N_{pq}|$ of buses in N_{pq}. Let

$$\theta := (\theta_j,\ j \in N)$$
$$|V| := (|V_j|,\ j \in N_{pq})$$

i.e., θ collects all unknown phase angles and $|V|$ collects all unknown voltage magnitudes. Rewrite (4.27) as (right-hand sides are given constants):

$$p_j(\theta, |V|) = p_j, \qquad j \in N$$
$$q_j(\theta, |V|) = q_j, \qquad j \in N_{pq}$$

where we have abused notation to use (p_j, q_j) to denote both power injections and as functions of $(\theta, |V|)$ given by:

$$p_j(\theta, |V|) := \sum_{k:k\sim j} \left(g^s_{jk} + g^m_{jk}\right) |V_j|^2 - \sum_{k:k\sim j} |V_j||V_k| \left(g^s_{jk} \cos\theta_{jk} + b^s_{jk} \sin\theta_{jk}\right) \tag{4.32a}$$

$$q_j(\theta, |V|) := -\sum_{k:k\sim j} \left(b^s_{jk} + b^m_{jk}\right) |V_j|^2 - \sum_{k:k\sim j} |V_j||V_k| \left(g^s_{jk} \sin\theta_{jk} - b^s_{jk} \cos\theta_{jk}\right) \tag{4.32b}$$

where $\theta_{jk} := \theta_j - \theta_k$. Define the function $f : \mathbb{R}^{N+N_{pq}} \to \mathbb{R}^{N+N_{pq}}$ by

$$f(\theta, |V|) := \begin{bmatrix} \Delta p(\theta, |V|) \\ \Delta q(\theta, |V|) \end{bmatrix} := \begin{bmatrix} p(\theta, |V|) - p \\ q(\theta, |V|) - q \end{bmatrix} \tag{4.33}$$

where $p := (p_j, j \in N)$, $q := (q_j, j \in N_{pq})$ are constants and

$$p(\theta, |V|) := \begin{bmatrix} p_1(\theta, |V|) \\ \vdots \\ p_N(\theta, |V|) \end{bmatrix}, \qquad q(\theta, |V|) := \begin{bmatrix} q_1(\theta, |V|) \\ \vdots \\ q_{N_{pq}}(\theta, |V|) \end{bmatrix}$$

Our goal is to compute a root of $f(\theta, |V|) = 0$ iteratively. The Jacobian of f is the $(N + N_{pq}) \times (N + N_{pq})$ matrix

$$J(\theta, |V|) := \begin{bmatrix} \frac{\partial p}{\partial \theta} & \frac{\partial p}{\partial |V|} \\ \frac{\partial q}{\partial \theta} & \frac{\partial q}{\partial |V|} \end{bmatrix} \tag{4.34}$$

Hence the Newton-Raphson algorithm is:

1 Choose an initial point $(\theta(0), |V|(0))$.
2 Iterate until converge (or the maximum number of iterations has been reached):
  1 Solve $(\Delta\theta(t), \Delta|V|(t))$ from

$$J\left(\theta(t), |V|(t)\right) \begin{bmatrix} \Delta\theta(t) \\ \Delta|V|(t) \end{bmatrix} = -\begin{bmatrix} \Delta p(\theta(t), |V|(t)) \\ \Delta q(\theta(t), |V|(t)) \end{bmatrix} \tag{4.35}$$

2 Set

$$\begin{bmatrix} \theta(t+1) \\ |V|(t+1) \end{bmatrix} := \begin{bmatrix} \theta(t) \\ |V|(t) \end{bmatrix} + \begin{bmatrix} \Delta\theta(t) \\ \Delta|V|(t) \end{bmatrix}$$

The right-hand side of (4.35) is defined in (4.33) and represents the mismatch in injections at iteration t. This mismatch is used to compute the increment $(\Delta\theta(t), \Delta|V|(t))$ that updates the current iterate $(\theta(t), |V|(t))$.

The Newton-Raphson algorithm is widely used in industry to compute power flow solution and solve optimal power flow problems. It converges, typically quadratically, to a solution if it starts close to a solution; see Kantorovich Theorem in Exercise 4.15. Like the Gauss-Seidel algorithm, it may not converge if the initial point is far away from a solution. The convergence of the Newton-Raphson algorithm is analyzed in Chapter **??**.

Remark 4.8. Usually the injection q_j at a PV bus j must be constrained within a range. After solving for $(\theta, |V|)$ and evaluating the resulting q_j at bus j, if it hits or exceeds its limit then q_j is set to the limit and bus j is re-classified as a PQ bus with $|V_j|$ (as well as θ_j) to be determined. The updated power flow equations are then re-solved for the remaining unknown quantities.

4.4.3 Fast decoupled algorithm

We now take a closer look at the Jacobian J in (4.34). Using (4.32) it can be shown that the diagonal blocks are (Exercise 4.16):

$$\frac{\partial p_j}{\partial \theta_k} = \begin{cases} -|V_j||V_k|\left(g^s_{jk}\sin\theta_{jk} - b^s_{jk}\cos\theta_{jk}\right), & j \sim k,\ j,k \in N \\ -q_j(\theta,|V|) \ - \left(\sum_{i:i\sim j} b^s_{ji} + b^m_{ji}\right)|V_j|^2, & j = k,\ j \in N \end{cases} \tag{4.36a}$$

$$\frac{\partial q_j}{\partial |V_k|} = \begin{cases} -|V_j|\left(g^s_{jk}\sin\theta_{jk} - b^s_{jk}\cos\theta_{jk}\right), & j \sim k,\ j,k \in N_{pq} \\ \frac{q_j(\theta,|V|)}{|V_j|} \ - \sum_{i:i\sim j}\left(b^s_{ji} + b^m_{ji}\right)|V_j|, & j = k,\ j \in N_{pq} \end{cases} \tag{4.36b}$$

and the off-diagonal blocks are:

$$\frac{\partial p_j}{\partial |V_k|} = \begin{cases} -|V_j|\left(g^s_{jk}\cos\theta_{jk} + b^s_{jk}\sin\theta_{jk}\right), & j \sim k,\ j \in N, k \in N_{pq} \\ \frac{p_j(\theta,|V|)}{|V_j|} + \sum_{i:i\sim j}\left(g^s_{ji} + g^m_{ji}\right)|V_j|, & j = k,\ j,k \in N_{pq} \end{cases} \tag{4.36c}$$

$$\frac{\partial q_j}{\partial \theta_k} = \begin{cases} |V_j||V_k|\left(g^s_{jk}\cos\theta_{jk} + b^s_{jk}\sin\theta_{jk}\right), & j \sim k,\ j \in N_{pq}, k \in N \\ p_j(\theta,|V|) - \sum_{i:i\sim j}\left(g^s_{ji} + g^m_{ji}\right)|V_j|^2, & j = k,\ j \in N_{pq} \end{cases} \tag{4.36d}$$

Hence the sparsity of the network graph induces a sparse Jacobian matrix J.

Moreover if line losses and angle differences θ_{jk} are small then it is reasonable to approximate $g^s_{jk} = g^m_{jk} = 0$ and $\sin\theta_{jk} = 0$ for all $(j,k) \in E$. In this case it can be verified that the off-diagonal blocks are approximately zero (Exercise 4.16), i.e.,

$$\frac{\partial p_j}{\partial |V_k|} \approx 0 \text{ and } \frac{\partial q_j}{\partial \theta_k} \approx 0, \qquad \forall j,k$$

This means that the voltage magnitudes and the real power injections (at the same or different buses) are approximately decoupled, and the voltage angles and the reactive power injections are approximately decoupled. This motivates a fast decoupled algorithm where an approximate Jacobian $\hat{J}$ matrix with the off-diagonal blocks of J set to zero is used in place of J in the Newton-Raphson's algorithm (step 2):

$$\hat{J}(\theta,|V|) := \begin{bmatrix} \frac{\partial p}{\partial \theta} & 0 \\ 0 & \frac{\partial q}{\partial |V|} \end{bmatrix}$$

Then equation (4.35) to compute the increments in the Newton-Raphson algorithm is replaced by the following equations that decouple active and reactive power:

$$\frac{\partial p}{\partial \theta}(\theta(t),|V|(t))\ \Delta\theta(t) = -\Delta p(\theta(t),|V|(t)) \tag{4.37a}$$

$$\frac{\partial q}{\partial |V|}(\theta(t),|V|(t))\ \Delta|V|(t) = -\Delta q(\theta(t),|V|(t)) \tag{4.37b}$$

There are other properties of J one can exploit to obtain symmetric matrices that saves storage and computation in executing the exact Newton-Raphson algorithm; see [1, p. 350–351]. The fast decoupled algorithm (4.37) can be further simplified with more approximations; see [1, p. 353–354].

4.4.4 Holomorphic Embedding Load-flow Method (HELM)

We now explain a solution method from [14] for solving power flow equations that adopts a very different approach from those in Chapters 4.4.1, 4.4.2 and 4.4.3.

Holomorphic functions.

A complex-valued function $f : \mathbb{C} \to \mathbb{C}$ is *complex differentiable* at $z \in \mathbb{C}$ if

$$f'(z) := \lim_{\substack{h \in \mathbb{C} \\ h \to 0}} \frac{f(z+h) - f(z)}{h} \tag{4.38}$$

exists. When $f'(z)$ exists we will call it the *complex derivative* or *derivative* of f at $z \in \mathbb{C}$. Note that $f'(z)$ is generally a complex number. If f is complex differentiable at every $z \in Z \subseteq \mathbb{C}$ then f is called *holomorphic* on Z. Complex differentiability in (4.38) is a much stronger notion than differentiability of real-valued functions because h must approach 0 from all directions in the complex plane; see Chapter A.9 for details. The most important property of holomorphic functions is that they are (complex) analytic, i.e., they can be expressed as a power series. Specifically a complex-valued function $f : Z \to \mathbb{C}$ on an open set $Z \subseteq \mathbb{C}$ is holomorphic on Z if and only if at every point $z_0 \in Z$ there is a neighborhood $B_\delta(z_0) := \{z \in Z : |z - z_0| < \delta\}$ around z_0 such that

$$f(z) = \sum_{k=0}^{\infty} a_k (z - z_0)^k, \qquad z \in B_\delta(z_0) \tag{4.39}$$

wher $a_k = \frac{f^{(k)}(z_0)}{k!}$, i.e., $f(z)$ can be expressed as a Taylor series on $B_\delta(z_0)$. The neighborhood $B_\delta(z_0)$ is called the region of convergence for (4.39).

Power flow equations.

Suppose the voltage phasor V_0 at bus 0 and power injections $s := (s_j, j \in N)$ at buses $j \neq 0$ are given. Bus 0 is referred to as a slack bus where its voltage V_0 is specified and its power injection s_0 is a variable. Our goal is to compute a solution $V := (V_j, j \in N) \in \mathbb{C}^N$ to the complex-form power flow equations:

$$\sum_{k=0}^{N} Y_{jk} V_k = \frac{\bar{s}_j}{\bar{V}_j}, \qquad j \in N \tag{4.40}$$

where Y_{jk} are the jkth entries of the admittance matrix $Y \in \mathbb{C}^{(N+1)\times(N+1)}$ and for $a \in \mathbb{C}$, $\bar{a}$ denotes its complex conjugate. Here is a summary of the HELM procedure (see [14] for details).

Holomorphic embedding

Introduce a new variable $\lambda \in \mathbb{C}$ and embed (4.40) in $\mathbb{C}^{N+1}$ so that the voltage $V := V(\lambda) := (V_j(\lambda), j \in N)$ becomes a vector function of λ, i.e., consider the polynomial equations

$$Y_{j0} V_0 + \sum_{k=1}^{N} Y_{jk} V_k(\lambda) = \frac{\lambda \bar{s}_j}{\bar{V}_j(\bar{\lambda})}, \qquad j \in N \tag{4.41}$$

Note that the denominator on the right-hand side is $\bar{V}_j(\bar{\lambda})$, not $\bar{V}_j(\lambda)$, in order for $V_j(\lambda)$ to be a holomorphic function. Instead of solving (4.40) for V, HELM solves (4.41) rewritten as:

$$Y_{j0}V_0 + \sum_{k=1}^{N} Y_{jk}V_k(\lambda) = \frac{\lambda \bar{s}_j}{\tilde{V}_j(\lambda)}, \quad \bar{Y}_{j0}\bar{V}_0 + \sum_{k=1}^{N} \bar{Y}_{jk}\tilde{V}_k(\lambda) = \frac{\lambda s_j}{V_j(\lambda)}, \quad j \in N \tag{4.42a}$$

$$\tilde{V}_j(\lambda) = \bar{V}_j(\bar{\lambda}), \qquad j \in N \tag{4.42b}$$

for two sets of complex-valued functions $(V(\lambda), \tilde{V}(\lambda)) := (V_j(\lambda), \tilde{V}_j(\lambda), j \in N)$.

At $\lambda = 0$, (4.42a) reduces to

$$Y_{j0}V_0 + \sum_{k=1}^{N} Y_{jk}V_k(0) = 0, \qquad \bar{Y}_{j0}\bar{V}_0 + \sum_{k=1}^{N} \bar{Y}_{jk}\tilde{V}_k(0) = 0, \qquad j \in N$$

Decomposing the admittance matrix $Y =: \begin{bmatrix} W_{00} & W_{10}^{\mathsf{T}} \\ W_{10} & W_{11} \end{bmatrix}$ according to V_0 and $V := (V_j, j \in N)$ where $W_{00} \in \mathbb{C}$ and $W_{11} \in \mathbb{C}^{N \times N}$, the system of equations above becomes

$$W_{11}V(0) = -V_0 W_{10}, \qquad \bar{W}_{11}\tilde{V}(0) = -\bar{V}_0 \bar{W}_{10}$$

where $\bar{W}_{11}$ and $\bar{W}_{10}$ are the componentwise complex conjugates of W_{11} and W_{10} respectively. If W_{11} is nonsingular then the unique solution is

$$V(0) = -V_0 W_{11}^{-1} W_{10}, \qquad \tilde{V}(0) = -\bar{V}_0 \bar{W}_{11}^{-1} \bar{W}_{10} \tag{4.43}$$

Note that the solution $(V(0), \tilde{V}(0))$ satisfies (4.42b) as well. This is the solution driven by the given voltage source V_0 at bus 0 and zero injections at other buses.

The solution to the original power flow equation (4.40) corresponds to a solution $(V(\lambda), \tilde{V}(\lambda))$ of (4.42) at $\lambda = 1$. HELM uses a continuation method to compute this solution, starting from $(V(0), \tilde{V}(0))$ in (4.43).

Power series.

To show that the functions $(V_j(\lambda), \tilde{V}_j(\lambda), j \in N)$ are holomorphic, Gröbner basis can be used to express $\tilde{V}_1, (V_2, \tilde{V}_2), \ldots, (V_N, \tilde{V}_N)$ in terms of V_1 and reduce (4.42a) to a polynomial equation in V_1:

$$\mathbb{P}(V_1) := \sum_{k=0}^{M} p_k(\lambda) V_1^k = 0 \tag{4.44}$$

The degree M of the polynomial in (4.44) is generally exponential in the number N of original variables. This defines an algebraic curve which then implies that $(V_j(\lambda), \tilde{V}_j(\lambda), j \in N)$ are indeed holomorphic functions everywhere except at a finite number of points.

Therefore, for each $j \in N$, we can write $V_j(\lambda)$ and $1/V_j(\lambda)$ as power series in a

neighborhood of $\lambda = 0$, from (4.39),

$$V_j(\lambda) \;=\; \sum_{i=0}^{\infty} a_{ji}\lambda^i, \qquad \frac{1}{V_j(\lambda)} \;=\; \sum_{i=0}^{\infty} b_{ji}\lambda^i, \qquad j \in N \tag{4.45}$$

for some sequences $(a_{ji}, i \geq 0, j \in N)$ and $(b_{ji}, i \geq 0, j \in N)$. Hence $1/\tilde{V}_j(\lambda) = \left(1/V_j(\lambda^{\mathsf{H}})\right)^{\mathsf{H}} = \sum_{i=0}^{\infty} \bar{b}_{ji}\lambda^i$. Substituting into (4.42) we have

$$Y_{j0}V_0 + \sum_{k=1}^{N} Y_{jk} \sum_{i=0}^{\infty} a_{ki}\lambda^i \;=\; \lambda \bar{s}_j \sum_{i=0}^{\infty} \bar{b}_{ji}\lambda^i, \qquad j \in N \tag{4.46a}$$

or in vector form

$$V_0 W_{10} + \sum_{i=0}^{\infty} (W_{11}a_i)\,\lambda^i \;=\; \sum_{i=0}^{\infty} \left(\bar{s} \odot \bar{b}_i\right) \lambda^{i+1} \tag{4.46b}$$

where $s := (s_j, j \in N)$ is the vector of injections at buses $j \neq 0$, and for $i \geq 0$, $a_i := (a_{ji}, j \in N)$ and $b_i := (b_{ji}, j \in N)$ are N-dimensional column vectors of coefficients. For two vectors x and y, $x \odot y$ is the column vector of componentwise products, i.e., $(x \odot y)_j := x_j y_j$. We can compute these coefficients $(a_i, b_i, i \geq 0)$ iteratively from (4.46), as follows. Setting $\lambda := 0$, (4.46) yields, when W_{11} is nonsingular,

$$V_0 W_{10} + W_{11} a_0 \;=\; 0, \qquad \Longrightarrow \qquad a_0 \;=\; -V_0 W_{11}^{-1} W_{10} \tag{4.47a}$$

Differentiating successively (4.46b) with respect to λ and setting $\lambda := 0$ yields

$$W_{11}a_1 \;=\; \bar{s} \odot \bar{b}_0, \qquad \cdots, \qquad W_{11}a_i \;=\; \bar{s} \odot \bar{b}_{i-1}, \qquad \cdots, \tag{4.47b}$$

Since $V_j(\lambda)\left(1/V_j(\lambda)\right) = 1$ for all λ, we have $1 = \left(\sum_{i=0}^{\infty} a_{ji}\lambda^i\right)\left(\sum_{i=0}^{\infty} b_{ji}\lambda^i\right)$ for all λ for $j \in N$, or in vector form

$$\mathbf{1}_N \;=\; \left(\sum_{i=0}^{\infty} a_i\lambda^i\right) \odot \left(\sum_{i=0}^{\infty} b_i\lambda^i\right)$$

where $\mathbf{1}_N$ is the column vector of all 1s of size N. Hence

$$\begin{aligned} \mathbf{1}_N \;&=\; a_0 \odot b_0 + (a_0 \odot b_1 + a_1 \odot b_0)\lambda + (a_0 \odot b_2 + a_1 \odot b_1 + a_2 \odot b_0)\lambda^2 + \cdots \\ &=\; \sum_{i \geq 0} \left(\sum_{k=0}^{i} a_k \odot b_{i-k}\right) \lambda^i, \qquad \forall \lambda \end{aligned} \tag{4.47c}$$

Since (4.47c) holds for all λ, the coefficients of λ^i must be equal on both sides for all $i \geq 0$. From (4.47) we can obtain $(a_i, b_i, i \geq 0)$ iteratively: a_0 from (4.47a) and then b_0 from (4.47c) by equating the coefficients of λ^0:

$$a_0 \;=\; -V_0 W_{11}^{-1} W_{10}, \qquad b_0 \;=\; \mathbf{1}_N \oslash a_0 \tag{4.48a}$$

where, for two vectors x and y, $x \oslash y$ is the column vector of componentwise division,

i.e., $(x \oslash y)_j := x_j/y_j$. For $i \geq 1$, we have from (4.47b) and (4.47b) by equating the coefficients of λ^i, assuming W_{11} is nonsingular,

$$a_i \;=\; W_{11}^{-1}\left(\bar{s} \odot \bar{b}_{i-1}\right), \qquad b_i \;=\; -\left(\sum_{k=1}^{i} a_k \odot b_{i-k}\right) \oslash a_0, \qquad i \geq 1 \tag{4.48b}$$

With the coefficients $(a_i, i \geq 0) = (a_{ji}, j \in N, i \geq 0)$ from (4.48), the solution $V_j(\lambda)$ is given by (4.45) as a power series in λ. In practice only an approximation $\hat{V}_j(\lambda) := \sum_{i=0}^{K} a_{ji}\lambda^i$ of $V_j(\lambda)$ with a finite number of terms is computed.

Analytic continuation.

We are interested in $V(\lambda) := (V_j(\lambda), j \in N)$ at $\lambda = 1$. Even though, for $\lambda \in B_\delta(0)$ in the region of convergence around $\lambda = 0$,

$$V_j(\lambda) \;=\; \sum_{i=0}^{\infty} a_{ji}\lambda^i, \qquad j \in N$$

and we have the coefficients $(a_i, i \geq 0) = (a_{ji}, j \in N, i \geq 0)$ from (4.48), the radius δ of convergence is typically much smaller than 1 so we may not be able to simply substitute $\lambda = 1$ into the power series as the infinite sum may not converge. To deal with this, Padé approximation is used to approximate the power series. Padé approximation approximates a power series by a rational function and typically has much better convergence properties than a power series (Taylor series). The power solution $V_j(\lambda)$ is computed as the analytic continuation of the Padé approximation, starting from $V_j(0)$ in (4.43). See [14] for details.

Example 4.8 (Two-bus system [14])**.** □

4.5 Properties of power flow solutions

Example 4.9 (Two-bus network)**.** Consider two buses 1 and 2 connected by a line with admittance $y = g + \mathbf{i}b$ with $g > 0, b < 0$. Assume zero charging admittances, and we ignore reactive powers. Assume $V_1 := 1\angle 0^\circ$ and $V_2 = e^{\mathbf{i}\theta}$, i.e., voltage magnitudes are fixed at 1 pu. Then the real power injections (p_1, p_2) depend on θ according to the power flow equations in polar form are:

$$p_1 \;:=\; p_1(\theta) := g \;-\; g\cos\theta \;-\; b\sin\theta \tag{4.49a}$$

$$p_2 \;:=\; p_2(\theta) := g \;-\; g\cos\theta \;+\; b\sin\theta \tag{4.49b}$$

or in vector form

$$P - g\mathbf{1} = A\begin{bmatrix}\cos\theta \\ \sin\theta\end{bmatrix} \tag{4.50}$$

where $\mathbf{1} := [1\ 1]^\mathsf{T}$ and A is an invertible (indeed negative definite) matrix:

$$A := \begin{bmatrix} -g & -b \\ -g & b \end{bmatrix}$$

Show that, as θ ranges from 0 to 2π, $(p_1(\theta), p_2(\theta))$ traces out an ellipse.

4.6 Linear power flow model

4.6.1 Laplacian matrix L

In this section we collect some basic properties of graph Laplacian matrix L that are useful in the analysis of linearized models such as the DC power flow model (4.55). In this section, L is taken to be a real symmetric matrix with zero row and column sums. It is the admittance matrix of the linearized power flow models. These properties are extensively used in, e.g., electricity market (Chapter 6.4), voltage control (Chapter 7.2), topology identification (Chapter 7.3), cascading failure, and other power system applications where a linearized model is applicable.

Consider a graph $G := (N, E)$ where $N := \{1, \ldots, n\}$ is a set of n nodes and $E \subseteq N \times N$ is a set of $m := |E|$ lines. For an undirected graph we refer to its line by $(j,k) \in E$ or $j \sim k \in E$. We assume there are no self-loops, i.e., $(j,j) \notin E$ for any $j \in N$. We sometimes endow the graph with an arbitrary orientation in which case we refer to a line in E by (j,k), $j \sim k$, or $j \to k$ interchangeably. With respect to this graph orientation, let $C \in \{-1, 0, 1\}^{n \times m}$ denote the node-by-line *incidence matrix* defined in (4.11) and reproduced here:

$$C_{jl} = \begin{cases} 1 & \text{if } l = j \to k \text{ for some bus } k \\ -1 & \text{if } l = i \to j \text{ for some bus } i \\ 0 & \text{otherwise} \end{cases}$$

Unless otherwise specified we usually assume G is connected.

Associated with each line $l := (j,k) \in E$ is a parameter b_l and let $B := \operatorname{diag}(b_l, l \in E)$. A key property we assume is that $b_l > 0$ for all $l \in E$, so B is positive definite and invertible. The *Laplacian matrix* L associated with G is defined to be

$$L := CBC^\mathsf{T} \tag{4.51a}$$

Since the Laplacian matrix L is symmetric it is often simpler to treat G as an undirected graph when working with L. The entries of L are given by (Exercise 4.17):

$$L_{jk} := \begin{cases} -b_{jk} & (j,k) \in E \\ \sum_{i:i\sim j} b_{ij} & j = k \\ 0 & \text{otherwise} \end{cases} \tag{4.51b}$$

The defining properties of the Laplacian matrix L are:

- It is real symmetric. For notational convenience we define, for each $(j,k) \in E$, both b_{jk} and b_{kj} with $b_{jk} = b_{kj}$.
- All row sums, and column sums, are zero.
- $b_l > 0$ for all $l \in E$.

For the DC power flow model studied in Chapter 4.6.2, row/column sums are zero because the shunt admittances $(\tilde{y}^m_{jk}, \tilde{y}^m_{kj})$ are assumed zero, and $b_{jk} > 0$ becasuse $b_{jk} := -\tilde{b}^s_{jk}|V_j||V_k|$ where $\tilde{b}^s_{jk} < 0$ are the series line susceptances and $|V_j|$ are given voltage magnitudes.

This leads to the following important property from which many other properties of L follow.

Lemma 4.12. For all $x \in \mathbb{R}^n$ we have $x^{\mathsf{T}}Lx = \sum_{(j,k)\in E} b_{jk}(x_j - x_k)^2 \geq 0$.

Proof We have from (4.51)

$$x^{\mathsf{T}}Lx = \sum_j \sum_k L_{jk}x_j x_k = \sum_j x_j \left(\sum_{i:i\sim j} b_{ij}x_j + \sum_{k:j\sim k} -b_{jk}x_k \right) = \sum_{(i,j)\in E} b_{ij}\left(x_i^2 - 2x_i x_j + x_j^2\right)$$
$$= \sum_{(i,j)\in E} b_{ij}\,(x_i - x_j)^2$$

where the third equality follows because we have defined both $b_{jk} = b_{kj}$ for each $(j,k) \in E$. □

An immediate consequence of the lemma is a set of useful properties in Theorem 4.13. Before presenting them we review the concept of pseudo-inverse (see Appendix A.7 for more details).

Spectral decomposition and pseudo-inverse.

An arbitrary complex matrix $A \in \mathbb{C}^{n\times n}$ has a singular value decomposition

$$A = V\Sigma W^{\mathsf{H}}$$

where $\Sigma = \mathrm{diag}(\sigma_j, j = 1,\ldots,n)$ is a diagonal matrix of singular values $\sigma_j \geq 0$, and V and W are unitary matrices whose columns are orthonormal sets of eigenvectors of AA^{H} and $A^{\mathsf{H}}A$ respectively (Theorem A.11 in Appendix A.6.1). The pseudo-inverse of A is defined to be

$$A^{\dagger} := W\Sigma^{\dagger}V^{\mathsf{H}}$$

where $\Sigma^{\dagger}$ is a diagonal matrix obtained by replacing the positive σ_j by $1/\sigma_j$ in Σ. The main properties of pseudo-inverse are summarized in Theorem A.19 and Corollary A.20 in Appendix A.7.

If $A \in \mathbb{C}^{n\times n}$ is a normal matrix then it has a spectral decomposition

$$A \;=\; U\Lambda U^{\mathsf{H}} \;=\; \sum_j \lambda_j u_j u_j^{\mathsf{H}}$$

where $\lambda_i \in \mathbb{C}$ are complex eigenvalues of A and the columns $(u_j, j = 1,\ldots,n)$ of the unitary matrix U are an orthonormal basis of $\mathbb{C}^n$ (Theorem A.15 of Appendix A.6.2). If $A \in \mathbb{C}^{n\times n}$ is positive semidefinite (necessarily Hermitian), then the eigenvalues $\lambda_j \geq 0$ are real and nonnegative. Moreover Theorem A.16 shows that the singular value decomposition coincides with the spectral decomposition of A, i.e., $A = V\Sigma W^{\mathsf{H}} = U\Lambda U^{\mathsf{H}}$ and $\sigma_j = \lambda_j \geq 0$. If $A \in \mathbb{R}^{n\times n}$ is a real positive semidefinite matrix (necessarily symmetric by definition), then U can be taken as a real and orthogonal matrix. In this case

$$A^\dagger \;=\; U\Lambda^\dagger U^{\mathsf{T}} \;=\; \sum_{j:\lambda_j>0} \frac{1}{\lambda_j} u_j u_j^{\mathsf{T}}$$

where $\Lambda^\dagger$ is a diagonal matrix obtained by replacing the positive λ_j by $1/\lambda_j$ in Λ. Let rank $A = n - k$ and

$$0 = \lambda_1 = \cdots = \lambda_k < \lambda_{k+1} \leq \cdots \leq \lambda_n$$

Then

$$A \;=\; U\Lambda U^{\mathsf{T}} \;=\; \sum_{j>k} \lambda_j u_j u_j^{\mathsf{T}}, \qquad A^\dagger \;=\; U\Lambda^\dagger U^{\mathsf{T}} \;=\; \sum_{j>k} \frac{1}{\lambda_j} u_j u_j^{\mathsf{T}} \tag{4.52}$$

Theorem 4.13 (Laplacian matrix L)**.** Suppose the graph $G = (N, E)$ consists of $K \geq 1$ connected components. Consider its Laplacian matrix L defined in (4.51).

1 L is positive semideifinite.
2 L is of rank $n - K$ with the null space of L spanned by vectors that have $x_j = x_k$ for all buses j, k in the same connected component. In particular if G is connected ($K = 1$) then L is of rank $n - 1$ with span($\mathbf{1}$) as its null space.
3 Suppose the graph G is connected, i.e., $K = 1$. Then
   - The pseudo-inverse $L^\dagger$ of L is given by

$$L^\dagger = \left(L + \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathsf{T}}\right)^{-1} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathsf{T}} \;=\; \sum_{j=2}^{N} \frac{1}{\lambda_j} v_j v_j^{\mathsf{T}} \tag{4.53}$$

     where $0 = \lambda_1 < \lambda_2 \leq \cdots \leq \lambda_n$ are the eigenvalues of L and v_j are the corresponding eigenvectors.
   - Both L and $L^\dagger$ are symmetric and have zero row (and hence column) sums.
   - We have

$$LL^\dagger \;=\; L^\dagger L = \mathbb{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathsf{T}}$$

     where $\mathbb{I}_n$ is the identity matrix of size n. Hence for all $x \in \mathbb{R}^n$ with $\mathbf{1}^{\mathsf{T}}x = 0$, we have $L^\dagger Lx = x$ and $LL^\dagger x = x$.

4 Suppose the graph G is connected, i.e., $K = 1$. Then
- Any $k \times k$ principal submatrix M of L is positive definite and hence invertible, $k \le n-1$.
- Moreover both M and its inverse M^{-1} are symmetric.

Proof 1 Lemma 4.12 implies that L is positive semidefinite since $b_l > 0$ for all $l \in E$.

2 First we claim that v is in the null space of L if and only if $v^\mathsf{T} L v = 0$. To see the sufficiency, we have from (4.52) that $v^\mathsf{T} L v = \sum_j \lambda_j (u_j^\mathsf{T} v)^2$. Hence $v^\mathsf{T} L v = 0$ implies that $u_j^\mathsf{T} v = 0$ for all j such that $\lambda_j > 0$, i.e., $v \in \text{null}(L)$ since $(u_j, \forall j)$ forms a basis of $\mathbb{R}^n$. Suppose $v \in \text{null}(L)$. Lemma 4.12 then implies that $v_i = v_j$ for all buses i, j in the same connected component. If $N_k \subseteq N$, $k = 1, \ldots, K$, are connected components of the graph G, then an orthonormal basis of the null space consists of K orthogonal vectors v^k whose entries are:

$$v_i^k := \frac{1(i \in N_k)}{\sqrt{|N_k|}}, \qquad i = 1, \ldots, n, \;\; k = 1, \ldots, K$$

where $1(\cdot)$ is the indicator function. Hence the null space of L has a dimension of K. Since $\dim(\text{null}(L)) + \text{rank}(L) = n$, $\text{rank}(L) = n - K$.

3 Suppose now $K = 1$. By definition, L is symmetric and has zero row sums. That $L^\dagger = \sum_{j \ge 2} (1/\lambda_j) v_j v_j^\mathsf{T}$ follows directly from (4.52). The formula (4.53) for $L^\dagger$ is proved in Exercise 4.18. The formula implies that $L^\dagger$ is also symmetric. Its row sum is

$$L^\dagger \mathbf{1} = \left(\left(L + \frac{1}{n} \mathbf{1}\mathbf{1}^\mathsf{T} \right)^{-1} - \frac{1}{n} \mathbf{1}\mathbf{1}^\mathsf{T} \right) \mathbf{1} \;=\; \left(L + \frac{1}{n} \mathbf{1}\mathbf{1}^\mathsf{T} \right)^{-1} \mathbf{1} \;-\; \mathbf{1}$$

To show that this is zero multiply both sides by $L + \frac{1}{n}\mathbf{1}\mathbf{1}^\mathsf{T}$ to get:

$$\left(L + \frac{1}{n} \mathbf{1}\mathbf{1}^\mathsf{T} \right) L^\dagger \mathbf{1} = \mathbf{1} \;-\; \left(L + \frac{1}{n} \mathbf{1}\mathbf{1}^\mathsf{T} \right) \mathbf{1} \;=\; \mathbf{1} \;-\; \mathbf{1} \;=\; 0$$

Since $L + \frac{1}{n}\mathbf{1}\mathbf{1}^\mathsf{T}$ is nonsingular, $L^\dagger \mathbf{1}$ must be a zero vector, i.e., row sums of $L^\dagger$ are all zero.

Finally, since v_j are orthonormal eigenvectors of L, we have from (4.53)

$$L L^\dagger \;=\; L \sum_{j \ge 2} \frac{1}{\lambda_j} v_j v_j^\mathsf{T} \;=\; \sum_{j \ge 2} v_j v_j^\mathsf{T} \;=\; \mathbb{I}_n - \frac{1}{n} \mathbf{1}\mathbf{1}^\mathsf{T}$$

where the last equality follows because $\sum_{j \ge 1} v_j v_j^\mathsf{T} = \mathbb{I}_n$ and $v_1 = \mathbf{1}/\sqrt{n}$. Similarly

$$L^\dagger L \;=\; \left(\sum_{j \ge 2} \frac{1}{\lambda_j} v_j v_j^\mathsf{T} \right) L \;=\; \sum_{j \ge 2} v_j v_j^\mathsf{T} \;=\; \mathbb{I}_n - \frac{1}{n} \mathbf{1}\mathbf{1}^\mathsf{T}$$

4 Consider a $k \times k$ principal submatrix M of L with $k \le n-1$. Without loss of

generality we assume M consists of the first k rows and columns of L. As in Lemma 4.12 we have for any nonzero $x \in \mathbb{R}^k$

$$\begin{aligned} x^{\mathsf{T}} M x &= \sum_{i=1}^{k}\sum_{j=1}^{k} L_{ij}x_i x_j = \sum_{i=1}^{k} L_{ii}x_i^2 + \sum_{i\le k}\sum_{\substack{j\le k\\ i\neq j}} L_{ij}x_i x_j \\ &= \sum_{i\le k}\left(\sum_{\substack{j'\le k\\ i\sim j'}} b_{ij'} + \sum_{\substack{j'>k\\ i\sim j'}} b_{ij'}\right)x_i^2 + \sum_{i\le k}\sum_{\substack{j\le k\\ i\sim j}} -b_{ij}x_i x_j \\ &= \sum_{\substack{(i,j)\in E\\ i,j\le k}} b_{ij}\left(x_i^2 - 2x_i x_j + x_j^2\right) + \sum_{i\le k}\sum_{\substack{j'>k\\ i\sim j'}} b_{ij'}x_i^2 \\ &= \sum_{\substack{(i,j)\in E\\ i,j\le k}} b_{ij}\left(x_i - x_j\right)^2 + \sum_{i\le k}\sum_{\substack{j'>k\\ i\sim j'}} b_{ij'}x_i^2 > 0 \end{aligned} \tag{4.54}$$

where the second to last equality follows because $b_{ij} = b_{ji}$ and the inequality follows because G is connected, $k < n$, and $x \neq 0$. Hence M is positive definite and hence invertible.

Since L is symmetric, so is the $k \times k$ principal submatrix M. The inverse of any symmetric nonsingular matrix is symmetric. To see this, first note that if M is a nonsingular square matrix and $M\hat{M} = I$, then $\hat{M}$ is unique because the jth column $\hat{M}_j$ of $\hat{M}$ is uniquely determined by $M\hat{M}_j = e_j$. Since the inverse of M satisfies $M\hat{M} = I$, $\hat{M}$ must be the inverse. If M is symmetric then $M\hat{M}^{\mathsf{T}} = (\hat{M}M^{\mathsf{T}})^{\mathsf{T}} = (\hat{M}M)^{\mathsf{T}} = I$ where the last equality follows because $\hat{M}$ is an inverse of M. This means that $\hat{M}^{\mathsf{T}}$ is also an inverse of M and hence $\hat{M}^{\mathsf{T}} = \hat{M}$, i.e., the inverse of M is symmetric.

□

Hence a strict principal submatrix M of L is always positive definite and invertible, but it is not necessarily strictly diagonally dominant (only diagonally dominant) even though $b_{jk} > 0$ for all $(j,k) \in E$ because strict diagonal dominance requires $\sum_{j\neq i} |M_{ij}| < |M_{ii}|$ for *all* rows i. The theorem is illustrated in Exercise 4.19.

Remark 4.9 (Comparison with complex symmetric admittance matrix)**.** To summarize:

1 For a complex symmetric admittance matrix Y, a strict principal submatrix Y_{22} is not always nonsingular. Theorems 4.5 and 4.6 provide sufficient conditions ($\mathrm{Re}(Y_22) \succ 0$ or $\mathrm{Im}(Y_{22}) \prec 0$) for a strict principal submatrix Y_{22} to be nonsingular.
2 For a complex symmetric admittance matrix Y for a connected radial network, a principal submatrix $\hat{Y}$ corresponding to removing any *leaf node* is always nonsingular and $\hat{Y}^{-1}$ has a simple structure, according to Theorem 4.10. By induction, this holds for any strict principal submatrix Y_{22} if the reduced network graph remains a (connected) tree.

3 For a real symmetric Laplacian matrix L with zero row and column sums, any strict principal submatrix M is nonsingular, according to Theorem 4.13. This is because all off-diagonal entries $L_{jk} = -b_{jk}$, $j \neq k$, are nonzero and of the same sign, resulting in a positive definite M (when $b_{jk} > 0$). Otherwise, it is possible for a real symmetric matrix Y with zero row sums whose off-diagonal entries Y_{jk} may be of different signs to have a rank strictly less than $n-1$ (see Exercise 4.2).

Indeed one can interpret Corollary 4.8 as an extension of the result here to a complex symmetric admittance matrix. Corollary 4.8 shows that, for a complex symmetric admittance matrix Y with zero row and column sums, if $g^s_{jk} > 0$ for all $(j,k) \in E$ or if $b^s_{jk} < 0$ for all $(j,k) \in E$, then indeed $\mathrm{Re}(Y_{22}) \succ 0$ or $\mathrm{Im}(Y_{22}) \prec 0$ respectively, and therefore Y_{22} is nonsingular. The proof that $\mathrm{Re}(Y_{22}) \succ 0$ or $\mathrm{Im}(Y_{22}) \prec 0$ is essentially the same as that for Theorem 4.13 for a real Laplacian matrix (compare (4.54) and (4.19)). In this sense we can regard the conditions $\mathrm{Re}(Y_{22}) \succ 0$ or $\mathrm{Im}(Y_{22}) \prec 0$ in Theorems 4.5 and 4.6 as the generalization of sign definiteness of off-diagonal entries Y_{jk} for a complex symmetric admittance matrix Y. □

4.6.2 DC power flow model

We again model a power network by a connected graph $G = (\overline{N}, E)$ of $N+1$ nodes and M lines, where $\overline{N} := \{0\} \cup N$, $N := \{1, 2, \ldots, N\}$ and $E \subseteq \overline{N} \times \overline{N}$. Each line $(j,k) \in E$ is characterized by series admittance and shunt admittances $(\tilde{y}^s_{jk}, \tilde{y}^m_{jk})$ and $(\tilde{y}^s_{kj}, \tilde{y}^m_{kj})$. In this section we assume $\tilde{y}^s_{jk} = \tilde{y}^s_{kj}$ (assumption C4.1) and $\tilde{y}^m_{jk} = \tilde{y}^m_{kj} = 0$. A popular linearized model, called the *DC power flow model*, makes the following additional assumptions:

- Line losses are negligible, i.e., the series conductances $\tilde{g}^s_{jk} \approx 0$, so $\tilde{y}^s_{jk} \approx \mathbf{i}\tilde{b}^s_{jk}$. The series susceptances $\tilde{b}^s_{jk} < 0$.
- Voltage angle differences are small across each line, i.e., $\sin(\theta_j - \theta_k) \approx \theta_j - \theta_k$ for all lines $(j,k) \in E$.
- Voltage magnitudes $|V_j|$ are given and fixed for all buses $j \in \overline{N}$.
- Ignore reactive power, so the variables in the DC power flow model are $(p_j, \theta_j,\ j \in \overline{N})$.

The DC power flow model is widely used in the industry, e.g., in economic dispatch of generators. The assumptions are reasonable for many problems in transmission networks where the voltage magnitudes are high and real power losses are small. The last two assumption in the model are justified because on transmission networks where loss is low, there is decoupling between voltage angle θ_j and reactive power q_k and between voltage magnitude $|V_j|$ and real power p_k; see Chapter 4.4.3. Hence it is implicitly assumed that reactive power injections q_k can be chosen to stabilize the voltage magnitudes $|V_j|$ separately from the determination of $(p_j, \theta_j,\ j \in \overline{N})$. These

assumptions are not suitable for distribution systems where voltages are much lower, the ratio of line resistance to reactance is high, and reactive power is often used to stabilize voltages. The linear branch flow model of Chapter 5.4 is more suitable for distribution systems.

Under these assumptions, the DC power flow model is defined by (substituting $\tilde{g}_{jk} = 0$, $\tilde{y}^m_{jk} = \tilde{y}^m_{kj} = 0$ and replace $\sin\theta_{jk}$ with $\theta_j - \theta_k$ in (4.27a)):

$$p_j = \sum_{k:j\sim k} (-\tilde{b}^s_{jk}|V_j||V_k|)(\theta_j - \theta_k) \;=:\; \sum_{k:j\sim k} b_l(\theta_j - \theta_k) \qquad j \in \overline{N} \quad (4.55a)$$

where $b_l := -\tilde{b}^s_{jk}|V_j||V_k| > 0$ where $|V_j|, |V_k|$ are given voltage magnitudes. Clearly $\sum_j p_j = \sum_j \sum_k b_l(\theta_j - \theta_k) = 0$. This is a consequence of the lossless assumption $\tilde{g}^s_{jk} = 0$ and $\tilde{y}^m_{jk} = \tilde{y}^m_{kj} = 0$.[7]

We can write the DC model (4.55a) in vector form, as follows. Let $B = \text{diag}(b_l, l \in E) \succ 0$ be the (weighted) susceptance matrix. Let $p := (p_j, j \in \overline{N})$ be the power injections at buses in $\overline{N}$. Let $\theta := (\theta_j, j \in \overline{N})$ be the voltage phase angles at these buses. Let $P := (P_l, l \in E)$ be the real power flows on line l. The DC power flow model is specified by the following equations in (p, P, θ):

$$p \;=\; CP, \qquad P \;=\; BC^{\mathsf{T}}\theta \qquad (4.55b)$$

Eliminate P to relate voltage angles θ directly to injections p:

$$p = CBC^{\mathsf{T}}\theta \;=:\; L\theta$$

where the $(N+1)\times(N+1)$ matrix $L := CBC^{\mathsf{T}}$ is the Laplacian matrix of the graph G. This is (4.55a). When G is connected, L has rank N and the null space is span($\mathbf{1}$) (Theorem 4.13). Hence, given an injection vector p that is orthogonal to span($\mathbf{1}$), i.e., power is balanced over the network $\mathbf{1}^{\mathsf{T}}p = \sum_{j\in\overline{N}} p_j = 0$, the DC power flow equation (4.55b) has a subspace of solutions (P, θ) given by:

$$P \;=\; BC^{\mathsf{T}}L^{\dagger}p, \qquad \theta \;=\; L^{\dagger}p \;+\; a\mathbf{1}, \quad a \in \mathbb{R} \qquad (4.55c)$$

For example we can choose a so that $\theta_0 = 0$ at bus 0. It is important that the line flows P are unique regardless of the choice of θ because $C^{\mathsf{T}}\mathbf{1} = 0$. The models (4.55a), (4.55b) and (4.55c) are equivalent models.

There is yet another way to specify the DC power flow model. Let $\hat{C}$ denote the $N \times M$ reduced incidence matrix obtained from C by removing the row corresponding to the reference bus 0. Let $\hat{L} := \hat{C}B\hat{C}^{\mathsf{T}}$ be the reduced Laplacian matrix. Hence $\hat{L}$ can be obtained from L by removing its row and column corresponding to bus 0. Then $\hat{L}$ is of rank N and invertible according to Theorem 4.13. Let $\hat{p} := (p_j, j \in N)$ and $\hat{\theta} := (\theta_j, j \in N)$ be the power injections and voltage angels at non-reference buses.

[7] For the special case of the flat voltage profile $V_j = V^{\text{flat}}$ for all $j \in \overline{N}$ where V^{flat} is a common nominal voltage, e.g., $V^{\text{flat}} = 1\angle 0°$, (4.55a) is also the linearization of the polar form power flow equation (4.27a) around the flat voltage profile and the resulting injections $(p^{\text{flat}}, q^{\text{flat}}) = (0, 0)$; see Exercise 7.8.

Then, given any $\hat{p}$, the solution of (4.55b) can also be expressed in terms of $\hat{L}^{-1}$ and $(\hat{p}, \hat{\theta})$ at non-reference buses as:

$$P = B\hat{C}^{\mathsf{T}}\hat{L}^{-1}\hat{p}, \qquad \hat{\theta} = \hat{L}^{-1}\hat{p} \tag{4.55d}$$

This solution is unique and assumes that bus 0 is the angle reference bus, i.e., $\theta_0 := 0$. It is a special case of the solution (4.55c) in terms of the pseudo-inverse $L^{\dagger}$ with a chosen so that $\theta_0 = 0$. The solution (4.55c) is therefore more flexible since it works for any reference bus whereas $\hat{L}$ in (4.55d) generally changes when a different bus is chosen as a reference. We will mostly use $L^{\dagger}$ in our analysis. The next result formally states this relation; in particular, it shows that the line flow P is independent of the choice of the angle reference bus or $\hat{L}$.

Lemma 4.14. Consider the DC power flow model (4.55). For any injections p with $\mathbf{1}^{\mathsf{T}}p = 0$ we have

$$P = B\hat{C}^{\mathsf{T}}\hat{L}^{-1}\hat{p} = BC^{\mathsf{T}}L^{\dagger}p, \qquad \hat{\theta} = \hat{L}^{-1}\hat{p} \tag{4.56}$$

when $\theta_0 := 0$. This implies $C^{\mathsf{T}}L^{\dagger}p = \hat{C}^{\mathsf{T}}\hat{L}^{-1}\hat{p}$ and $C^{\mathsf{T}}L^{\dagger}C = \hat{C}^{\mathsf{T}}\hat{L}^{-1}\hat{C}$.

Proof Write

$$C = \begin{bmatrix} c_0^{\mathsf{T}} \\ \hat{C} \end{bmatrix}, \quad p = \begin{bmatrix} p_0 \\ \hat{p} \end{bmatrix}, \quad \theta = \begin{bmatrix} \theta_0 \\ \hat{\theta} \end{bmatrix}$$

where c_0^{T} is the first row of C corresponding to bus 0. Then

$$L = \begin{bmatrix} c_0^{\mathsf{T}}Bc_0 & c_0^{\mathsf{T}}B\hat{C}^{\mathsf{T}} \\ \hat{C}Bc_0 & \hat{L} \end{bmatrix}$$

with $\hat{L} = \hat{C}B\hat{C}^{\mathsf{T}}$ and the power flow equations (4.55b) become:

$$p_0 = c_0Bc_0^{\mathsf{T}}\theta_0 + c_0BC^{\mathsf{T}}\hat{\theta}, \quad \hat{p} = \hat{C}Bc_0^{\mathsf{T}}\theta_0 + \hat{L}\hat{\theta} \tag{4.57a}$$

$$P = B\begin{bmatrix} c_0^{\mathsf{T}} & \hat{C}^{\mathsf{T}} \end{bmatrix}\begin{bmatrix} \theta_0 \\ \hat{\theta} \end{bmatrix} = Bc_0^{\mathsf{T}}\theta_0 + B\hat{C}^{\mathsf{T}}\hat{\theta} \tag{4.57b}$$

The power flow solution (4.56) corresponds to choosing a in (4.55c) so that $\theta_0 = 0$ ($P = BC^{\mathsf{T}}L^{\dagger}p$ is independent of the choice of a because $L^{\dagger}\mathbf{1} = 0$). Hence (4.57) implies $P = BC^{\mathsf{T}}L^{\dagger}p = B\hat{C}^{\mathsf{T}}\hat{L}^{-1}p$.

Finally equating P in (4.55) and (4.57) gives $B^{-1}P = C^{\mathsf{T}}L^{\dagger}p = \hat{C}^{\mathsf{T}}\hat{L}^{-1}\hat{p}$ for any p with $\mathbf{1}^{\mathsf{T}}p = 0$. Substituting $p := C_j$ and $\hat{p} := \hat{C}_j$ to be the jth columns of C and $\hat{C}$ respectively (which satisfies $\mathbf{1}^{\mathsf{T}}p = 0$), we have $C^{\mathsf{T}}L^{\dagger}C_j = \hat{C}^{\mathsf{T}}\hat{L}^{-1}\hat{C}_j$. Since this holds for all j we have $C^{\mathsf{T}}L^{\dagger}C = \hat{C}^{\mathsf{T}}\hat{L}^{-1}\hat{C}$. This completes the proof. □

The quantities in the lemma are illustrated in Exercise 4.20. The lemma is generalized in Chapter 6.4.3.4 to the case where there can be a reference bus for angle and a different reference (slack) bus for pricing electricity (both are taken to be bus 0 here). It is shown in Theorem 6.3 that the line flows P, and the optimal dispatch and LMP (p^*, λ^*) are independent of the choices of reference buses.

Remark 4.10 (Loop flow and uniqueness of P)**.** We call a line flow vector P a *loop flow* if it satisfies power balance with zero injections, i.e., $CP = 0$. Hence P_σ is a loop flow if and only if it is in the null space of C. Given any balanced injection vector p with $\sum_j p_j = 0$, the line flows P that satisfy $p = CP$ are not unique. If P satisfies $p = CP$, so does $P + P_\sigma$ for any loop flow P_σ. The DC power flow model (4.55b) requires both $p = CP$ and $P = BC^\mathsf{T}\theta$. The second equation ensures that loop flow $P_\sigma = 0$ and the line flows P in a DC power flow solution are unique. To see this, suppose both (P, θ) and $(P + P_\sigma, \tilde{\theta})$ are power flow solutions, i.e., they satisfy

$$\begin{aligned} p &= CP, & P &= BC^\mathsf{T}\theta \\ p &= C(P + P_\sigma), & P + P_\sigma &= BC^\mathsf{T}\tilde{\theta} \end{aligned}$$

This implies $CP_\sigma = 0$ and $B^{-1}P_\sigma = C^\mathsf{T}(\tilde{\theta} - \theta)$ and hence P_σ and $B^{-1}P_\sigma$ are in orthogonal subspaces, i.e., $P_\sigma^\mathsf{T}\left(B^{-1}P_\sigma\right) = 0$ yielding $P_\sigma = 0$ since B is positive definite. □

Power loss.

The DC power flow model assumes zero real power loss. It is possible to augment the basic equation (4.55) by adding a loss term, as the next example shows.

Example 4.10 (Loss in linear mdoel)**.** Suppose $\tilde{y}^s_{jk} = \tilde{y}^s_{kj}$ for all lines $(j,k) \in E$ (assumption C4.1) and $\tilde{y}^m_{jk} = \tilde{y}^m_{kj} = 0$. Write $V_j := |V_j| e^{\mathbf{i}\theta_j}$ and $\tilde{y}^s_{jk} =: \tilde{g}^s_{jk} + \mathbf{i}\tilde{b}^s_{jk}$. Then the total real power loss over a network is given by (Exercise 4.12):

$$c(\theta) := \sum_{j \in \overline{N}} p_j = \sum_{j \to k \in E} \tilde{g}^s_{jk} \left|V_j - V_k\right|^2 = \sum_{j \to k \in E} \tilde{g}^s_{jk} \left(|V_j|^2 + |V_k|^2 - 2|V_j||V_k| \cos\theta_{jk}\right)$$

where $\theta_{jk} := \theta_j - \theta_k$. As in the DC power flow model (4.55) we assume here voltage magnitudes $|V_j|$ are fixed and the total loss c is a function of the voltage angles θ.

Recall the flat voltage profile where $V_j^{\text{flat}} = \mu e^{\mathbf{i}\theta^{\text{flat}}}$ for all $j \in \overline{N}$, so that the resulting power injection is $\left(p^{\text{flat}}, q^{\text{flat}}\right) = (0,0)$. To compute the Taylor expansion of $c(\theta)$ around the flat voltage profile we have:

$$\begin{aligned} c(\theta^{\text{flat}}) &= 0 \\ \frac{\partial c}{\partial \theta_i}(\theta^{\text{flat}}) &= \sum_{i \to k \in E} 2\mu^2 \tilde{g}^s_{ik} \sin\theta^{\text{flat}}_{ik} + \sum_{j \to i \in E} -2\mu^2 \tilde{g}^s_{ji} \sin\theta^{\text{flat}}_{ji} = 0 \\ \frac{\partial^2 c}{\partial \theta_i \partial \theta_j}(\theta^{\text{flat}}) &= \begin{cases} -2\mu^2 \tilde{g}^s_{ij} \cos\theta^{\text{flat}}_{ij} = -2\mu^2 \tilde{g}^s_{ij} & \text{if } i \to j \in E \\ -2\mu^2 \tilde{g}^s_{ji} \cos\theta^{\text{flat}}_{ji} = -2\mu^2 \tilde{g}^s_{ji} & \text{if } j \to i \in E \\ \sum_{k:(i,k) \text{ or } (k,i) \in E} 2\mu^2 \tilde{g}^s_{ik} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Hence the second derivative $\frac{\partial^2 c}{\partial \theta^2}$ is a real symmetric Laplacian matrix with zero row and column sums, and is therefore positive semidefinite. Let $g_l := 2\mu^2 \tilde{g}^s_l$ for $l \in E$ and

$G := \mathrm{diag}\,(g_l, l \in E)$. Define

$$L_{\text{loss}} := \frac{\partial^2 c}{\partial \theta^2}(\theta^{\text{flat}}) = CGC^{\mathsf{T}} \tag{4.58a}$$

where C is the incidence matrix of the network graph. Then a loss term can be taken as the second-order Taylor expansion of $c(\theta)$ around the flat voltage profile (the perturbation variable θ now denotes the deviations from θ^{flat}):

$$\hat{c}(\theta) = c(\theta^{\text{flat}}) + \frac{\partial c}{\partial \theta}(\theta^{\text{flat}})\,\theta + \frac{1}{2}\theta^{\mathsf{T}} L_{\text{loss}}\theta = \frac{1}{2}\theta^{\mathsf{T}} L_{\text{loss}}\theta \tag{4.58b}$$

Since the matrix L_{loss} in (4.58a) is positive semidefinite the loss $\hat{c}(\theta)$ is a convex quadratic function of θ. □

4.6.3 Distribution factors

4.7 Bibliographical notes

The description of LU decomposition to solve $I = YV$ and algorithms to compute power flow solutions are adapted from [1]. For properties of complex symmetric matrices such as the admittance matrix Y, see [15, Chapter 4.4]. For invertibility of Y, the first part of Theorem 4.2 is from [16, Lemma 1] though we have used properties of Schur complement to simplify its proof. See also [17**?**].

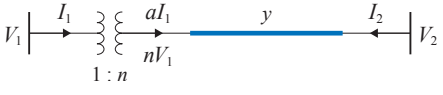
The DC power flow model has been widely used in applications, e.g., for formulating DC OPF [18, 19].

The use of Newton-Raphson algorithm for solving power flow problems is first proposed in [20]. An implementation at BPA is reported in [21] with major improvements, especially a heuristic to optimize the order of Gaussian elimination of the Jacobian matrix in solving $J(x(t))\Delta x(t) = -f(x(t))$. A method is introduced in [22] that computes a new voltage solution $V' = V + \sum_l i_{j_l k_l}(e_{j_l} - e_{k_l})$ to $I = Y'V'$ in terms of the old voltage solution V to $I = YV$ when the admittance matrix changes from Y to Y' (line changes). The quantities $i_{j_l k_l}$ are called compensation currents and are computed from using the old admittance matrix Y. This method, well explained in [23], has the advantage of not having to factorize new matrix Y' into its LU decomposition when relatively few number of lines are changed. The Fast Decoupled algorithm is proposed in [18].

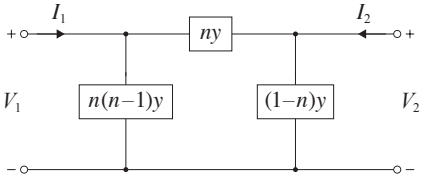
4.8 Problems

Chapter 4.2

Exercise 4.1 (Ideal transformer and transmission line)**.** Consider the cascade in the one-line diagram of Figure 4.13(a) of an *ideal* transformer with voltage gain n and a transmission line modeled by a series admittance y (and zero shunt admittances). Show that its external behavior is equivalent to that of the Π circuit model in Figure



(a) One-line diagram



(b) Equivalent Π circuit model

Figure 4.13 An ideal transformer with turns ratio $a = n^{-1}$ followed by a transmission line modeled by a series admittance y.

4.13(b).

Exercise 4.2 (Real Laplacian matrix)**.** Suppose the $n \times n$ admittance matrix Y of a connected graph is real symmetric with zero row sums (e.g., Y is the admittance matrix of a DC network), i.e., $Y_{jk} = Y_{kj}$ for all $j \neq k$ and $Y_{jj} = -\sum_{k:j\neq k} Y_{jk}$ for all j.

1 If Y_{jk} have the same sign for all $(j,k) \in E$, show that rank $Y = n-1$ and hence Y is not invertible and null(Y) = span$(\mathbf{1})$.
2 If Y_{jk} have the same sign for all $(j,k) \in E$, show that the $(n-1)\times(n-1)$ matrix Y' obtained from Y by removing the jth row and column, for any j, has rank $n-1$ and is hence invertible.
3 If Y_{jk} may have different signs for $(j,k) \in E$, give a counterexample to part 1.

Exercise 4.3 (Unitary diagonalizability of Y)**.** Suppose condition C4.1 holds. Let the bus admittance matrix $Y := G + \mathbf{i}B$ where G and B are real matrices (whose rows may not sum to zero).

1 Show that Y is normal (i.e., $YY^H = Y^H Y$) and hence unitarily diagonalizable if and only if G and B commute, or if and only if BG is symmetric.
2 Suppose all lines have the same RX ratio, i.e., for some real α, $b^s_{jk} = \alpha g^s_{jk}$ for all $(j,k) \in E$ and $b^m_{jj} = \alpha g^m_{jj}$ for all $j \in \overline{N}$ (or all shunt elements are zero). Show that Y is normal. (Hint: Use part 1.)

Exercise 4.4 (Inverse of Y)**.** Consider a complex matrix $A =: G + \mathbf{i}B$ where $G, B \in \mathbb{R}^{n\times n}$. Show that, even if both G and B are singular, its inverse $A^{-1} =: R + \mathbf{i}X$ may exist though not given by the formulae (4.13b) or (4.14b). This is the case even if G and B are symmetric.

Exercise 4.5 (Invertibility of Y)**.** Prove part 2 of Theorem 4.2.

Exercise 4.6 (Invertibility of Y, [16])**.** This is an alternative proof from [16, Lemma 1] of (part of) Theorem 4.2: a complex symmetric matrix Y is nonsingular if $\mathrm{Re}(Y) \succ 0$ or if $\mathrm{Im}(Y) \prec 0$. Prove the claim by showing that there exists no nonzero vector α such that $Y\alpha = 0$.

Exercise 4.7 (Invert Y using matrix inversion lemma)**.** Recall that, under condition C4.1, the admittance matrix Y can be written in terms of the incidence matrix C as (from (4.12)):

$$Y = C\, D^s_y C^\mathsf{T} \;+\; D^m_y$$

where $D^s_y := \mathrm{diag}\left(y^s_l, l \in E\right)$ and $D^m_y := \mathrm{diag}\left(y^m_{jj}, j \in \overline{N}\right)$. Suppose $y^s_l \neq 0$ for all l and $y^m_{jj} \neq 0$ for all j so that the diagonal matrices Y^s and Y^m are invertible.

1 Show that Y is invertible if and only if the $M \times M$ matrix

$$\hat{E} := \left(D^s_y\right)^{-1} + C^\mathsf{T}\left(D^m_y\right)^{-1} C$$

is invertible.
2 If Y is invertible then

$$Y^{-1} = \left(D^m_y\right)^{-1} \;-\; (D^m)^{-1}\left(C\left(\hat{E}\right)^{-1} C^\mathsf{T}\right)\left(D^m_y\right)^{-1}$$

(Hint: For part 1 use the property that a matrix is nonsingular if and only if a principal submatrix and its Schur complement are both nonsingular, according to Theorem A.4 in Appendix A.3. For part 2 use the matrix inversion lemma in Appendix A.3.2.)

Exercise 4.8 (Invertibility of complex symmetric vs psd matrices)**.** Let $A \in \mathbb{C}^{n\times n}$.

1 Prove that A is invertible if $v^H Av \neq 0$ for all nonzero $v \in \mathbb{C}^n$.
2 Show that the converse is not true by providing a counterexample A that is Hermitian (including real symmetric) and a counterexample A that is complex symmetric. (Hint: Consider 2×2 diagonal matrices.)
3 Suppose A is (Hermitian and) positive semidefinite. Then the following are equivalent:
   - A is invertible
   - $v^H Av \neq 0$ for all nonzero $v \in \mathbb{C}^n$.
   - A is positive definite.
4 Why Lemma 4.12 applies to real Laplacian matrices but not complex Laplacian matrices?

Exercise 4.9 (Alternative proof of Theorem 4.3)**.** Consider the complex symmetric admittance matrix $Y \in \mathbb{C}^{(N+1)\times(N+1)}$. Let λ be an eigenvalue of Y and $\alpha \in \mathbb{C}^{N+1}$ a corresponding eigenvector. Then $\alpha^{\mathsf{H}} Y\alpha = \lambda||\alpha||^2$ where $||\cdot||$ denotes the Euclidean norm. A sufficient (but not necessary) condition for Y to be invertible is that $\alpha^{\mathsf{H}} Y\alpha \neq 0$ for all nonzero vectors $\alpha \in \mathbb{C}^{N+1}$. Let $y^s_{jk} =: g^s_{jk} + \mathbf{i}b^s_{jk}$, $y^m_{jj} =: g^m_{jj} + \mathbf{i}b^m_{jj}$.

1 Suppose condition C4.1 holds. Show that

$$\alpha^{\mathsf{H}} Y\alpha = \left(\sum_{(j,k)\in E} g^s_{jk}\left|\alpha_j - \alpha_k\right|^2 + \sum_{j\in\overline{N}} g^m_{jj}\left|\alpha_j\right|^2\right) + \mathbf{i}\left(\sum_{(j,k)\in E} b^s_{jk}\left|\alpha_j - \alpha_k\right|^2 + \sum_{j\in\overline{N}} b^m_{jj}\left|\alpha_j\right|^2\right)$$

2 Show that the conditions in Theorem 4.3 imply that $\alpha^{\mathsf{H}} Y\alpha > 0$ for all nonzero vectors $\alpha \in \mathbb{C}^{N+1}$.

Exercise 4.10 (Kron reduction)**.** Suppose condition C4.1 holds so that an admittance matrix Y is complex symmetric. Consider its Kron-reduction Y/Y_{22} (assume Y_{22} is invertible):

$$Y =: \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{12}^{\mathsf{T}} & Y_{22} \end{bmatrix}, \qquad Y/Y_{22} := Y_{11} - Y_{12}Y_{22}^{-1}Y_{12}^{\mathsf{T}}$$

1 Show that Y_{22}^{-1} and Y/Y_{22} are symmetric.
2 Show that if Y has zero row (and hence column) sums, i.e., $y^m_{jk} = y^m_{kj} = 0$ for $(j,k) \in E$, so does Y/Y_{22}.
3 Show that the converse does not necessarily hold. (Hint: Consider Example 4.5.)

Exercise 4.11 (Radial Network: inverses of $\hat{C}$ and $\hat{Y}$)**.** Prove Theorem 4.10. (Hint: Let B be the matrix defined in (4.23) and verify directly that $\hat{C}B$ equals the identity matrix. Use part 1 to derive $\hat{Z}_{jk}$.)

Chapter 4.3

Exercise 4.12 (Real power loss)**.** Let $(p_j, j \in \overline{N})$ denote the real nodal power injections. For each line $(j,k) \in E$, let its series and shunt admittances be $y^s_{jk} = g^s_{jk} + \mathbf{i}b^s_{jk}$ and $y^m_{jk} = g^m_{jk} + \mathbf{i}b^m_{jk}$, and similarly for (y^s_{kj}, y^m_{kj}). Define the total real power loss over the network, as a function of V: injection $L_1(V) := \sum_j p_j(V)$. Suppose $y^s_{jk} = y^s_{kj}$ for all $(j,k) \in E$ (assumption C4.1).

1 Show that

$$L_1(V) \;=\; \sum_{(j,k)\in E} \left(g^s_{jk} \left|V_j - V_k\right|^2 + g^m_{jk} \left|V_j\right|^2 \;+\; g^m_{kj} \left|V_k\right|^2 \right)$$

If C4.1 does not hold, why will the loss depend also on series subsceptances (b^s_{jk}, b^s_{kj})?

2 A popular concept is the thermal loss on transmission or distribution lines. Define the total thermal loss as:

$$L_2(V) \;:=\; \sum_{(j,k)\in E} r^s_{jk} \left|I_{jk}(V)\right|^2$$

where $z^s_{jk} = r^s_{jk} + \mathbf{i}x^s_{jk} := 1/y^s_{jk}$ and $I_{jk}(V)$ is the sending-end current on line (j,k) from j to k. Show that $L_1(V)$ reduces to $L_2(V)$ when $g^m_{jk} = g^m_{kj} = 0$.

Chapter 4.4

Exercise 4.13 (Gauss algorithm)**.** Consider solving for the roots of

$$g(x) = ax^2 - x \tag{4.59}$$

i.e., finding x such that $g(x) = 0$. An x is a root of g if and only if it is a fixed point of $f(x) := ax^2$, i.e., if and only if $x = f(x)$. The Gauss algorithm computes a fixed point of $f(x)$ by performing the fixed-point iteration:

$$x(t+1) := f(x(t)) \tag{4.60}$$

Let $X \subseteq \mathbb{R}$ be closed and convex and suppose f maps X into X. We say f is a *contraction mapping* on X if there exists an $\alpha \in [0,1)$ such that

$$|f(y) - f(x)| \le \alpha\, |y - x|, \qquad \text{for all } x, y \in X \tag{4.61}$$

If f is a contraction mapping on X then there is a unique fixed point $x^* \in X$ and the fixed-point iteration (4.60) always converges to x^*, starting from any initial point $x(0) \in X$.

1 What are the roots of g in (4.59)?
2 Whenever $|a| < 1$, f maps $X := [-1, 1]$ into X. Show that f is a contraction mapping on X if and only if $|a| < 1/2$. In that case, what is the root of g that (4.60) computes?
3 Show that (4.60) converges to $x^* = 0$ if and only if $x(0)$ satisfies $|ax(0)| < 1$, in which case the convergence is quadratic (i.e., the error ratio $|x(t+1)|/|x^2(t+1)| = |a|$ a constant).
4 Use part 3 to argue that f being a contraction mapping is not necessary for the Gauss algorithm (4.60) to compute a root of g? What is the advantage, if any, if f is indeed a contraction mapping?

Exercise 4.14 (Newton algorithm)**.** The Newton algorithm solves iteratively for $x \in \mathbb{R}^n$ such that $g(x) = 0$ where $g : \mathbb{R}^n \to \mathbb{R}^n$. In each iteration, it approximates g by its linearization at the current iterate $x(t)$ and moves to $x(t+1)$ where the linearization vanishes. Show that if g is linear, $g(x) = Ax + b$ where A is invertible, then the Newton algorithm solves $g(x) = 0$ in one step wherever it starts.

Exercise 4.15 (Kantorovich Theorem)**.** The Newton algorithm converges if the initial point is close to a solution. This is made precise by the Kantorovich Theorem. Consider $g : D \to \mathbb{R}^n$ where $D \subseteq \mathbb{R}^n$ is an open convex set. Suppose g is differentiable on D and ∇g is Lipschitz on D, i.e., there is an L such that

$$\|\nabla g(y) - \nabla g(x)\| \le L\|y - x\|, \qquad \text{for all } x, y \in D$$

where $[\nabla g(x)]_{ij} := \frac{\partial g_i}{\partial x_j}(x)$. Suppose $x_0 \in D$ and that $\nabla g(x_0)$ is invertible. Let

$$\beta \ \ge \ \left\|(\nabla g(x_0))^{-1}\right\|, \ \eta \ \ge \ \left\|(\nabla g(x_0))^{-1}\, g(x_0)\right\|$$

$$h \ := \ \beta\eta L, \ r \ := \ \frac{1 - \sqrt{1 - 2h}}{h}\,\eta$$

The Kantorovich Theorem says that if the closed ball $B_r(x_0) \subseteq D$ and $h \le 1/2$ then the Newton iteration

$$x(t+1) := x(t) \ - \ (\nabla g(x(t)))^{-1}\, g(x(t))$$

converges to a solution x^* of $g(x) = 0$ in the closed ball $B_r(x_0)$.

1 Apply the Kantorovich Theorem to $g(x) := ax^2 - x$ to prove that the Newton

iterates converge to a root of g if the initial point x_0 satisfies either of the following conditions, assuming $a > 0$:

$$x_0 \leq \frac{1}{2a}\left(1 - \frac{1}{\sqrt{2}}\right) \text{ or } x_0 \geq \frac{1}{2a}\left(1 + \frac{1}{\sqrt{2}}\right)$$

Which root will the Newton iteration compute in each case?

2 The Kantorovich Theorem provides only a sufficient condition for convergence of the Newton iterates. Show that, for $g(x) := ax^2 - x$, as long as $x_0 \neq (2a)^{-1} = \min_x g(x)$, the Newton iterates will converge. (Hint: use part 1.)

Exercise 4.16 (Fast decoupled algorithm)**.** 1 Use (4.32) to prove (4.36).

2 Show that if $g^s_{jk} = g^m_{jk} = 0$ and $\sin\theta_{jk} = 0$ for all $(j,k) \in E$ then the Jacobian reduces to the approximating block-diagonal matrix $\hat{J}(\theta, |V|) := \begin{bmatrix} \frac{\partial p}{\partial \theta} & 0 \\ 0 & \frac{\partial q}{\partial |V|} \end{bmatrix}$.

Chapter 4.6

Exercise 4.17 (Laplacian matrix L)**.** Show that the entries of Laplacian matrix $L := CBC^\mathsf{T}$ are given by:

$$L_{ij} := \begin{cases} -b_{ij} & i \sim j \ (i \neq j) \\ \sum_{k \sim i} b_{ik} & i = j \\ 0 & \text{otherwise} \end{cases}$$

Exercise 4.18 (Pseudo-inverse of a psd matrix)**.** Consider an positive semidefinite (and necessarily Hermitian) matrix $A \in \mathbb{C}^{n \times n}$ with rank $n - k$. Let its eigenvalues be

$$0 = \lambda_1 = \cdots = \lambda_k < \lambda_{k+1} \leq \cdots \leq \lambda_n$$

and a set of corresponding orthonormal eigenvectors be $u_1, \ldots, u_n$. Then $A = U\Lambda U^\mathsf{T}$ and $A^\dagger = U\Lambda^\dagger U^\mathsf{T}$ where the columns of U are u_i. Show that

$$A^\dagger = \left(A + \sum_{i \leq k} u_i u_i^\mathsf{T}\right)^{-1} - \sum_{i \leq k} u_i u_i^\mathsf{T} \tag{4.62}$$

(Hint: Use (4.52) to verify the inverse of $A + \sum_{i \leq k} u_i u_i^\mathsf{T}$.)

Exercise 4.19 (Laplacian matrix L)**.** Consider the Laplacian matrix

$$L := \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

Compute its spectral decomposition, $L^\dagger$, $LL^\dagger$ and $L^\dagger L$.

Exercise 4.20 (DC power flow model)**.** Consider the 3-bus network shown in Figure 4.14. Assuming the (weighted) susceptance matrix $B = \mathbb{I}_3$ is the identity matrix.

1 Write down the incidence matrix C and reduced incidence matrix $\hat{C}$ using the graph orientation shown in the figure and bus 0 as the reference bus.
2 Write down the Laplacian matrix L and its pseudo-inverse $L^\dagger$, the reduced Laplacian matrix $\hat{L}$ and its inverse $\hat{L}^{-1}$.
3 Write down the line flows P in terms of the injections p with $\sum_j p_j = 0$, and evaluate P when $p = (2, -1, -1)$.
4 Suppose the injection is changed from $p = (2, -1, -1)$ to $\tilde{p} = (2, 0, -2)$. Calculate the new line flows $\tilde{P}$.

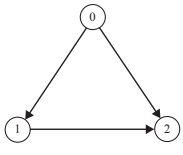


Figure 4.14 Exercise 4.20.

Exercise 4.21 (DC power flow model)**.**

5 Branch flow models: radial networks

In Chapter 5.1 we introduce branch flow models for radial networks with a tree topology. They are useful for modeling distribution systems as most distribution systems are radial. Whereas bus injection models of Chapter 4 consist of only nodal variables (nodal voltages and nodal power or current injections), branch flow models involve also branch power flows and branch currents. In Chapter 5.2 we prove their equivalence by first extending branch flow models to general networks with cycles. Branch flow models are most useful for radial networks where they enjoy two important advantages: a fast iterative algorithm studied in Chapter 5.3, called the backward forward sweep, for power flow computation, and a linearized model studied in Chapter 5.4 that admits an explicit solution and bounds on nonlinear branch powers and voltage magnitudes.

Except in Chapter 5.2 or otherwise specified we will focus in this chapter on radial networks without cycles.

5.1 BFM for radial networks

5.1.1 Line model

We use the same line model as that in Chapter 4.2.2 where a power network with $N+1$ buses and M lines is represented as a connected undirected graph $G = (\overline{N}, E)$ where $\overline{N} := \{0\} \cup N$, $N := \{1, 2, \ldots, N\}$ and $E \subseteq \overline{N} \times \overline{N}$; see Figure 4.7. For each bus $j \in \overline{N}$ let V_j denote its voltage phasor and s_j its complex power injection. For each line $(j,k) \in E$, let (I_{jk}, I_{kj}) denote the *sending-end* line currents from buses j to k and buses k to j respectively. Similarly let (S_{jk}, S_{kj}) denote the *sending-end* line power flows in each direction. Let $V := (V_j, j \in \overline{N})$, $s := (s_j, j \in \overline{N})$, $I := (I_{jk}, I_{kj}, (j,k) \in E)$, and $S := (S_{jk}, S_{kj}, (j,k) \in E)$.

Each line $(j,k) \in E$ is characterized by two pairs of series and shunt admittances, $\left(y^s_{jk}, y^m_{jk}\right) \in \mathbb{C}^2$ from j to k and $\left(y^s_{kj}, y^m_{kj}\right) \in \mathbb{C}^2$ from k to j. It may model a transmission or distribution line, a single-phase transformer, the per-phase model of a three-phase transformer in balanced setting, and may contain admittances of sources and loads. Specifically when (j,k) models a transmission or distribution line, the line parame-

ters $(y^s_{jk} = y^s_{kj}, y^m_{jk}, y^m_{kj})$ are the series and shunt admittances of the transmission or distribution line. When (j,k) models a transformer, the line parameters (y^s_{jk}, y^m_{jk}) and (y^s_{kj}, y^m_{kj}) are given by (4.6) in terms of transformer voltage gain and leakage and shunt admittances $(K(n), \tilde{y}^s_{jk}, \tilde{y}^m_{jk})$. Hence y^s_{kj} and y^s_{jk} may be different, and (y^m_{jk}, y^m_{kj}) are generally different and nonzero even if the transformer shunt admittance $\tilde{y}^m_{jk} = 0$. Let $z^s_{jk} := \left(y^s_{jk}\right)^{-1}$ and $z^s_{kj} := \left(y^s_{kj}\right)^{-1}$.

We will often restrict ourselves to the special case where the series admittances are equal $y^s_{jk} = y^s_{kj}$, and characterize a line by three admittances $\left(y^s_{jk}, y^m_{jk}, y^m_{kj}\right)$. This model can be represented as a Π circuit and behaves like a transmission or distribution line though with generally different y^m_{jk} and y^m_{kj}; see Figure 5.1. It is not suitable as

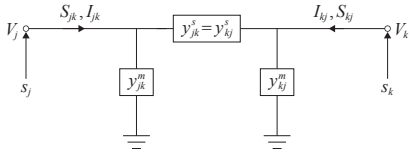


Figure 5.1 Line model under assumption C5.1.

the per-phase model of a balanced three-phase transformer in ΔY or $Y\Delta$ configuration that has a complex voltage gain $K(n)$, but is still widely used as an approximation.

As in Chapter 4.2.2 we label the following assumption and will explicitly state it when it is required:

> C5.1: The series admittances $y^s_{jk} = y^s_{kj}$ or equivalently the series impedances $z^s_{jk} = z^s_{kj}$ for every line $(j,k) \in E$.

In this section we assume the network graph G is a (connected) tree.

5.1.2 With shunt admittances

Transformers are important devices in a distribution system, especially three-phase transformers in ΔY or $Y\Delta$ configuration whose per-phase equivalent circuit does not satisfy assumption C5.1. Their shunt admittances y^m_{jk} and y^m_{kj} may not be negligible even when the transformer shunt admittance $\tilde{y}^m_{jk} = 0$ (see (4.6)). This motivates a branch flow model that includes shunt admittances and allows $y^s_{jk} \neq y^s_{kj}$.

The key feature of a branch flow model for radial networks is that it does not involve phase angles of voltage and current phasors. For each bus j let

- $s_j := (p_j, q_j)$ and $s_j := (p_j + \mathbf{i} q_j)$ represent the real and reactive power injections at bus j. Let $s := (s_j, j \in \overline{N})$.[1]
- v_j represent the squared voltage magnitude at bus j. Let $v := (v_j, j \in \overline{N})$.

For each line (j,k) let

- ℓ_{jk} represent the squared magnitude of the *sending-end* current from bus j to bus k, and ℓ_{kj} represent the squared current magnitude from k to j. Let $\ell := (\ell_{jk}, \ell_{kj}, (j,k) \in E)$.
- $S_{jk} = (P_{jk}, Q_{jk})$ and $S_{jk} = P_{jk} + \mathbf{i} Q_{jk}$ represent the *sending-end* real and reactive branch power flow from bus j to bus k, and S_{kj} represent the sending-end power from k to j. Let $S := (S_{jk}, S_{kj}, (j,k) \in E)$.

We will introduce power flow equations below in terms of the real vector $x := (s, v, \ell, S) \in \mathbb{R}^{3(N+1)+6M}$ that does not involve voltage and current phase angles as variables. The vector v includes v_0 and s includes s_0. The angle information is however embedded in, and can be recovered from, x; see (5.12) below.

Define for each $(j,k) \in E$

$$\alpha_{jk} := 1 + z^s_{jk} y^m_{jk}, \qquad \alpha_{kj} := 1 + z^s_{kj} y^m_{kj}$$

Note that $\alpha_{jk} = \alpha_{kj}$ if and only if $z^s_{jk} y^m_{jk} = z^s_{kj} y^m_{kj}$ and $\alpha_{jk} = \alpha_{kj} = 1$ if and only if $y^m_{jk} = y^m_{kj} = 0$ since $|z^s_{jk}| \neq 0$. A branch flow model for radial networks that allows shunt admittances of lines is:

$$s_j = \sum_{k: j \sim k} S_{jk}, \qquad j \in \overline{N} \tag{5.1a}$$

$$|\alpha_{jk}|^2 v_j - v_k = 2\mathrm{Re}\left(\alpha_{jk} \bar{z}^s_{jk} S_{jk}\right) - |z^s_{jk}|^2 \ell_{jk}, \qquad (j,k) \in E \tag{5.1b}$$

$$|\alpha_{kj}|^2 v_k - v_j = 2\mathrm{Re}\left(\alpha_{kj} \bar{z}^s_{kj} S_{kj}\right) - |z^s_{kj}|^2 \ell_{kj}, \qquad (j,k) \in E \tag{5.1c}$$

$$\left|S_{jk}\right|^2 = v_j \ell_{jk}, \qquad \left|S_{kj}\right|^2 = v_k \ell_{kj}, \qquad (j,k) \in E \tag{5.1d}$$

$$\bar{\alpha}_{jk} v_j - \bar{z}^s_{jk} S_{jk} = \left(\bar{\alpha}_{kj} v_k - \bar{z}^s_{kj} S_{kj}\right)^{\mathsf{H}}, \qquad (j,k) \in E \tag{5.1e}$$

These equations express four properties that a power flow solution $x := (s, v, \ell, S)$ satisfies:

1 *Power balance*: (5.1a) enforces power balance at each bus and is the consequence of KCL.
2 *Ohm's law and KCL*: (5.1b) and (5.1c) originates from the Ohm's law and KCL $I_{jk} = y^s_{jk}(V_j - V_k) + y^m_{jk} V_j$ and similarly for I_{kj} in the opposite direction; see (5.23) in the proof of Theorem 5.2.

[1] We abuse notation and use s to denote both the complex power injection $s = (p + \mathbf{i} q)$ and the real pair $s = (p, q)$, depending on the context. Similarly for $S = (P + \mathbf{i} Q)$ and $S = (P, Q)$, and for $z = (r + \mathbf{i} x)$ and $z = (r, x)$.

3 *Apparent power*: (5.1d) defines the apparent powers and is obtained from $S_{jk} = V_j I_{jk}^{\mathsf{H}}$ and $S_{kj} = V_k I_{kj}^{\mathsf{H}}$.
4 *Cycle condition*: We call (5.1e) a *cycle condition* and it ensures that the line angles implied by a power flow solution x can indeed be realized by nodal voltage angles; see Chapter 5.1.4. It says $V_j V_k^{\mathsf{H}} = \left(V_k V_j^{\mathsf{H}}\right)^{\mathsf{H}}$ where (V_j, V_k) are not part of the model but can be recovered from a power flow solution (see (5.25) in the proof of Theorem 5.2).

The complex notation in (5.1) is only a shorthand for a system of $2(N+1)+6M = 8N+2$ real equations in the vector x of $3(N+1)+6M = 9N+3$ real variables (recall that $M = N$ for a tree). For instance (5.1a) is a shorthand for $p_j = \sum_{k:j\sim k} P_{jk}$ and $q_j = \sum_{k:j\sim k} Q_{jk}$ and (5.1d) is a shorthand for $v_j \ell_{jk} = P_{jk}^2 + Q_{jk}^2$ and $v_k \ell_{kj} = P_{kj}^2 + Q_{kj}^2$. All equations are linear in x except (5.1d) which are quadratic. Given $(2N+1)$ of these variables (e.g., given $v_0 = 1$ and non-slack bus injections (p_j, q_j), $j \in N$), the power flow problem is to determine the remaining $7N+2$ real variables from these equations. There can be zero, one or more than one solutions. In this example there are more (nonlinear) equations than the number of variables, but see Example 5.6 for a linear example where the resulting set of equations is not linearly independent. As mentioned above, this model does not require assumption C5.1 and allows nonzero shunt admittances (y_{jk}^m, y_{kj}^m), and therefore is suitable for modeling transformers as well as distribution lines (see Example 5.1).

Example 5.1 (Two buses connected by a transformer)**.** Consider two buses j and k connected by a transformer characterized by its voltage gain K (possibly complex, e.g., $K = \sqrt{3} n e^{\mathbf{i}\pi/6}$), a series admittance $\tilde{y}^s$ and a shunt admittance $\tilde{y}^m$. The bus injection model of this 2-bus network is given by (4.26a) in complex form. Derive the branch flow model (5.1) in terms of transformer parameters $(K, \tilde{y}^s, \tilde{y}^m)$. (We will show in Chapter 5.2 that the branch flow model and the bus injection model are equivalent.)

Solution. The abstract line parameters in terms of the transformer parameters are given by (4.6) reproduced here:

$$y_{jk}^s := \frac{\tilde{y}^s}{K}, \qquad y_{jk}^m := \left(1 - \frac{1}{K}\right)\tilde{y}^s,$$
$$y_{kj}^s := \frac{\tilde{y}^s}{\bar{K}}, \qquad y_{kj}^m := \frac{1}{|K|^2}\left((1-K)\tilde{y}^s + \tilde{y}^m\right),$$

Define $\tilde{z}^s := (\tilde{y}^s)^{-1}$ and $\tilde{\alpha} := 1 + \tilde{z}^s \tilde{y}^m$. Then

$$z_{jk}^s := (y_{jk}^s)^{-1} = K\tilde{z}^s, \quad z_{kj}^s := (y_{kj}^s)^{-1} = \bar{K}\tilde{z}^s, \quad \alpha_{jk} = K, \quad \alpha_{kj} = \tilde{\alpha}/K$$

For a single line we can substitute $S_{jk} = s_j$ and $S_{kj} = s_k$ and the branch flow model

(5.1) becomes:

$$\begin{aligned} v_j - v_k \left/ |K|^2 \right. &= 2\mathrm{Re}\left((\tilde{z}^s)^{\mathsf{H}} s_j\right) - |\tilde{z}^s|^2 \ell_{jk} \\ |\tilde{\alpha}/K|^2 v_k - v_j &= 2\mathrm{Re}\left(\tilde{\alpha}\,(\tilde{z}^s)^{\mathsf{H}} s_k\right) - |K\tilde{z}^s|^2 \ell_{kj} \\ |s_j|^2 &= v_j \ell_{jk}, \qquad |s_k|^2 = v_k \ell_{kj} \\ v_j - (\tilde{z}^s)^{\mathsf{H}} s_j &= \left(\tilde{\alpha}/|K|^2\right) v_k - \tilde{z}^s \bar{s}_k \end{aligned}$$

This is a system of 6 real (nonlinear) equations in 8 real variables $(s_j, s_k, v_j, v_k, \ell_{jk}, \ell_{kj})$. □

5.1.3 Without shunt admittances

Consider a radial network where lines have zero shunt admittances and hence $\alpha_{jk} = \alpha_{kj} = 1$. Moreover we suppose assumption C5.1 holds. This is a reasonable model if (j,k) models a (short) transmission line or a distribution line. It may be unsuitable if (j,k) models a transformer because, as noted above, the shunt admittances (y^m_{jk}, y^m_{kj}) corresponding to a single-phase nonideal transformer are generally nonzero (see Example 5.1).

A consequence of substituting $z^s_{jk} = z^s_{kj}$ and $y^m_{jk} = y^m_{kj} = 0$ into (5.1) for all lines $(j,k) \in E$ is the relation between the sending-end power flows S_{jk} and S_{kj} (see Exercise 5.3):

$$S_{jk} + S_{kj} = z^s_{jk} \ell_{jk}, \qquad \ell_{jk} = \ell_{kj} \tag{5.2}$$

It says that the sum of sending-end power flows is equal to the complex line loss across the series impedance z^s_{jk}. Hence $-S_{kj} = S_{jk} - z^s_{jk}\ell_{jk}$ is the receiving-end power from j to k. For each line $(j,k) \in E$, we can use (5.2) to eliminate from (5.1) the branch variables (ℓ_{kj}, S_{kj}) in the direction k to j. This leads to a simpler set of equations based on a directed, rather than undirected, graph G, as we now explain. In particular the linear cycle condition (5.1e) becomes vacuous.

In this subsection we assume $G = (\overline{N}, E)$ is directed. We denote a line in E from bus j to bus k either by $(j,k) \in E$ or $j \to k \in E$. Associated with each line $j \to k \in E$ are branch variables (ℓ_{jk}, S_{jk}). It is important to remember that, unlike models in the previous sections, (ℓ_{kj}, S_{kj}) in the opposite direction are not defined in the models in this subsection, unless otherwise specified. Let $(s,v) := (s_j, v_j, j \in \overline{N})$ and $(\ell, S) := (\ell_{jk}, S_{jk}, j \to k \in E)$. In particular the vector v includes v_0 and s includes s_0. Let $x := (s, v, \ell, S)$ in $\mathbb{R}^{3(N+1+M)}$ with $M = N$ since G is a tree. To simplify notation we sometimes omit the superscript on z^s_{jk} and write $z_{jk} = (r_{jk}, x_{jk}) = \left(y^s_{jk}\right)^{-1}$ as the series impedance of line (j,k). Then the branch flow model (5.1) reduces to what is

called the DistFlow equations as follows:

$$\sum_{k:j\to k} S_{jk} = \sum_{i:i\to j} \left(S_{ij} - z_{ij}^s \ell_{ij}\right) + s_j, \qquad j \in \overline{N} \tag{5.3a}$$

$$v_j - v_k = 2\,\mathrm{Re}\left(\bar{z}_{jk}^s S_{jk}\right) - |z_{jk}^s|^2 \ell_{jk}, \qquad j \to k \in E \tag{5.3b}$$

$$v_j \ell_{jk} = |S_{jk}|^2, \qquad j \to k \in E \tag{5.3c}$$

This model is first proposed in [24, 25] for radial networks and is the most commonly used branch flow model in the literature. These equations express the same properties as (5.1) and can be derived by substituting (5.2) into (5.1) to eliminate (ℓ_{kj}, S_{kj}) on each line $j \to k \in E$ (Exercise 5.4):

1 *Power balance*: (5.1a) reduces to (5.3a).
2 *Ohm's law*: (5.1b)(5.1c) reduce to (5.3b).
3 *Apparent power*: (5.1d) reduces to (5.3c).
4 *Cycle condition*: (5.1e) becomes vacuous under assumption C5.1 and when $y_{jk}^m = y_{kj}^m = 0$.

Comparing with (5.1), the inclusion of nonzero shunt admittances (y_{jk}^m, y_{kj}^m) introduces two requirements in modeling: the need for line variables in both directions and for the cycle condition (5.1e).

Despite the complex notation, (5.3) is a set of $2(N+1+M)$ real equations in $3(N+1+M)$ real variables $x = (p_i, q_i, v_i, \ell_{jk}, P_{jk}, Q_{jk})$ and a shorthand for:

$$\sum_{k:j\to k} P_{jk} = \sum_{i:i\to j} \left(P_{ij} - r_{ij}\ell_{ij}\right) + p_j, \qquad j \in \overline{N}$$

$$\sum_{k:j\to k} Q_{jk} = \sum_{i:i\to j} \left(Q_{ij} - x_{ij}\ell_{ij}\right) + q_j, \qquad j \in \overline{N}$$

$$v_j - v_k = 2\left(r_{jk}P_{jk} + x_{jk}Q_{jk}\right) - (r_{jk}^2 + x_{jk}^2)\ell_{jk}, \qquad j \to k \in E$$

$$v_j \ell_{jk} = P_{jk}^2 + Q_{jk}^2, \qquad j \to k \in E$$

Since $M = N$, there are $(4N+2)$ equations in $(6N+3)$ real variables. Given $(2N+1)$ of these variables (e.g., given $v_0 = 1$ and non-slack bus injections (p_j, q_j), $j \in N$), the power flow problem is to determine the remaining $4N+2$ variables from these equations. There can be zero, one or more than one solutions.

This model can also be written compactly in vector form in terms of the $(N+1)\times N$ incidence matrix C defined as:

$$C_{jl} = \begin{cases} 1 & \text{if } l = j \to k \text{ for some bus } k \\ -1 & \text{if } l = i \to j \text{ for some bus } i \\ 0 & \text{otherwise} \end{cases} \tag{5.4}$$

Let $C^+ := \max\{C, 0\}$ and $C^- := \min\{C, 0\}$ denote the matrices containing only the

source nodes and destination nodes respectively of the (directed) lines. Then (5.3) is:

$$s = CS - C^{-} z\ell \tag{5.5a}$$

$$C^{\mathsf{T}} v = 2\mathrm{Re}\left(z^{\mathsf{H}} S\right) - \bar{z} z \ell \tag{5.5b}$$

$$|S|^2 = \mathrm{diag}\left((C^{+})^{\mathsf{T}} v \ell^{\mathsf{T}}\right) \tag{5.5c}$$

where $z := \mathrm{diag}(z_{jk}, j \to k \in E)$, $\bar{z}$ is the componentwise complex conjugate of the diagonal matrix z, and $|S|^2$ is the vector $|S|^2 := (|S_{jk}|^2, j \to k \in E)$.

Example 5.2 (Graph orientation)**.** Intuitively nodal injections and voltages (s, v) should not depend on the orientation of the graph while branch currents and powers (ℓ, S) do, since branch variables are defined only in the direction of the lines, not in the opposite direction. We can formally relate the power flow solutions defined for opposite graph orientations. Specifically, consider the opposite orientation where the direction of every line is reversed from that in (5.3). The resulting power flow equations are:

$$\sum_{k:j\to k} \hat{S}_{jk} = \sum_{i:i\to j} \left(\hat{S}_{ij} - z^s_{ij} \hat{\ell}_{ij}\right) + \hat{s}_j, \qquad j \in \overline{N} \tag{5.6a}$$

$$\hat{v}_k - \hat{v}_j = 2\mathrm{Re}\left(\bar{z}^s_{jk} \hat{S}_{kj}\right) - |z^s_{jk}|^2 \hat{\ell}_{kj}, \qquad k \to j \in E \tag{5.6b}$$

$$\hat{v}_k \hat{\ell}_{kj} = |\hat{S}_{kj}|^2, \qquad k \to j \in E \tag{5.6c}$$

An example is the down and up orientations below. Then it can be shown that (5.3) and (5.6) are equivalent in the sense that there is a bijection g such that x is a power flow solution of (5.3) if and only if $\hat{x} := g(x)$ is a power flow solution of (5.6) (Exercise 5.5). Indeed $\hat{x} = g(x)$ is given by:

$$\hat{s}_j := s_j, \qquad \hat{v}_j := v_j, \qquad \hat{\ell}_{kj} := \ell_{jk}, \qquad \hat{S}_{kj} := -\left(S_{jk} - z^s_{jk} \ell_{jk}\right) \tag{5.7}$$

□

Without loss of generality we take bus 0 as the root of the tree. Two particularly convenient graph orientations are where every line points *away from* bus 0 and where every line points *towards* bus 0; see Figure 5.2. For every bus j there is a unique node i that is adjacent to j on the path from bus 0 to bus j. We present two equivalent sets of power flow equations, one for each graph orientation.

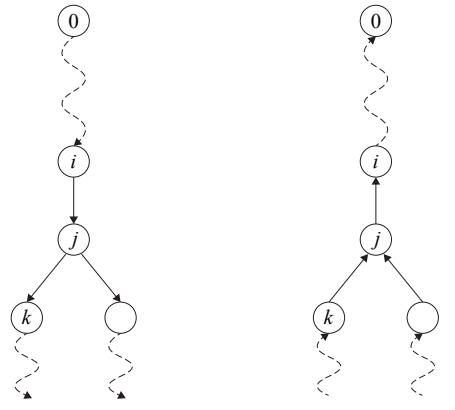
Down orientation: lines point away from bus 0.

When all lines point away from bus 0, the DistFlow equations (5.3) reduce to:

$$\sum_{k:j\to k} S_{jk} = S_{ij} - z^s_{ij} \ell_{ij} + s_j, \qquad j \in \overline{N} \tag{5.8a}$$

$$v_j - v_k = 2\mathrm{Re}\left(\bar{z}^s_{jk} S_{jk}\right) - |z^s_{jk}|^2 \ell_{jk}, \qquad j \to k \in E \tag{5.8b}$$

$$v_j \ell_{jk} = |S_{jk}|^2, \qquad j \to k \in E \tag{5.8c}$$



(a) Down orientation (b) Uporientation

Figure 5.2 Graph orientations for radial networks.

where, in (5.8a), $S_{ij} - z_{ij}\ell_{ij}$ is the receiving-end power at bus j from i, and bus $i := i(j)$ denotes the unique adjacent node of j on the path from node 0 to node j, with the understanding that when $j = 0$ then $S_{i0} = 0$ and $\ell_{i0} = 0$. When j is a leaf node[2], all $S_{jk} = 0$ in (5.8a).

Up orientation: lines point towards bus 0.

When the graph orientation is opposite to that in Case 1, BFM is specified by the following equations in $\overline{x} := (\overline{s}, \overline{v}, \overline{\ell}, \overline{S}) \in \mathbb{R}^{3(2N+1)}$:

$$\overline{S}_{ji} = \sum_{k:k\to j} \left(\overline{S}_{kj} - z^s_{kj}\overline{\ell}_{kj}\right) + \overline{s}_j, \qquad j \in \overline{N} \tag{5.9a}$$

$$\overline{v}_k - \overline{v}_j = 2\mathrm{Re}\left(\overline{z}^s_{kj}\overline{S}_{kj}\right) - |z^s_{kj}|^2\overline{\ell}_{kj}, \qquad k \to j \in E \tag{5.9b}$$

$$\overline{v}_k\overline{\ell}_{kj} = |\overline{S}_{kj}|^2, \qquad k \to j \in E \tag{5.9c}$$

where $i := i(j)$ in (5.9a) denotes the node adjacent to j on the unique path between node 0 and node j. The boundary condition is defined by $\overline{S}_{ji} = 0$ in (5.9a) when $j = 0$ and $\overline{S}_{kj} = 0, \overline{\ell}_{kj} = 0$ in (5.9a) when j is a leaf node. For an advantage of this orientation see Remark 5.2.

[2] A node j is a *leaf* node if there exists no k such that $j \to k \in E$.

5.1.4 Angle recovery

We now explain how to obtain voltage and current angles $(\angle V_j, \angle I_{jk})$ from a power flow solution x of (5.1). It applies to a solution x of the DistFlow equations (5.3), (5.8) or (5.9) with $\alpha_{jk} := 1$ in (5.10).

Given any x define the vector $\beta(x) \in \mathbb{R}^{2M}$ of line angles as a function of x by

$$\beta_{jk}(x) := \angle\left(\bar{\alpha}_{jk} v_j - \bar{z}^s_{jk} S_{jk}\right), \qquad (j,k) \in E \tag{5.10a}$$

$$\beta_{kj}(x) := \angle\left(\bar{\alpha}_{kj} v_k - \bar{z}^s_{kj} S_{kj}\right), \qquad (j,k) \in E \tag{5.10b}$$

It can be shown that, if x is a power flow solution of (5.1), then $(\beta_{jk}(x), \beta_{kj}(x))$ are voltage angle differences across line (j,k) (Exercise 5.1), i.e.,

$$\beta_{jk}(x) \;=\; \angle V_j - \angle V_k, \qquad \beta_{kj}(x) \;=\; \angle V_k - \angle V_j, \qquad (j,k) \in E \tag{5.11}$$

This implies in particular that $\beta_{jk}(x) = -\beta_{kj}(x)$, even in the absence of assumption C5.1.

Recall the $(N+1) \times N$ incidence matrix C defined in (5.4). It is proved in Theorem 5.2 below that the cycle condition (5.1e) is equivalent to:

$$\exists \theta \in \mathbb{R}^{N+1} \quad \text{s.t.} \quad \beta(x) \;=\; C^{\mathsf{T}} \theta \tag{5.12a}$$

where $\beta(x) := (\beta_{jk}(x), (j,k) \in E)$. When the network graph G is a (connected) tree, its incidence matrix C^{T} has rank $N = M$. The null space of C^{T} is span$(\mathbf{1})$ and its pseudo-inverse $\left(C^{\mathsf{T}}\right)^{\dagger} = C\left(C^{\mathsf{T}} C\right)^{-1}$ (Exercise 5.2 shows that C^{T} has full row rank and its pseudo-inverse is therefore given by Corollary A.20.2 of Appendix A.7). Given a power flow solution x of (5.1), a solution of (5.12a) is therefore

$$\theta = C\left(C^{\mathsf{T}} C\right)^{-1} \beta(x) \;+\; \phi \mathbf{1} \tag{5.12b}$$

for an arbitrary angle $\phi \in \mathbb{R}$. The angle ϕ can be fixed by choosing (say) bus 0 as a reference for voltage angles, i.e., setting $\theta_0 := 0$. An equivalent way to compute θ is to use (5.11) iteratively. Let P_j denote the unique path from bus 0 to bus j in the directed graph with orientation pointing away from bus 0. Set $\angle\theta_0$ to an arbitrary value. For $j = 1, \ldots, N+1$,

$$\angle\theta_j \;:=\; \angle\theta_0 \;-\; \sum_{(i,k) \in \mathbb{P}_j} \angle\beta_{ik} \tag{5.12c}$$

The voltage and current phasors can then be recovered from (5.11) and (5.12a)(5.12b). Pick any solution $\theta(x)$ in (5.12b), and without loss of generality, we can project it to $\theta_j(x) \in (-\pi, \pi]$. The voltage and current phasors (V, I) can then be obtained in terms of x as:

$$V_j \;:=\; \sqrt{v_j}\, e^{\mathbf{i}\theta_j(x)}, \qquad I_{jk} \;:=\; \sqrt{\ell_{jk}}\, e^{\mathbf{i}\left(\theta_j(x) - \angle S_{jk}\right)} \tag{5.12d}$$

where $\angle S_{jk} := \tan^{-1}(Q_{jk}/P_{jk})$ is the power factor angle.

5.1.5 Power flow solutions

In this section we first illustrate the solution of the branch flow model (5.9) using a simple two-bus network. The power flow solutions in the example lie on the surface of an ellipse. We prove that this feature of hollow solution set is general.

Example 5.3 (Two buses connected by a line)**.** Consider two buses 0 and 1 connected by a line characterized by a series impedance $z = r + \mathbf{i}x$ and zero shunt admittances. The power balance at bus 0 (noting that $S_{0k} := 0$) and the other DistFlow equations over line $1 \to 0$ are given by:

$$p_0 - r\ell = -p_1, \qquad q_0 - x\ell = -q_1 \tag{5.13a}$$

$$v_1 - v_0 = 2(rp_1 + xq_1) - (r^2 + x^2)\ell \tag{5.13b}$$

$$p_1^2 + q_1^2 = v_1\ell \tag{5.13c}$$

where the voltage v_0 and the injections p_1, q_1 are given. Suppose $r = x = 1$, $v_0 = 1$pu and $q_1 = 0$.

1 Show that power flow solutions (p_0, q_0, v_1, ℓ) exist if and only if

$$\frac{1}{2}\left(1 - \sqrt{2}\right) \le p_1 \le \frac{1}{2}\left(1 + \sqrt{2}\right)$$

2 For each injection value p_1 that satisfies the condition in part 1, find (p_0, q_0, v, ℓ) and show in particular that there are two voltage solutions v_1 given by

$$v_1 = \frac{1}{2}\left(1 + 2p_1 \mp \sqrt{\Delta}\right)$$

where $\Delta := 4p_1(1 - p_1) + 1$.

3 Show that the locus (v_1, p_1) that satisfies (5.13) is a (rotated) ellipse. Plot the two solutions for v_1 in Part 2 as functions of p_1. These two curves form the ellipse.

4 Show that the lowest voltage solution is $v_1 = 0$ pu attained at $p_1 = 0$ pu and the highest voltage solution is $v_1 = 2$ pu attained at $p_1 = 1$ pu.

Solution.

1 Since (p_1, q_1, v_0) are given and we are to solve for (p_0, q_0, v_1, ℓ), substitute v_1 from (5.13b) into (5.13c) to get (noting $q_1 = 0$ and $v_0 = r = x = 1$):

$$2\ell^2 - (1 + 2p_1)\ell + p_1^2 = 0 \tag{5.14}$$

There is a solution for ℓ if and only

$$(1 + 2p_1)^2 - 8p_1^2 = 1 + 4p_1 - 4p_1^2 \ge 0$$

or if and only if

$$\frac{1}{2}\left(1 - \sqrt{2}\right) \le p_1 \le \frac{1}{2}\left(1 + \sqrt{2}\right)$$

2 Let $\Delta := 4p_1(1-p_1)+1$. We have from (5.14)

$$\ell = \frac{1}{4}\left(1+2p_1 \pm \sqrt{\Delta}\right)$$

Hence

$$\begin{aligned} p_0 = \ell - p_1 &= \frac{1}{4}\left(1-2p_1 \pm \sqrt{\Delta}\right) \\ q_0 = \ell &= \frac{1}{4}\left(1+2p_1 \pm \sqrt{\Delta}\right) \\ v_1 = 1+2p_1-2\ell &= \frac{1}{2}\left(1+2p_1 \mp \sqrt{\Delta}\right) \end{aligned}$$

3 The set of points $x \in \mathbb{R}^n$ that satisfy

$$(x-c)^T A(x-c) = x^T Ax - 2c^T x + \|c\|^2 = 1$$

is an ellipse if $c \in \mathbb{R}^n$ and A is a real (symmetric) positive definite matrix. Substitute $v_1\ell = p_1^2 + q_1^2$ into (5.13b) to get $v_1 - 1 = 2p_1 - 2\frac{p_1^2}{v_1}$, i.e.,

$$\begin{aligned} \left(2p_1^2 - 2p_1v_1 + v_1^2\right) - v_1 &= 0 \\ \begin{bmatrix} p_1 & v_1 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} p_1 \\ v_1 \end{bmatrix} - \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} p_1 \\ v_1 \end{bmatrix} &= 0 \\ \begin{bmatrix} p_1 & v_1 \end{bmatrix} \underbrace{\begin{bmatrix} 8 & -4 \\ -4 & 4 \end{bmatrix}}_{A} \begin{bmatrix} p_1 \\ v_1 \end{bmatrix} - 2\underbrace{\begin{bmatrix} 0 & 2 \end{bmatrix}}_{c^T} \begin{bmatrix} p_1 \\ v_1 \end{bmatrix} + 1 &= 1 \end{aligned}$$

Since $A > 0$ is positive definite, (p_1, v_1) traces out an ellipse. It is shown in Figure 5.3 as the high voltage solution and the low voltage solution for v_1 as functions of p_1.

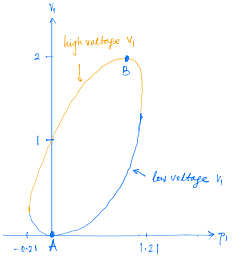


Figure 5.3 High and low voltage solutions v_1 as functions of injection p_1.

4 The figure confirms that the lowest voltage solution is attained at $v_1 = 0$ pu (point A when $p_1 = 0$) and the highest voltage is attained at $v_1 = 2$ pu (point B when $p_1 = 1$ pu). This can also be proved analytically, as follows.

Let $\underline{u}(p_1)$ and $\overline{u}(p_1)$ denote the low voltage solution and the high voltage

solution respectively:

$$\underline{u}(p_1) := \frac{1}{2}\left(1+2p_1 \; - \; \sqrt{4p_1(1-p_1)+1}\right)$$
$$\overline{u}(p_1) := \frac{1}{2}\left(1+2p_1 \; + \; \sqrt{4p_1(1-p_1)+1}\right)$$

Their derivatives are:

$$\underline{u}'(p_1) := 1 \; - \; \frac{1-2p_1}{\sqrt{4p_1(1-p_1)+1}}$$
$$\overline{u}'(p_1) := 1 \; + \; \frac{1-2p_1}{\sqrt{4p_1(1-p_1)+1}}$$

Therefore $\underline{u}'(p_1) = 0$ if and only if

$$1-2p_1 = \sqrt{4p_1(1-p_1)+1} \tag{5.15}$$

Taking square on both sides (which may introduce spurious solution for p_1), $\underline{u}'(p_1) = 0$ *only if*

$$p_1(p_1-1) = 0$$

i.e., $p_1 = 0$ or 1. Clearly, $p_1 = 1$ does not satisfy (5.15) and hence is not a solution. Moreover it can be checked that $\underline{u}'(0) = 0$, $\underline{u}(p_1)$ is decreasing for $p_1 \le 0$ and increasing for $p_1 \ge 0$. Hence $p_1 = 0$ is a minimum and $\underline{u}(0) = 0$ pu.

Similarly $\overline{u}'(p_1) = 0$ if and only if

$$2p_1 - 1 = \sqrt{4p_1(1-p_1)+1} \tag{5.16}$$

Taking square on both sides, $\overline{u}'(p_1) = 0$ only if

$$p_1(p_1-1) = 0$$

i.e., $p_1 = 0$ or 1. Clearly, $p_1 = 0$ does not satisfy (5.16) and hence is not a solution. Moreover it can be checked that $\underline{u}'(1) = 0$, $\underline{u}(p_1)$ is increasing for $p_1 \le 1$ and decreasing for $p_1 \ge 1$. Hence $p_1 = 1$ is a maximum and $\overline{u}(1) = 2$ pu. □

For the two-bus network in Example 5.3 power flow solutions, when projected onto the (p_1, v_1) coordinate, form an ellipse without the interior. This feature of hollow solution set is generally true for the DistFlow model (5.3), (5.8), or (5.9) as the following result shows. Let

$$\mathbb{X}_{\text{df}} \; := \; \{x : (s, v, \ell, S) \in \mathbb{R}^{6N+3} \mid x \text{ satisfies (5.3)}\}$$

Theorem 5.1 (Hollow solution set)**.** Suppose the network graph G is connected. If $\hat{x}$ and $\tilde{x}$ are distinct power flow solutions in $\mathbb{X}_{\text{df}}$ with the same voltage $\hat{v}_0 = \tilde{v}_0$ at the root bus 0, then no convex combination of $\hat{x}$ and $\tilde{x}$ can be in $\mathbb{X}_{\text{df}}$. In particular $\mathbb{X}_{\text{df}}$ is nonconvex.

Proof Suppose $\hat{x} \neq \tilde{x}$ are distinct power flow solutions in $\mathbb{X}_{\rm df}$. Fix any $a \in (0,1)$ and consider $x := a\hat{x} + (1-a)\tilde{x}$. We now show that if $x \in \mathbb{X}_{\rm df}$ then $\hat{x} = \tilde{x}$, contradicting that $\hat{x}$ and $\tilde{x}$ are distinct.

Suppose $x \in \mathbb{X}_{\rm df}$. In particular $v_j \ell_{jk} = |S_{jk}|^2$ by (5.3c). Substituting $x := (\hat{x}+\tilde{x})/2$, we have

$$\frac{1}{4}(\hat{v}_j + \tilde{v}_j)(\hat{\ell}_{jk} + \tilde{\ell}_{jk}) = \frac{1}{4}\left|\hat{S}_{jk} + \tilde{S}_{jk}\right|^2, \qquad j \to k \in E$$

Substituting $\hat{v}_j \hat{\ell}_{jk} = |\hat{S}_{jk}|^2$ and $\tilde{v}_j \tilde{\ell}_{jk} = |\tilde{S}_{jk}|^2$ yeilds

$$\hat{v}_j \tilde{\ell}_{jk} + \tilde{v}_j \hat{\ell}_{jk} = 2\,\mathrm{Re}\left(\hat{S}_{jk}^{\mathsf{H}} \tilde{S}_{jk}\right) \tag{5.17a}$$

The right-hand side satisfies

$$2\,\mathrm{Re}\left(\hat{S}_{jk}^{\mathsf{H}} \tilde{S}_{jk}\right) \leq 2|\tilde{S}_{jk}||\hat{S}_{jk}| \tag{5.17b}$$

with equality if and only if $\angle \hat{S}_{jk} = \angle \tilde{S}_{jk} \pmod{2\pi}$. The left-hand side of (5.17a) is

$$\hat{v}_j \tilde{\ell}_{jk} + \tilde{v}_j \hat{\ell}_{jk} = \eta_j |\tilde{S}_{jk}|^2 + \eta_j^{-1} |\hat{S}_{jk}|^2 \;\geq\; 2|\tilde{S}_{jk}||\hat{S}_{jk}| \tag{5.17c}$$

with equality if and only if $\eta_j |\tilde{S}_{jk}| = |\hat{S}_{jk}|$, where for $j \in \overline{N}$, $\eta_j := \hat{v}_j / \tilde{v}_j$. But (5.17) implies that equalities are attained in both (5.17b) and (5.17c), and hence

$$\eta_j \tilde{S}_{jk} \;=\; \hat{S}_{jk} \text{ and } \eta_j \tilde{\ell}_{jk} \;=\; \hat{\ell}_{jk}, \qquad j \in N \tag{5.18}$$

(The second equation in (5.18) follows from (5.17c): $\eta_j \tilde{\ell}_{jk} + \hat{\ell}_{jk} = 2|\tilde{S}_{jk}||\hat{S}_{jk}|/\tilde{v}_j = 2\sqrt{\eta_j \tilde{\ell}_{jk} \hat{\ell}_{jk}}$ and squaring both sides yields the equation.) Define $\eta_0 := \hat{v}_0 / \tilde{v}_0 = 1$. Then for each line $j \to k \in E$ we have, using (5.3b),

$$\begin{aligned}
\eta_k = \frac{\hat{v}_k}{\tilde{v}_k} \;&=\; \frac{\hat{v}_j - 2\mathrm{Re}(z_{jk}^{\mathsf{H}} \hat{S}_{jk}) + |z_{jk}|^2 \hat{\ell}_{jk}}{\tilde{v}_j - 2\mathrm{Re}(z_{jk}^{\mathsf{H}} \tilde{S}_{jk}) + |z_{jk}|^2 \tilde{\ell}_{jk}} \\
&= \frac{\eta_j \left(\tilde{v}_j - 2\mathrm{Re}(z_{jk}^{\mathsf{H}} \tilde{S}_{jk}) + |z_{jk}|^2 \tilde{\ell}_{jk}\right)}{\tilde{v}_j - 2\mathrm{Re}(z_{jk}^{\mathsf{H}} \tilde{S}_{jk}) + |z_{jk}|^2 \tilde{\ell}_{jk}} \;=\; \eta_j
\end{aligned}$$

where the third equality follows from (5.18). This implies, since the network graph G is connected, that $\eta_j = \eta_0 = 1$ for all $j \in \overline{N}$, i.e. $\hat{v}_j = \tilde{v}_j$, $j \in \overline{N}$.

We have thus shown that $\hat{S} = \tilde{S}$, $\hat{\ell} = \tilde{\ell}$, $\hat{v} = \tilde{v}$, and hence, by (5.3a), $\hat{s} = \tilde{s}$, i.e., $\hat{x} = \tilde{x}$. This completes the proof. □

This property of the power flow solution set is illustrated vividly in several numerical examples in [26, 27, 28, 29]. It is used in Theorem 11.1 of Chapter 11.3 to prove that if any convex relaxation of OPF on a radial network is exact in a strong sense, then the optimal solution of the relaxation is unique.

5.2 Equivalence

The branch flow models for radial networks are (5.1) with shunt admittances and without assumption C5.1 and the DistFlow equations(5.3), (5.8) and (5.9), when shunt admittances are zero and assumption C5.1 holds. They are defined by different sets of power flow equations from the bus injection model (4.26a) studied in Chapter 4.3, reproduced here:

$$s_j = \sum_{k:j\sim k} \left(y^s_{jk}\right)^{\mathsf{H}} \left(|V_j|^2 - V_j V_k^{\mathsf{H}}\right) + \left(y^m_{jj}\right)^{\mathsf{H}} |V_j|^2, \qquad j \in \overline{N} \tag{5.19}$$

Yet all of them are models of Kirchhoff's and Ohm's laws. In this section we show that these models are equivalent in a precise sense.

To this end we first extend the branch flow model (5.1) to general networks. We then use these generalized branch flow models, (5.20) and (5.21) below, as a bridge to relate BFM (5.1), (5.3), (5.8), (5.9) for radial networks to BIM (5.19) for general networks.

5.2.1 Extension to general networks

Complex form.

The branch flow model for a general network possibly with cycles in the complex form is defined by the following power flow equations in the variables $(s, V, I, S) \in \mathbb{C}^{2(N+1)+4M}$ (from (4.1)(4.2)):

$$s_j = \sum_{k:j\sim k} S_{jk}, \qquad j \in \overline{N} \tag{5.20a}$$

$$I_{jk} = \tilde{y}_{jk} V_j - y^s_{jk} V_k, \qquad I_{kj} = \tilde{y}_{kj} V_k - y^s_{kj} V_j, \qquad (j,k) \in E \tag{5.20b}$$

$$S_{jk} = V_j I_{jk}^{\mathsf{H}}, \qquad S_{kj} = V_k I_{kj}^{\mathsf{H}}, \qquad (j,k) \in E \tag{5.20c}$$

where in (5.20b),

$$\tilde{y}_{jk} := y^s_{jk} + y^m_{jk}, \qquad \tilde{y}_{kj} := y^s_{kj} + y^m_{kj}$$

Equation (5.20a) imposes power balance at each bus, (5.20b) describes the Ohm's law and KCL, and (5.20c) defines branch power in terms of the associated voltage and current. For convenience we include V_0 in the vector variable $V := (V_j, j \in \overline{N})$ with the understanding that $V_0 := 1\angle 0°$ is fixed. This model does not require assumption C5.1 and allows nonzero shunt admittances (y^m_{jk}, y^m_{kj}). It serves as a bridge between the bus injection model (5.19) in complex form and the branch flow models in the real domain.

Real form.

The following branch flow model relaxes the angles of voltages and currents and are applicable to general networks:

$$s_j = \sum_{k:j\sim k} S_{jk}, \qquad j \in \overline{N} \quad (5.21a)$$

$$\left|\alpha_{jk}\right|^2 v_j - v_k = 2\mathrm{Re}\left(\alpha_{jk}\left(z^s_{jk}\right)^{\mathsf{H}} S_{jk}\right) - \left|z^s_{jk}\right|^2 \ell_{jk}, \quad (j,k) \in E \quad (5.21b)$$

$$\left|\alpha_{kj}\right|^2 v_k - v_j = 2\mathrm{Re}\left(\alpha_{kj}\left(z^s_{kj}\right)^{\mathsf{H}} S_{kj}\right) - \left|z^s_{kj}\right|^2 \ell_{kj}, \quad (j,k) \in E \quad (5.21c)$$

$$\left|S_{jk}\right|^2 = v_j \ell_{jk}, \quad \left|S_{kj}\right|^2 = v_k \ell_{kj}, \qquad (j,k) \in E \quad (5.21d)$$

$$\exists \theta \in \mathbb{R}^{N+1} \text{ s.t. } \beta_{jk}(x) = \theta_j - \theta_k, \quad \beta_{kj}(x) = \theta_k - \theta_j, \qquad (j,k) \in E \quad (5.21e)$$

where $\beta_{jk}(x)$ and $\beta_{kj}(x)$ are defined in (5.10) and reproduced here:

$$\beta_{jk}(x) := \angle\left(\alpha^{\mathsf{H}}_{jk} v_j - \left(z^s_{jk}\right)^{\mathsf{H}} S_{jk}\right), \qquad \beta_{kj}(x) := \angle\left(\alpha^{\mathsf{H}}_{kj} v_k - \left(z^s_{kj}\right)^{\mathsf{H}} S_{kj}\right)$$

Compared with (5.1) for radial networks, the model (5.21) differs only in its cycle condition: the linear cycle condition (5.1e) for radial networks becomes a nonlinear cycle condition (5.21e) for general networks. It ensures that the line angles $\beta(x) := (\beta_{jk}(x), (j,k) \in E)$ implied by a power flow solution x of (5.21) is consistent with voltage angles in model (5.20). Since (5.21e) implies that $\beta(x) = C^{\mathsf{T}}\theta$ and $\beta_{jk}(x) = -\beta_{kj}(x)$, the nodal voltage angles θ are also given by (5.12).

The model (5.21) does not require assumption C5.1 and allows nonzero shunt admittances (y^m_{jk}, y^m_{kj}). Let $x := (s, v, \ell, S) = (p_j, q_j, v_j, \ell_{jk}, \ell_{kj}, P_{jk}, P_{kj}, Q_{jk}, Q_{kj},\ j \in \overline{N},\ (j,k) \in E)$. Then (5.21) is a set of $2(N+1)+6M$ real equations in the $3(N+1)+6M$ real variables in x and $N+1$ variables in θ. The power flow problem is: given $2(N+1)$ of these variables (e.g., $(p_j, q_j, j \in N)$ and (v_0, θ_0)), determine the remaining $2(N+1)+6M$ variables from (5.21). Equations (5.21d) are quadratic, the cycle condition (5.21e) is nonlinear, and the rest are linear in x. The major simplification for radial networks is the replacement of the nonlinear cycle condition (5.21e) for general networks by the linear cycle condition (5.1e). When shunt admittances are assumed zero and assumption C5.1 holds, then the cycle condition becomes vacuous for radial networks as in the DistFlow equations.

5.2.2 Equivalence of BFM and BIM

Let the set of solutions (s, V) of BIM be:

$$\mathbb{V} := \mathbb{V}(\theta_0) := \{(s,V) \in \mathbb{C}^{2(N+1)} \mid (s,V) \text{ satisfies (5.19)}\}$$

where we have fixed a reference angle $\angle V_0 = \theta_0$. Let the sets of solutions of BFM be:

$$\tilde{\mathbb{X}} := \tilde{\mathbb{X}}(\theta_0) := \{\tilde{x} : (s, V, I, S) \in \mathbb{C}^{2(N+1)+4M} \mid \tilde{x} \text{ satisfies (5.20)}\}$$
$$\mathbb{X}_{\text{meshed}} := \mathbb{X}_{\text{meshed}}(\theta_0) := \{x : (s, v, \ell, S) \in \mathbb{R}^{3(N+1)+6M} \mid x \text{ satisfies (5.21)}\}$$
$$\mathbb{X}_{\text{tree}} := \mathbb{X}_{\text{tree}}(\theta_0) := \{x : (s, v, \ell, S) \in \mathbb{R}^{3(N+1)+6M} \mid x \text{ satisfies (5.1)}\}$$
$$\mathbb{X}_{\text{df}} := \mathbb{X}_{\text{df}}(\theta_0) := \{x : (s, v, \ell, S) \in \mathbb{R}^{3(N+1+M)} \mid x \text{ satisfies (5.3) under C5.1 and } y^m_{jk} = y^m_{kj} = 0\}$$

where a reference angle $\angle V_0 = \theta_0$ is fixed so that voltage phasors can be uniquely recovered from power flow solutions in $\mathbb{X}_{\text{meshed}}(\theta_0)$, $\mathbb{X}_{\text{tree}}(\theta_0)$ and $\mathbb{X}_{\text{df}}(\theta_0)$. We say two sets A and B are *equivalent*, denoted by $A \equiv B$, if there is a bijection between them. The equivalence of these power flow models is clarified in the following theorem and illustrated in Figure 5.4.

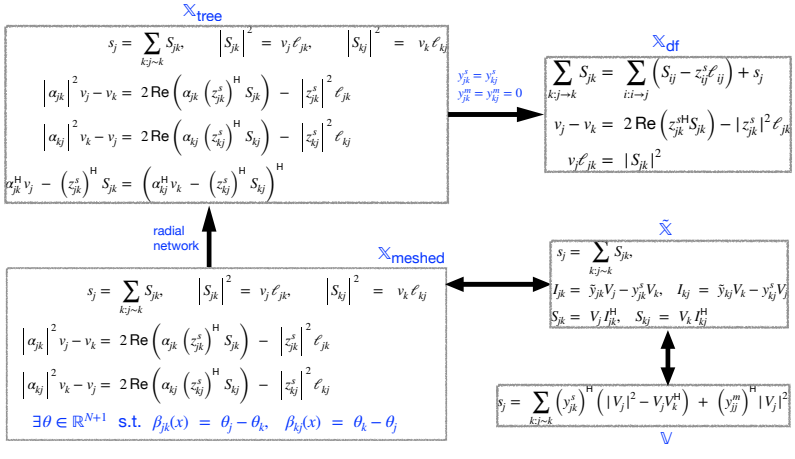


Figure 5.4 Equivalence of BFM and BIM. Proof focuses on $\tilde{\mathbb{X}} \equiv \mathbb{X}_{\text{meshed}}$ and $\mathbb{X}_{\text{meshed}} \equiv \mathbb{X}_{\text{tree}}$.

Theorem 5.2 (Equivalence)**.** Suppose the network G is connected.

1 $\mathbb{V} \equiv \tilde{\mathbb{X}} \equiv \mathbb{X}_{\text{meshed}}$.
2 If G is a tree then $\mathbb{X}_{\text{meshed}} \equiv \mathbb{X}_{\text{tree}}$.
3 Suppose $y^s_{jk} = y^s_{kj}$ (assumption C5.1) and $y^m_{jk} = y^m_{kj} = 0$ for all lines (j,k). If G is a tree then $\mathbb{X}_{\text{tree}} \equiv \mathbb{X}_{\text{df}}$.

Proof *Part 1:* $\mathbb{V} \equiv \tilde{\mathbb{X}} \equiv \mathbb{X}_{meshed}$. It is obvious $\mathbb{V} \equiv \tilde{\mathbb{X}}$ since, given $(s,V) \in \mathbb{V}$, define I by (5.20b) and S by (5.20c) and the resulting $(s,V,I,S) \in \tilde{\mathbb{X}}$. Conversely given $(s,V,I,S) \in \tilde{\mathbb{X}}$, substituting (5.20b)(5.20c) into (5.20a) shows $(s,V) \in \mathbb{V}$. Clearly these two mappings are the inverses of each other.

To show $\tilde{\mathbb{X}} \equiv \mathbb{X}_{\text{meshed}}$, fix an $\tilde{x} := (s,V,I,S) \in \tilde{\mathbb{X}}$. Define (v,ℓ) by:

$$v_j := |V_j|^2, \qquad \ell_{jk} := |I_{jk}|^2, \qquad \ell_{kj} := |I_{kj}|^2 \tag{5.22}$$

We now show that $x := (s, v, \ell, S) \in \mathbb{X}_{\text{meshed}}$. That x satisfies (5.21a) follows from (5.20a). Taking the squared magnitude on both sides of (5.20c) gives (5.21d). For (5.21b) rewrite the first equation in (5.20b) as

$$V_k = \alpha_{jk} V_j - z^s_{jk} \left(\frac{S_{jk}}{V_j} \right)^{\mathsf{H}} \tag{5.23}$$

where we have substituted $I_{jk} := S^{\mathsf{H}}_{jk} / V^{\mathsf{H}}_j$ from (5.20c). Taking the squared magnitude on both sides gives

$$v_k = \left|\alpha_{jk}\right|^2 v_j + \left|z^s_{jk}\right|^2 \ell_{jk} - 2\text{Re}\left(\alpha_{jk} \left(z^s_{jk} \right)^{\mathsf{H}} S_{jk} \right)$$

which is (5.21b). Similarly (5.21c) can be derived from the second equation in (5.20b). From (5.20b) and (5.20c) we have

$$V_j V^{\mathsf{H}}_k = \alpha^{\mathsf{H}}_{jk} |V_j|^2 - \left(z^s_{jk} \right)^{\mathsf{H}} S_{jk}, \qquad V_k V^{\mathsf{H}}_j = \alpha^{\mathsf{H}}_{kj} |V_k|^2 - \left(z^s_{jk} \right)^{\mathsf{H}} S_{kj}$$

The definitions of $\beta_{jk}(x)$ and $\beta_{kj}(x)$ in (5.10) then imply that $\beta_{jk}(x) = \angle V_j - \angle V_k = -\beta_{kj}(x)$ and hence the cycle condition (5.21e) holds with $\theta_j := \angle V_j$. This shows $x \in \mathbb{X}_{\text{meshed}}$.

Conversely fix an $x := (s, v, \ell, S) \in \mathbb{X}_{\text{meshed}}$, i.e., x satisfies (5.21). Since $\beta_{jk}(x)$ defined in (5.10) satisfy (5.21e), i.e., $\beta(x) = C^{\mathsf{T}}\theta$ for some θ, we can construct (V, I) from x as:

$$V_j := \sqrt{v_j}\, e^{\mathbf{i}\theta_j}, \qquad I_{jk} := \sqrt{\ell_{jk}}\, e^{\mathbf{i}(\theta_j - \angle S_{jk})} \tag{5.24}$$

We now verify that $\tilde{x} := (s, V, I, S)$ satisfies (5.20). Clearly (5.20a) is (5.21a). For (5.20c), we have from (5.21d) and the construction (5.24) of (V, I) that

$$|S_{jk}| = \left| V_j I^{\mathsf{H}}_{jk} \right|, \qquad \angle S_{jk} = \angle V_j - \angle I_{jk}$$

Hence $S_{jk} = V_j I^{\mathsf{H}}_{jk}$. Similarly $S_{kj} = V_k I^{\mathsf{H}}_{kj}$. We next show that (5.20b) follows from (5.21b)(5.21c). First note that (5.20b) is equivalent to $z^s_{jk} \left(S_{jk} / V_j \right)^{\mathsf{H}} = \alpha_{jk} V_j - V_k$ which is equivalent to

$$V_j V^{\mathsf{H}}_k = \alpha^{\mathsf{H}}_{jk} v_j - z^{s\mathsf{H}}_{jk} S_{jk} \tag{5.25}$$

We now show that (5.21b) implies that the quantities on both sides of (5.25) have equal magnitudes and angles, thus establishing their equality. For their angles, the definition of $\beta_{jk}(x)$ in (5.10) implies

$$\angle \left(\alpha^{\mathsf{H}}_{jk} v_j - z^{s\mathsf{H}}_{jk} S_{jk} \right) = \beta_{jk}(x) = \theta_j - \theta_k = \angle \left(V_j V^{\mathsf{H}}_k \right)$$

where the last two equalities follow from the construction (5.24) of V_j, V_k. The squared magnitude of the right-hand side of (5.25) is

$$\begin{aligned} \left| \alpha^{\mathsf{H}}_{jk} v_j - z^{s\mathsf{H}}_{jk} S_{jk} \right|^2 &= \left|\alpha_{jk}\right|^2 v_j^2 - 2 v_j \,\text{Re}\left(\alpha_{jk} z^{s\mathsf{H}}_{jk} S_{jk} \right) + \left|z^s_{jk}\right|^2 \left|S_{jk}\right|^2 \\ &= v_j \left(\left|\alpha_{jk}\right|^2 v_j - 2\text{Re}\left(\alpha_{jk} z^{s\mathsf{H}}_{jk} S_{jk} \right) + \left|z^s_{jk}\right|^2 \ell_{jk} \right) = v_j v_k \end{aligned}$$

which is the squared magnitude of the quantity on the left-hand side of (5.25). The second equality above follows from $|S_{jk}|^2 = v_j \ell_{jk}$ from (5.21d) and the last equality follows from (5.21b). Similarly for I_{kj} in the opposite direction and hence (5.20b) follows from (5.21b)(5.21c). This proves $\tilde{x} \in \tilde{\mathbb{X}}$. Finally the mappings defined by (5.22) and (5.24) are inverses of each other, given a fixed reference angle $\angle V_0 = \theta_0$. We hence conclude $\tilde{\mathbb{X}} \equiv \mathbb{X}_{\text{meshed}}$.

Part 2: $\mathbb{X}_{meshed} \equiv \mathbb{X}_{tree}$. Suppose G is a tree. We will show that $x := (s, v, \ell, S)$ satisfies (5.21) if and only if it satisfies (5.1). It suffices to show that x satisfies (5.21e) if and only if it satisfies (5.1e). Suppose x satisfies (5.21e) which implies that $\beta_{jk}(x) = -\beta_{kj}(x)$. Using (5.10) we have

$$\angle\left(\alpha_{jk}^{\mathsf{H}} v_j \;-\; z_{jk}^{\mathsf{H}} S_{jk}\right) \;=\; \beta_{jk}(x) = -\beta_{kj}(x) \;=\; -\angle\left(\alpha_{kj}^{\mathsf{H}} v_k \;-\; z_{kj}^{\mathsf{H}} S_{kj}\right)$$

i.e., the quantities on both sides of (5.1e) have equal angles. We now show that they have equal magnitudes as well. Indeed

$$\left|\alpha_{jk}^{\mathsf{H}} v_j \;-\; z_{jk}^{\mathsf{H}} S_{jk}\right|^2 = |\alpha_{jk}|^2 v_j^2 + |z_{jk}|^2 |S_{jk}|^2 - 2\,\mathrm{Re}\left(\alpha_{jk} z_{jk}^{\mathsf{H}} v_j S_{jk}\right) \;=\; v_j v_k$$

where the last equality follows from multiplying both sides of (5.1b) by v_j and then substituting (5.1d). Similarly

$$\left|\alpha_{kj}^{\mathsf{H}} v_k \;-\; z_{kj}^{\mathsf{H}} S_{kj}\right|^2 = v_k v_j \;=\; \left|\alpha_{jk}^{\mathsf{H}} v_j \;-\; z_{jk}^{\mathsf{H}} S_{jk}\right|^2$$

This shows that $\alpha_{jk}^{\mathsf{H}} v_j \;-\; z_{jk}^{\mathsf{H}} S_{jk} = \left(\alpha_{kj}^{\mathsf{H}} v_k \;-\; z_{kj}^{\mathsf{H}} S_{kj}\right)^{\mathsf{H}}$. Hence x satisfies (5.1e). Conversely suppose x satisfies (5.1e). Adopt an arbitrary orientation of the network graph and define $\beta_{jk}(x) := \angle\left(\alpha_{jk}^{\mathsf{H}} v_j - z_{jk}^{\mathsf{H}} S_{jk}\right)$ for each directed line $j \to k$ (only). Since G is a tree, the $(N+1) \times N$ incidence matrix C has a full column rank of N and therefore $\theta := C\left(C^{\mathsf{T}} C\right)^{-1} \beta(x) + \phi \mathbf{1}$ as given by (5.12b) exists and is unique given a reference angle θ_0. Moreover θ is a solution to (5.21e) since (5.1e) implies that $\beta_{kj}(x) = -\beta_{jk}(x)$. This shows that $\mathbb{X}_{\text{meshed}} \equiv \mathbb{X}_{\text{tree}}$.

Part 3: $\mathbb{X}_{tree} \equiv \mathbb{X}_{df}$. This can be proved by substituting (5.2) into (5.1) to eliminate (ℓ_{kj}, S_{kj}) from (5.1) (see Exercise 5.4). □

Given the bijection between the solution sets of BIM and BFM, any result in one model is in principle derivable in the other. Some results however are much easier to state or derive in one model than the other. For instance BIM, which is widely used in transmission network problems, allows a much cleaner formulation of semidefinite program (SDP) relaxation (see Chapter 10). BFM for radial networks has a convenient recursive structure that allows a more efficient computation of power flows and leads to a useful linear approximation; see Chapters 5.3 and 5.4. The sufficient condition for exact relaxation in Chapter **??** provides intricate insights on power flows that are hard to formulate or prove in BIM. BFM for radial networks seems to be much more stable numerically than BIM as the network size scales up. Finally, since BFM directly models branch flows S_{jk} and currents I_{jk}, it is easier to use for some applications. One

should freely use either model depending on which is more convenient for the problem at hand.

5.3 Backward forward sweep

General iterative methods for solving power flow equations are studied in Chapter 4.4. These methods can be used not only for solving bus injection models but also branch flow models of this chapter. Tree topology however induces a spatially recursive structure in power flow equations and this structure allows an efficient computation method for solving power flow equations, called a backward forward sweep (BFS), that is unique to radial networks. The Newton-Raphson algorithm of Chapter 4.4.2 needs to compute Jacobian or solve a linear system in each iteration, a significant computational burden for large networks. The Fast Decoupled Algorithm of Chapter 4.4.3 reduces the computational effort of the Newton-Raphson algorithm, but assumes line losses are small, which is a good approximation for high-voltage transmission networks but not for distribution systems. In contrast BFS is simple, accurate, and tends to converge quickly in practice.

An outline of BFS is as follows. A power flow solution is partitioned into two groups of variables x and y. Starting from an initial vector y, the components x_i can be successively computed starting from leaf nodes and propagating towards the root (backward sweep). Given the newly updated vector x, the components y_i are then updated successively starting from the root and propagating towards the leaf nodes (forward sweep). A BFS method iterates on a backward sweep followed by a forward sweep, until convergence. It can be interpreted as a special Gauss-Seidel algorithm that exploits a spatially recursive structure enabled by tree topology.

Different BFS algorithms differ in their choices of variables x and y and the associated power flow equations. In the following we first provide in Chapter 5.3.1 a general formulation of BFS and then illustrate in Chapters 5.3.2 and 5.3.3 BFM algorithms using the complex form BFM and the DistFlow model. Their convergence of these two algorithms will be analyzed in Chapter 8.6 as examples of convergence analysis of iterative algorithms.

5.3.1 General BFS

The method of backward forward sweep can be interpreted as a Gauss-Seidel algorithm studied in Chapter 4.4.1 to compute a fixed point, with two special features.

Outer loop.

First it partitions a power flow variable into two vectors $x \in F^{n_1}$ and $y \in F^{n_2}$ where F is either $\mathbb{C}$ or $\mathbb{R}$. BFS consists of an outer loop which updates $(x(t), y(t))$ from $(x(t-1), y(t-1))$ and, for each outer iteration, two inner loops, one updating successively each component $x_i(t)$ using the Gauss-Seidel method with components of $y(t-1)$ held fixed and the other updating successively each component $y_i(t)$ using the Gauss-Seidel method with the newly updated $x(t)$ held fixed. We represent the outer iteration as a fixed-point iteration:

$$\text{Outer loop:} \quad x(t) := f(x(t); y(t-1)), \qquad y(t) := g(x(t); y(t)) \tag{5.26a}$$

where $f : F^{n_1+n_2} \to F^{n_1}$ and $g : F^{n_1+n_2} \to F^{n_2}$. By this notation we mean that each outer iteration in (5.26a) is computed iteratively in two inner loops that update components $x_j(t)$ and then $y_j(t)$ in turn, always using the latest available values, i.e.,

$$\text{Inner loop 1:} \quad x_j(t) := f_j(x_1(t), \ldots, x_{j-1}(t), x_j(t-1), \ldots, x_{n_1}(t-1); y(t-1)), \quad j = 1, \ldots, n_1 \tag{5.26b}$$

$$\text{Inner loop 2:} \quad y_j(t) := g_j(x(t); y_1(t), \ldots, y_{j-1}(t), y_j(t-1), \ldots, y_{n_2}(t-1)), \quad j = 1, \ldots, n_2 \tag{5.26c}$$

Inner loops (backward and forward sweeps).

Second the inner loops make use of a spatially recursive structure enabled by the tree topology. Specifically the partitions x and y are chosen so that, given a vector y, the update function f_j in (5.26b) for each component x_j depends only on $(x_1, \ldots, x_{j-1})$, but not other components of x. This means that, starting from $x_k(t)$ at leaf nodes k and propagating towards the root of the tree, $x_j(t)$ at nodes at successive layers are updated according to (backward sweep):

$$x_j(t) := f_j(x_1(t), \ldots, x_{j-1}(t); y(t-1)), \qquad j = 1, \ldots, n_1$$

Similarly, given an x, the update function g_j in (5.26c) for each component y_j depends only on $(y_1, \ldots, y_{j-1})$. Starting from the root and propagating towards leaf nodes, $y_j(t)$ are updated successively according to (forward sweep):

$$y_j(t) := g_j(x(t); y_1(t), \ldots, y_{j-1}(t)), \qquad j = 1, \ldots, n_2$$

We can visualize the two inner loops using the tree topology. Consider a tree network $G := (\overline{N}, E)$ where $\overline{N} := \{0, 1, \ldots, N\}$ with its root at bus $j = 0$ (instead of $j = 1$). Fix any graph orientation (it is sometimes convenient to use the up orientation if v_0 is fixed and s_0 is variable). Due to the tree topology we can always identify variables associated with a line $j \to k$, such as the line current I_{jk} or power flow S_{jk}, by either the from node j or the to node k depending on the design of (f, g) (see Chapters 5.3.2 and 5.3.3).

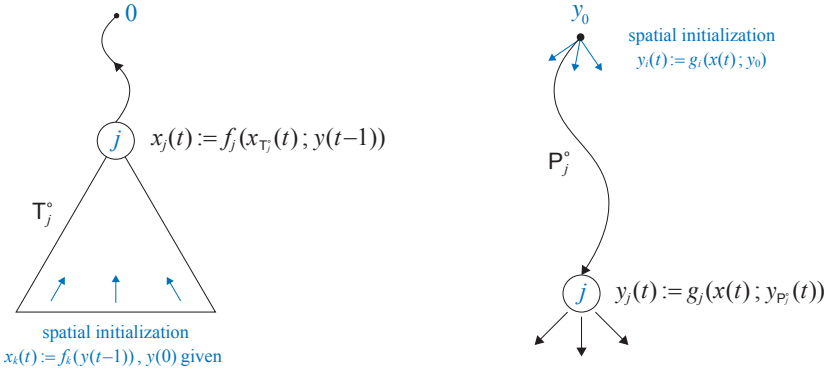
Typically the partitioning of variables into (x, y) and the update functions (f, g) are designed so that x_j depends only on x_k at its child nodes k (i.e., k is adjacent to j and farther away from the root than j regardless of the graph orientation). More generally let T_j° denote the set of buses in the subtree rooted at bus j, *not* including j. Let $x_{\mathsf{T}_j^\circ} := \left(x_k, k \in \mathsf{T}_j^\circ\right)$ denote the variables x_k in the subtree T_j°. We say that the function $f := (f_j, \forall j)$ is *spatially recursive* if, given y, f_j depends only on $x_{\mathsf{T}_j^\circ}$, but not other components of x:

$$x_j = f_j\left(x_{\mathsf{T}_j^\circ};\ y\right), \qquad j \in \overline{N}$$

This means that, at each outer iteration t, starting from the leaf nodes and propagating towards the root (bus 0) in the reverse breadth-first search order, x_j can be successively updated given vector $y(t-1)$:

Backward sweep at t: $$x_j(t) \ := \ f_j\left(x_{\mathsf{T}_j^\circ}(t);\ y(t-1)\right), \qquad j \in \overline{N}$$

as illustrated in Figure 5.5(a). The recursion is initialized at leaf nodes j where $\mathsf{T}_j^\circ := \emptyset$ so that $x_j(t) := f_j(\emptyset, y(t-1)) =: f_j(y(t-1))$ with a given $y(0)$ for outer iteration $t = 0$.



(a) Backward sweep (b) Forward sweep

Figure 5.5 General backward forward sweep.

Similarly (x, y) and (f, g) are chosen so that, given x, the components y_j depends only on y_i in the path from the root to j, not on variables at other buses further away from the root. Specifically let P_j° denote the set of buses in the unique path from the root to bus j, including the root bus 0 but *not* including j. Let $y_{\mathsf{P}_j^\circ} := \left(y_i, i \in \mathsf{P}_j^\circ\right)$. The function $g := (g_j, \forall j)$ is *spatially recursive* if, given x, g_j depends only on $y_{\mathsf{P}_j^\circ}$, but not other components of y:

$$y_j = g_j\left(x;\ y_{\mathsf{P}_j^\circ}\right), \qquad j \in N, \qquad y_0 \ \text{given}$$

At each outer iteration t, starting from the children of the root and propagating towards leaf nodes, y_j can be successively updated given vector $x(t)$:

Forward sweep at t: $$y_j(t) := g_j\left(x(t);\, y_{\mathsf{P}_j^\circ}(t)\right), \qquad j \in N$$

as illustrated in Figure 5.5(b). The recursion is initialized at children i of the root bus 0 where $\mathsf{P}_i^\circ := \{0\}$ so that $y_i(t) := g_i\left(x(t);\, y_0(t)\right) := g_i\left(x(t);\, y_0\right)$ for all outer iterations t, given y_0.

Summary.

Let $x := (x_j, j \in \overline{N})$ and $y := (y_j(t), j \in \overline{N})$. A pair (x, y) is a power flow solution if it satisfies the following power flow equations that have a spatially recursive structure:

$$x_j = f_j\left(x_{\mathsf{T}_j^\circ};\, y\right), \quad j \in \overline{N}, \qquad y_j = g_j\left(x;\, y_{\mathsf{P}_j^\circ}\right), \quad j \in N \tag{5.27a}$$

$$\mathsf{T}_i^\circ = \emptyset \text{ for all leaf nodes } j \qquad y_0 \text{ given} \tag{5.27b}$$

A BFS algorithm is a special Gauss-Seidel algorithm that computes a fixed point of (5.27) in which each outer iteration t consists of two inner loops:

$$\text{Backward sweep at } t: \quad x_j(t) := f_j\left(x_{\mathsf{T}_j^\circ}(t);\, y(t-1)\right), \qquad j \in \overline{N} \tag{5.28a}$$

$$\text{Forward sweep at } t: \quad y_j(t) := g_j\left(x(t);\, y_{\mathsf{P}_j^\circ}(t)\right), \qquad j \in N \tag{5.28b}$$

starting from the spatial initial conditions in (5.27b) and given temporal initial conditions $y(0)$ and $y_0(t) = y_0$ for all t. A more detailed description is in Algorithm 1. If the algorithm converges and the update functions (f, g) are continuous then the limit point is a fixed point of (5.27) and therefore a power flow solution. An advantage of BFS is that it does not need to calculate derivatives of power flow equations and tends to converge quickly in practice.

The design of BFS boils down to the choice of (f, g) and the partitioning (x, y) that define the power flow equations in (5.27). Given (f, g) with the spatial recursive structure in (5.27), the iterative algorithm is defined by the inner loops (5.28). These design choices are not unique and may have different convergence properties. We will study two examples in Chapters 5.3.2 and 5.3.3. Most BFS algorithms compute line currents or power flows in the backward sweep and voltages in the forward sweep. Typically the voltage at the substation (the root of the tree) is specified and that the line current or power out of a leaf node is zero. These two boundary conditions mean that the computation of line currents or powers must start from the leaf nodes and propagate backward, while that of voltages must start from the root and propagate forward.

Remark 5.1. 1 We assume for notational simplicity that each x_j or y_j is a scalar, but the description remains unchanged if x_j and y_j are vectors and the update functions f_j and g_j are vector-valued; see Example 5.4 below.

Algorithm 1: Backward forward sweep

Input: $(f_j, \mathsf{T}_j^\circ,\ j \in \overline{N})$, $(g_j, \mathsf{P}_j^\circ,\ j \in \overline{N})$, y_0 and $y(0)$.
Output: a solution (x, y) of (5.27).

1. **Initiatization**:
   - $\mathsf{T}_j^\circ := \emptyset$ for all leaf notes j.
   - $y_0(t) \leftarrow y_0$ for $t = 0, 1, \ldots$.
   - $t \leftarrow 0$.
2. **while** *stopping criterion not met* **do**
   1. $t \leftarrow t + 1$;
   2. *Backward sweep*: **for** *j starting from leaf nodes and iterating towards bus 0* **do**

$$x_j(t) \leftarrow f_j\left(x_{\mathsf{T}_j^\circ}(t);\ y(t-1)\right), \qquad j \in \overline{N}$$

   3. *Forward sweep*: **for** *j starting from children of bus 0 and iterating towards leaf nodes* **do**

$$y_j(t) \leftarrow g_j\left(x(t);\ y_{\mathsf{P}_j^\circ}(t)\right), \qquad j \in N$$

3. **Return**: $x := x(t)$, $y := y(t)$.

2 If (f_j, g_j) in (5.27a) depend not only on $(x_{\mathsf{T}_j^\circ}, y_{\mathsf{P}_j^\circ})$, but also on (x_j, y_j), then the update functions (f_j, g_j) in (5.28) become:

$$\begin{aligned} x_j(t) &= f_j\left(x_{\mathsf{T}_j^\circ}(t), x_j(t-1);\ y(t-1)\right), && j \in \overline{N} \\ y_j(t) &:= g_j\left(x(t);\ y_{\mathsf{P}_j^\circ}(t), y_j(t-1)\right), && j \in \overline{N} \end{aligned}$$

i.e., f_j only needs its own state and the state x_k at its child nodes, but not at upstream nodes and similarly for g_j.

3 In most applications, T_j° contains only the children of j and P_j° contains only the parent of j. □

In the next two subsections we illustrate this general BFS formulation using the complex form BFM (5.20) of Chapter 5.2.1 and the DistFlow model (5.9) of Chapter 5.1.3. The convergence of these two BFS algorithms will be analyzed in Chapter 8.6 as applications of general convergence analysis of iterative algorithms for solving systems of equations. These equations often arise as optimality conditions (e.g. the KKT condition) and we will therefore postpone the convergence analysis of iterative algorithms to after we have introduced a basic theory of optimization.

5.3.2 Complex form BFM

We consider the complex form BFM (5.20) of Chapter 5.2.1 but assume that the network graph $G := (\overline{N}, E)$ is radial and C5.1 holds ($y^s_{jk} = y^s_{kj}$). We can then adopt a directed graph G and need to involve line variables such as (I_{jk}, S_{jk}) only in the direction of the line $j \to k$, but not variables (I_{kj}, S_{kj}) in the opposite direction, as explained at the beginning of Chapter 5.1.3. Without loss of generality we assume the down orientation where each line points *away* from the root (and reference) bus 0.

With these assumptions the complex form BFM (5.20) reduces to

$$s_j \;=\; \sum_{k:j\sim k} V_j\, I^{\mathsf{H}}_{jk}, \qquad I_{jk} \;=\; \tilde{y}_{jk} V_j - y^s_{jk} V_k \tag{5.29}$$

Suppose V_0 and injections s_j at all non-reference buses $j \neq 0$ are given. To solve (5.29) for $(s_0, V_j, I_{jk}, j \in N, j \to k \in E)$, instead of I_{jk}, we will first compute the currents I^s_{jk} through the series admittances y^s_{jk} :

$$I^s_{jk} := I_{jk} \;-\; y^m_{jk} V_j$$

as well as V_j. All other variables in (5.20), such as the injection s_0 and the sending-end branch flows (I_{jk}, S_{jk}), can be computed once (V_j, I^s_{jk}) for all $j \in \overline{N}$ and all $j \to k \in E$ are determined. Instead of I^s_{jk}, we can also design a BFS algorithm that computes the branch current I_{jk} directly (Exercise 5.6).

To this end we will choose two sets of power flow equations in (V_j, I^s_{jk}) that are spatially recursive. For each bus j, let $i(j)$ denote the parent of bus j (i.e., $i := i(j)$ is the bus adjacent to j on the unique path from bus 0 to j). By Ohm's law we have $V_j - V_k = z^s_{jk} I^s_{jk}$ where $z^s_{jk} := 1/y^s_{jk}$ is the series impedance of line (j,k). Under assumption C5.1, the receiving-end current at bus j from $i := i(j)$ is $I^s_{ij} - y^m_{ji} V_j$.[3] The current injection at bus j is $(s_j/V_j)^{\mathsf{H}}$. Hence KCL at each non-reference bus j is (see Figure 5.6)

$$\left(\frac{s_j}{V_j}\right)^{\mathsf{H}} + \left(I^s_{ij} - y^m_{ji} V_j\right) \;=\; \sum_{k:j\to k} \left(I^s_{jk} + y^m_{jk} V_j\right), \qquad j \in N$$

This is the basis for the BFS algorithm of [30] which adopts the power flow equations:

$$I^s_{ij} \;=\; \sum_{k:j\to k} I^s_{jk} - \left(\left(\frac{s_j}{V_j}\right)^{\mathsf{H}} - y^m_{jj} V_j\right) \;=:\; f_j, \qquad j \in N \tag{5.30a}$$

$$V_j \;=\; V_i \;-\; z^s_{ij} I^s_{ij} \;=:\; g_j, \qquad j \in N \tag{5.30b}$$

where $i := i(j)$ denotes the unique parent of j and $y^m_{jj} := y^m_{ji} + \sum_{k:j\to k} y^m_{jk}$ is the total shunt admittance incident on bus j. The boundary conditions are

$$\{k : j \to k\} := \emptyset \text{ for leaf nodes } j, \quad V_0 \text{ is given}, \quad V_j(0) := V_0,\; j \in \overline{N} \tag{5.30c}$$

[3] Note that the received power at bus j from $i(j)$ is $V_j \left(I^s_{ij} - y^m_{ji} V_j\right)^{\mathsf{H}}$, not $V_i \left(I^s_{ij} - y^m_{ji} V_j\right)^{\mathsf{H}}$.

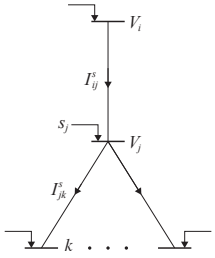


Figure 5.6 Spatially recursive structure of power flow equations (5.30).

This defines the partitioning (x, y) and the update functions (f, g) in (5.27) (recall that the injections s_j at all non-reference buses j are given):

- $x_j := I^s_{ij}$ for $j \in N$ are the complex line currents across the series impedance z^s_{ij} from buses i to j. The backward sweep functions f_j are given by (5.30a). Let $x := \left(I^s_{i(j)j}, j \in N\right) = (I^s_{jk}, j \to k \in E)$ and $f := (f_j, j \in N)$.
- $y_j := V_j$ for $j \in N$ are the complex voltages at buses j. The forward sweep functions g_j are given by (5.30b). Let $y := \left(V_j, j \in N\right)$ and $g := (g_j, j \in N)$.
- The initialization is given by (5.30c).

The update function f is linear in x given y, but not jointly linear in (x, y). The function g is linear in (x, y).

The functions (f, g) are spatially recursive because f_j depends on $x := \left(I^s_{i(j)j}, j \in N\right)$ only through $x_{\mathsf{T}^\circ_j}$ and g_j depends on $y := \left(V_j, j \in N\right)$ only through $y_{\mathsf{P}^\circ_j}$. This translates automatically into a BFS algorithm defined by the inner loops (5.28) and Algorithm 1. Given voltages $y(t-1)$, propagating (5.30a) backward from the leaf nodes towards the root (bus 0) in the reverse breadth-first search order, the current $I^s_{ij}(t)$ can be updated once all the currents $I^s_{jk}(t)$ in the previous level have been determined; see Figure 5.6. In the forward direction, given currents $x(t)$, propagating (5.30b) from the root towards the leaf nodes, the voltage $V_j(t)$ can be updated once its parent $V_i(t)$ has been determined. The detailed instantiation of Algorithm 1 for (5.30) is given in Algorithm 2. A stopping criterion for Algorithm 2 can be based on the discrepancy between the given injections s_j and the injections $s_j(t)$ implied by $x(t) := (I^s_{jk}(t), j \to k \in E)$ and $y(t) := (V_j(t), j \in N)$ at the end of each outer iteration t. Motivated by (5.30a), let

$$s_j(t) \;:=\; V_j(t)\left(\sum_{k:j\to k} I^s_{jk}(t) - I^s_{ij}(t)\right)^{\mathsf{H}} + y^{m\mathsf{H}}_{jj}\left|V_j(t)\right|^2, \qquad j \in N$$

Then a stopping criterion can be

$$\|s(t) - s\|_2 \;:=\; \sum_{j\in N}\left(s_j(t) - s_j\right)^2 < \epsilon$$

Algorithm 2: BFS for (5.30)

Input: voltage V_0 and injections $(s_i, i \in N)$.

Output: currents $x := (I^s_{jk}, j \to k \in E)$ and voltages $y := \left(V_j, j \in N\right)$ that are a solution of (5.30).

1. **Initialization**:
   - $\{k : j \to k\} := \emptyset$ for leaf nodes j
   - $V_j(0) := V_0$ at all buses $j \in N$.
   - $V_0(t) := V_0$ at bus $j = 0$ for all $t = 0, 1, \ldots$.
2. **while** *stopping criterion not met (see below)* **do**
   1. $t \leftarrow t+1$;
   2. *Backward sweep*: **for** j *starting from leaf nodes and iterating towards bus 0* **do**

$$I^s_{ij}(t) \leftarrow \sum_{k:j\to k} I^s_{jk}(t) \; - \left(\left(\frac{s_j}{V_j(t-1)}\right)^{\mathsf{H}} - y^m_{jj} V_j(t-1)\right), \quad i \to j \in E$$

      where $y^m_{jj} := y^m_{ji} + \sum_{k:j\sim k} y^m_{jk}$ and $i := i(j)$ is the unique parent of j.

   3. *Forward sweep*: **for** j *starting from children of bus 0 and iterating towards leaf nodes* **do**

$$V_j(t) = V_i(t) \; - z^s_{ij} I^s_{ij}(t), \qquad j \in N$$

      where $z^s_{ij} := \left(y^s_{ij}\right)^{-1}$ and $i := i(j)$ is the unique parent of j.

3. **Return**: $x := \left(I^s_{jk}(t), j \to k \in E\right)$, $y := \left(V_j(t), j \in N\right)$.

for a given tolerance $\epsilon > 0$.

The convergence of Algorithm 2 is analyzed in Example 8.18 of Chapter 8.6.2 as a Gauss-Seidel algorithm.

5.3.3 DistFlow model

The BFS algorithm defined by (5.30) assumes all power injections s_j at non-reference buses j are given and computes I^s_{jk} in the backward sweep. If some buses have their voltage magnitudes $|V_j|$ and real power p_j given instead (i.e., these are PV buses), we can develop BFS algorithms based on the DistFlow model of Chapter 5.1.3. The advantage of the DistFlow model is that the BFS algorithms need not compute the voltage angles θ_j. Phase angles can be recovered using (5.12) after BFS has produced a solution. As in Chapter 5.1.3, we assume $z^s_{jk} = z^s_{kj}$ (assumption C5.1) and $y^m_{jk} = y^m_{kj} = 0$.

We will present two algorithms, one where V_0 and $(s_j, j \in N)$ are given, as in Chapter 5.3.2, and the other where $(V_0, v_j, j \in N)$ and $(p_j, j \in N)$ are given. In both

cases only v_0 is needed in BFS but the angle $\angle V_0$ ensure a unique angle vector θ in (5.12) from the solution of BFS. It will be convenient to adopt a graph orientation where every line $k \to j$ points *towards* the root bus 0.

Example 5.4 (Given (V_0, s_j))**.** Suppose the complex voltage V_0 and $(s_j, j \in N)$ for all non-reference buses j are given. We will use the DistFlow equation (5.9) for the up orientation to compute $\big(S_{kj}, \ell_{kj}, k \to j \in E\big)$ and $(v_j, j \in N)$.

The equations (5.9a) and (5.9c) lead to the following backward sweep to compute $\big(S_{kj}, \ell_{kj}, k \to j\big)$:

$$S_{ji} \;=\; s_j \;+\; \sum_{k:k\to j} \Big(S_{kj} - z^s_{kj}\ell_{kj}\Big), \qquad j \in N \tag{5.31a}$$

$$\ell_{ji} \;=\; \frac{|S_{ji}|^2}{v_j}, \qquad j \in N \tag{5.31b}$$

where $i := i(j)$ in (5.31a) denotes the parent node of j on the unique path between node 0 and node j. The equation (5.9b) leads to a forward sweep to compute $(v_j, j \in N)$:

$$v_j \;=\; v_i \;+\; 2\mathrm{Re}\Big(z^{s\mathsf{H}}_{ji} S_{ji}\Big) - |z^s_{ji}|^2 \ell_{ji}, \qquad j \in N \tag{5.31c}$$

The boundary conditions are

$$\{k : k \to j\} := \emptyset \text{ for leaf nodes } j, \quad V_0 \text{ given}, \quad v_j(0) := |V_0|^2,\ j \in \overline{N} \tag{5.31d}$$

This defines the partitioning (x, y) and the update functions (f, g) in (5.27):

- $x := \big(S_{ji(j)}, \ell_{ji(j)}, j \in N\big)$. The backward sweep functions $f := (f_j, j \in N)$ are given by (5.31a)(5.31b).
- $y := (v_j, j \in N)$. The forward sweep functions $g := (g_j, j \in N)$ are given by (5.31c).
- The initialization is given by (5.31d).

The update function f is linear in x given y, but not jointly linear in (x, y). The function g is linear in (x, y). Since (f, g) are spatially recursive, (5.31) translates automatically into a BFS algorithm defined by the inner loops (5.28); see Algorithm 1. □

Example 5.5 (Given (v_j, p_j))**.** Suppose the complex voltage V_0, squared voltage magnitudes $(v_j, j \in N)$ and real power injections $(p_j, j \in N)$ for all non-reference buses j are given. We will compute the reactive power injections $(q_j, j \in N)$ as well as the line flows $(S_{ji}, j \to i \in E)$. All other variables can then be determined.

Eliminating ℓ_{kj} from (5.31a)(5.31b) we can compute $S_{ji} := \big(P_{ji}, Q_{ji}\big)$ in a backward sweep and q_j in a forward sweep:

$$S_{ji} \;=\; s_j \;+\; \sum_{k:k\to j} \left(S_{kj} - z^s_{kj}\frac{|S_{kj}|^2}{v_k}\right), \qquad j \in N \tag{5.32a}$$

$$q_j \;=\; Q_{ji} \;-\; \sum_{k:k\to j} \left(Q_{kj} - x^s_{kj}\frac{|S_{kj}|^2}{v_k}\right), \qquad j \in N \tag{5.32b}$$

where $z^s_{kj} =: r^s_{kj} + \mathbf{i}x^s_{kj}$. The boundary conditions are

$$\{k : k \to j\} := \emptyset \text{ for leaf nodes } j, \quad v_j \text{ given } j \in \overline{N}, \quad q_j(0) \text{ given } j \in N \tag{5.32c}$$

This defines the partitioning (x, y) and the update functions (f, g) in (5.27):

- $x := \big(S_{ji(j)}, j \in N\big)$. The backward sweep functions $f := (f_j, j \in N)$ are given by (5.32a).
- $y := (q_j, j \in N)$. The forward sweep functions $g := (g_j, j \in N)$ are given by (5.32b).
- The initialization is given by (5.32c).

Both functions f and g are nonlinear in x (f is linear in and g is independent of y). Since the functions (f, g) are spatially recursive, (5.32) translates automatically into a BFS algorithm defined by the inner loops (5.28); see Algorithm 1. □

5.4 Linear power flow models

We now present linear approximations of BFM for radial networks when the line losses $z^s_{jk}\ell_{jk}$ are small compared with the line flows S_{jk}. The linear models have two advantages. Given injections s, the voltages v^{lin}_j and line flows S^{lin}_{jk} of the linearized model can be solved explicitly in terms of s. Moreover the linear solution $(v^{\text{lin}}, S^{\text{lin}})$ provides bounds on line flows S and voltages v of nonlinear branch flow models (5.8)(5.9).

5.4.1 With shunt admittances

Recall the general branch flow model (5.1) in Chapter 5.1.2 for a radial network with $N+1$ buses and M lines where shunt admittances (y^m_{jk}, y^m_{kj}) may be nonzero and y^s_{jk} and y^s_{kj} may be unequal (i.e., assumption C5.1 may not hold). A linear approximation is the following model obtained from (5.1) by setting $\ell_{jk} = \ell_{kj} = 0$ in (5.1):

$$s_j = \sum_{k:j\sim k} S_{jk}, \qquad j \in \overline{N} \tag{5.33a}$$

$$|\alpha_{jk}|^2 v_j - v_k = 2\,\mathrm{Re}\left(\alpha_{jk}\,\bar{z}^s_{jk} S_{jk}\right), \qquad (j,k) \in E \tag{5.33b}$$

$$|\alpha_{kj}|^2 v_k - v_j = 2\,\mathrm{Re}\left(\alpha_{kj}\,\bar{z}^s_{kj} S_{kj}\right), \qquad (j,k) \in E \tag{5.33c}$$

$$\bar{\alpha}_{jk} v_j \; - \; \bar{z}^s_{jk} S_{jk} = \left(\bar{\alpha}_{kj} v_k \; - \; \bar{z}^s_{kj} S_{kj}\right)^{\mathsf{H}}, \qquad (j,k) \in E \tag{5.33d}$$

It is a set of $2(N+1)+4M = 6N+2$ linear real equations in $3(N+1)+4M = 7N+3$ real variables $x := (s_j, v_j, S_{jk}, S_{kj}, j \in \overline{N}, (j,k) \in E)$. Given $2N+1$ variables, e.g., $(v_0, p_j, q_j, j \in N)$, the linear power flow problem solves the remaining $5N+2$ variables

from the set of $6N+2$ linear equations (5.33). Even though there are more equations than variables these equations are typically linearly dependent, as the next example shows.

Example 5.6 (Two buses connected by a transformer)**.** For the two-bus network in Example 5.1, $S_{jk} = s_j$ and $S_{kj} = s_k$. Hence the linear approximation (5.33) is a set of 4 equations in 6 variables (s, v):

$$\begin{aligned} v_j - v_k \Big/ |K|^2 &= 2\,\mathrm{Re}\left((\tilde{z}^s)^{\mathsf{H}} s_j\right) \\ |\tilde{\alpha}/K|^2 v_k - v_j &= 2\,\mathrm{Re}\left(\tilde{\alpha}\,(\tilde{z}^s)^{\mathsf{H}} s_k\right) \\ v_j - (\tilde{z}^s)^{\mathsf{H}} s_j &= \left(\tilde{\alpha}/|K|^2\right) v_k - \tilde{z}^s \bar{s}_k \end{aligned}$$

where K is the voltage gain (possibly complex), $\tilde{\alpha} := (1 + \tilde{z}^s \tilde{y}^m)$ and $\tilde{z}^s$ and $\tilde{y}^m$ are the leakage and shunt admittance of the transformer. Let $\tilde{r} + \mathbf{i}\tilde{x} := \tilde{z}^s$ denote the resistance and reactance of the leakage impedance of the transformer. Then this system of linear equations can be written as

$$\begin{bmatrix} 1 & 1/|K|^2 \\ -1 & |\tilde{\alpha}/K|^2 \\ 1 & -\mathrm{Re}(\tilde{\alpha})/|K|^2 \\ 0 & -\mathrm{Im}(\tilde{\alpha})/|K|^2 \end{bmatrix} \begin{bmatrix} v_j \\ v_k \end{bmatrix} = \begin{bmatrix} 2\tilde{r} & 2\tilde{x} & 0 & 0 \\ 0 & 0 & 2\mathrm{Re}\left(\tilde{\alpha}^{\mathsf{H}} \tilde{z}^s\right) & 2\mathrm{Im}\left(\tilde{\alpha}^{\mathsf{H}} \tilde{z}^s\right) \\ \tilde{r} & \tilde{x} & -\tilde{r} & -\tilde{x} \\ -\tilde{x} & \tilde{r} & -\tilde{x} & \tilde{r} \end{bmatrix} \begin{bmatrix} p_j \\ q_j \\ p_k \\ q_k \end{bmatrix}$$

We now demonstrate that the system of linear equations are typically linearly dependent.

Suppose $\tilde{y}^m = 0$ so that $\tilde{\alpha} = 1$. Suppose further that (p_k, q_k, v_j) are given and we are to solve (p_j, q_j, v_k). Then (p_j, q_j, v_k) satisfies four equations (only three of which are linearly independent):

$$\underbrace{\begin{bmatrix} 2\tilde{r} & 2\tilde{x} & -1/|K|^2 \\ 0 & 0 & 1/|K|^2 \\ \tilde{r} & \tilde{x} & 1/|K|^2 \\ -\tilde{x} & \tilde{r} & 1/|K|^2 \end{bmatrix}}_{A} \begin{bmatrix} p_j \\ q_j \\ v_k \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 2\tilde{r} & 2\tilde{x} & 1 \\ \tilde{r} & \tilde{x} & 1 \\ \tilde{x} & -\tilde{r} & 0 \end{bmatrix} \begin{bmatrix} p_k \\ q_k \\ v_j \end{bmatrix}$$

Elementary row operation reduces the matrix A to the following rank-3 matrix:

$$\begin{bmatrix} (\tilde{r}/\tilde{x})(\tilde{r}^2 + \tilde{x}^2) & 0 & 0 \\ 0 & \tilde{r}^2 + \tilde{x}^2 & 0 \\ 0 & 0 & 1/|K|^2 \\ 0 & 0 & 0 \end{bmatrix}$$

□

5.4.2 Without shunt admittances

Suppose $y^s_{jk} = y^s_{kj}$ (assumption C5.1) and $y^m_{jk} = y^m_{kj} = 0$. Then we can consider a directed graph with an arbitrary orientation. To simplify notation we sometimes omit the superscript and write y_{jk} and z_{jk} for y^s_{jk} and z^s_{jk} respectively.

The linear approximation from [25] is obtained by setting $\ell_{jk} := 0$ in the DistFlow equation (5.3) of Chapter 5.1.3:

$$\sum_{k:j\to k} S_{jk} = \sum_{i:i\to j} S_{ij} + s_j, \qquad j \in \overline{N} \tag{5.34a}$$

$$v_j - v_k = 2\,\mathrm{Re}\left(\bar{z}^s_{jk} S_{jk}\right), \qquad j \to k \in E \tag{5.34b}$$

The linear model (5.34) can also be derived from (5.33) by setting $\alpha_{jk} = \alpha_{kj} = 1$ (i.e., $y^m_{jk} = y^m_{kj} = 0$) and $y^s_{jk} = y^s_{kj}$ in (5.33) so that $S_{kj} = -S_{jk}$ and the cycle condition (5.33d) becomes (5.34b). We can also write (5.34a)(5.34b) in vector form in terms of the bus-by-line incidence matrix C defined in (5.4). Let $D_r := \mathrm{diag}(r_l, l \in E) \succ 0$ and $D_x := \mathrm{diag}(x_l, l \in E) \succ 0$ be the $N \times N$ positive definite diagonal matrices of line resistances and reactances. Let $s := (s_j, j \in \overline{N})$, $v := (v_j, j \in \overline{N})$ and $S := (S_l, l \in E)$. Then the linear model in vector form is:

$$s \;=\; C\,S, \qquad C^{\mathsf{T}} v \;=\; 2\,(D_r P + D_x Q) \tag{5.34c}$$

The matrix C is of rank N since the graph is connected, i.e., its columns are linearly independent. The null space of C^{T} is span($\mathbf{1}$). Any $N \times N$ submatrix of C obtained by removing any row of C is invertible (Theorem A.36 of Appendix A.11).

5.4.3 Linear solution and its properties

Suppose the reference bus voltage v_0 and the injections $\hat{s} := (s_j, j \in N)$ at other buses are given.

5.4.3.1 Linear solution

The linear model (5.34) can be solved explicitly for non-reference bus voltages $\hat{v} := (v_j, j \in N)$ and line flows S, from which s_0 can also be determined. Recall the decomposition in (4.22a) of the incidence matrix C into the row c_0^{T} corresponding to bus 0 and the remaining reduced incidence matrix $\hat{C}$, reproduced here:

$$C =: \begin{bmatrix} c_0^{\mathsf{T}} \\ \hat{C} \end{bmatrix}$$

Then the linear model (5.34) when $(v_0, p_j, q_j, j \in N)$ are given is:

$$\hat{s} = \hat{C}\,S, \quad s_0 \;=\; c_0^{\mathsf{T}} S \tag{5.35a}$$

$$v_0\, c_0 \;+\; \hat{C}^{\mathsf{T}} \hat{v} = 2\,(D_r P \;+\; D_x Q) \tag{5.35b}$$

Let P_j denote the unique path from bus 0 to bus j, including both buses 0 and j. We use "$l \in \mathsf{P}_j$" to refer to a directed line l in the path P_j that points away from bus 0 and "$-l \in \mathsf{P}_j$" to refer to a directed line l in P_j that points towards bus 0. Theorem 4.10 shows that the reduced incidence matrix $\hat{C}$ is nonsingular and

$$\left[\hat{C}^{-1}\right]_{lj} = \begin{cases} -1 & l \in \mathsf{P}_j \\ 1 & -l \in \mathsf{P}_j \\ 0 & \text{otherwise} \end{cases}, \qquad \hat{C}^{-\mathsf{T}} c_0 = -\mathbf{1}$$

where $\hat{C}^{-\mathsf{T}} := \left(\hat{C}^{\mathsf{T}}\right)^{-1}$. Then (5.35) can be solved using Theorem 4.10.

Theorem 5.3 (Linear solution)**.** Suppose the network graph G is a (connected) tree, assumption C5.1 holds and $y^m_{jk} = y^m_{kj} = 0$. Fix any v_0 and $\hat{s} = (\hat{p}, \hat{q}) \in \mathbb{R}^{2N}$. Then

1 The solution to (5.35) is

$$S = \hat{C}^{-1}\hat{s}, \quad s_0 = c_0^{\mathsf{T}}\hat{C}^{-1}\hat{s} \tag{5.36a}$$

$$\hat{v} = v_0\mathbf{1} + 2(R\hat{p} + X\hat{q}) \tag{5.36b}$$

where $R := \hat{C}^{-\mathsf{T}} D_r \hat{C}^{-1}$ and $X := \hat{C}^{-\mathsf{T}} D_x \hat{C}^{-1}$.

2 $R \succ 0$ and $X \succ 0$ are positive definite matrices and

$$R_{jk} = \sum_{l \in \mathsf{P}_j \cap \mathsf{P}_k} r_l, \qquad X_{jk} = \sum_{l \in \mathsf{P}_j \cap \mathsf{P}_k} x_l \tag{5.36c}$$

The solution (5.36a)(5.36b) can be obtained by multiplying both sides of (5.35) by $\hat{C}^{-1}$. The positive definiteness of R and X follows from $D_r \succ 0$ and $D_x \succ 0$. The explicit expressions in (5.36c) follow from Theorem 4.10 and have a simple interpretation: the (j,k) entries of R and X are the total resistance and reactance respectively in the common segment of the paths from bus 0 to buses j and k. If we interpret $\hat{L} := \hat{C}(D_r^{-1})\hat{C}^{\mathsf{T}}$ as a reduced Laplacian matrix, then $R = \hat{L}^{-1}$ (similarly for X).

5.4.3.2 Analytical properties

We now study some analytical properties of the linear model (5.34). These properties hold for general graph orientations but are particularly transparent in two special orientations.

Down orientation: lines point away from bus 0.

The linear model (5.34) reduces to:

$$\sum_{k:j\to k} S^{\text{lin}}_{jk} = S^{\text{lin}}_{ij} + s_j, \quad j \in \overline{N} \tag{5.37a}$$

$$v^{\text{lin}}_j - v^{\text{lin}}_k = 2\,\text{Re}\left(z^{\mathsf{H}}_{jk} S^{\text{lin}}_{jk}\right), \quad j \to k \in E \tag{5.37b}$$

where bus $i := i(j)$ in (5.37a) denotes the bus adjacent to j on the unique path from bus 0 to bus j. The boundary condition is: $S^{\text{lin}}_{i0} := 0$ in (5.37a) when $j = 0$ and $S^{\text{lin}}_{jk} = 0$ in (5.37a) when j is a leaf node.

Up orientation: lines point towards bus 0.

The linear model (5.34) reduces to:

$$\overline{S}^{\text{lin}}_{ji} = \sum_{k:k\to j} \overline{S}^{\text{lin}}_{kj} + s_j, \quad j \in \overline{N} \tag{5.38a}$$

$$\overline{v}^{\text{lin}}_k - \overline{v}^{\text{lin}}_j = 2\,\text{Re}\left(z^{\mathsf{H}}_{kj}\overline{S}^{\text{lin}}_{kj}\right), \; k \to j \in E \tag{5.38b}$$

where $i := i(j)$ in (5.38a) denotes the node adjacent to j on the unique path between node 0 and node j. The boundary condition is defined by $\overline{S}^{\text{lin}}_{ji} = 0$ in (5.38a) when $j = 0$ and $\overline{S}^{\text{lin}}_{kj} = 0, \ell_{kj} = 0$ in (5.38a) when j is a leaf node.

Denote by T_j the subtree rooted at bus j, including j. We write "$l \in \mathsf{T}_j$" to mean either a bus l or a line l in the subtree T_j, depending on the context. If there is danger of confusion we write "$(j,k) \in \mathsf{T}_j$" to mean line $l := (j,k)$ in T_j. The following corollary is proved in Exercise 5.8.

Corollary 5.4 (Linear solutions)**.** Under the assumptions of Theorem 5.3 let $(v^{\text{lin}}, S^{\text{lin}}) \in \mathbb{R}^{N+2M}$ be the solution of (5.37) and $(\overline{v}^{\text{lin}}, \overline{S}^{\text{lin}}) \in \mathbb{R}^{N+2M}$ the solution of (5.38). Then

1 For $(i,j) \in E$

$$S^{\text{lin}}_{ij} = -\sum_{k\in\mathsf{T}_j} s_k, \qquad i \to j$$

$$\overline{S}^{\text{lin}}_{ji} = \sum_{k\in\mathsf{T}_j} s_k, \qquad j \to i$$

Hence $S^{\text{lin}}_{ij} = -\overline{S}^{\text{lin}}_{ji}$.

2 For $j \in \overline{N}$, $v^{\text{lin}}_j = \overline{v}^{\text{lin}}_j = v_0 + 2\sum_k \left(R_{jk}p_k + X_{jk}q_k\right)$ where R_{jk} and X_{jk} are given in (5.36c).

Corollary 5.4 says that, on each line $(i,j) \in E$, the power flow S_{ij} from i to j, or the power flow $\overline{S}_{ji}$ in the opposite direction, equals the total load $-\sum_{k\in\mathsf{T}_j} s_k$ in the subtree rooted at node j. These linear line flows neglect line losses and underestimate the required power to supply these loads. With zero line loss, we have $S^{\text{lin}}_{ij} = -\overline{S}^{\text{lin}}_{ji}$. Since all entries of R and X are nonnegative, both real and reactive power injections (p,q) always increase voltage magnitudes v according to the linear approximation.

This is not the case for solutions of nonlinear power flow equations (5.8) or (5.9). Indeed fix any v_0 and injections $\hat{s} \in \mathbb{R}^{2N}$ at non-reference buses in N. We can recurse

on the power flow equations (5.8), starting from the leaf nodes for S_{ij} and bus 0 for v_j, to show that any solution (v, ℓ, S) of (5.8) must satisfy (Exercise 5.9):

$$S_{ij} = -\sum_{k \in \mathsf{T}_j} s_k + \left(z_{ij}\ell_{ij} + \sum_{l \in \mathsf{T}_j} z_l \ell_l \right) \tag{5.39a}$$

$$v_j = v_0 - \sum_{l \in \mathsf{P}_j} \left(2\mathrm{Re}\left(z_l^{\mathsf{H}} S_l \right) - |z_l|^2 \ell_l \right) \tag{5.39b}$$

Similarly we can recurse on (5.9) to show that

$$\overline{S}_{ji} = \sum_{k \in \mathsf{T}_j} s_k - \sum_{l \in \mathsf{T}_j} z_l \overline{\ell}_l \tag{5.39c}$$

$$\overline{v}_j = v_0 + \sum_{l \in \mathsf{P}_j} \left(2\mathrm{Re}\left(z_l^{\mathsf{H}} \overline{S}_l \right) - |z_l|^2 \overline{\ell}_l \right) \tag{5.39d}$$

Summing (5.39a) and (5.39c) shows that

$$S_{ij} + \overline{S}_{ji} = z_{ij}\ell_{ij}$$

as we saw earlier in (5.2). Note that given v_0 and $s \in \mathbb{R}^{2N}$, Corollary 5.4 provides the unique solution $(v^{\mathrm{lin}}, S^{\mathrm{lin}})$ to (5.37) (or unique solution $(\overline{v}^{\mathrm{lin}}, \overline{S}^{\mathrm{lin}})$ to (5.38)). For nonlinear model (5.8) or (5.9), the solutions (v, ℓ, S) or $(\overline{v}, \overline{s}, \overline{S})$ may not be unique. Any nonlinear solution however must satisfy (5.39).

It is proved in Exercise 5.8 that, for $j \in N$, the linear solutions satisfy:

$$v_j^{\mathrm{lin}} = v_0 - \sum_{l \in \mathsf{P}_j} 2\mathrm{Re}\left(z_l^{\mathsf{H}} S_l^{\mathrm{lin}} \right) \tag{5.40a}$$

$$\overline{v}_j^{\mathrm{lin}} = v_0 + \sum_{l \in \mathsf{P}_j} 2\mathrm{Re}\left(z_l^{\mathsf{H}} \overline{S}_l^{\mathrm{lin}} \right) \tag{5.40b}$$

Comparing these relations and (5.39) leads to bounds on the nonlinear solutions in the following corollary (proved in Exercise 5.10). Recall that, by definition, x is a power flow solution only if $v \geq 0$ and $\ell \geq 0$ componentwise (assuming $z_l = (r_l, x_l) > 0$ for any line $l \in E$.).

Corollary 5.5 (Bounds on nonlinear solutions)**.** Suppose the network graph G is a (connected) tree, assumption C5.1 holds and $y_{jk}^m = y_{kj}^m = 0$. Fix any v_0 and $\hat{s} \in \mathbb{R}^{2N}$. Let (v, ℓ, S) and $(\overline{v}, \overline{\ell}, \overline{S})$ in $\mathbb{R}^{N+3M}$ be any (possibly nonunique) solutions of (5.8) and (5.9) respectively. Let $(v^{\mathrm{lin}}, S^{\mathrm{lin}})$ and $(\overline{v}^{\mathrm{lin}}, \overline{S}^{\mathrm{lin}})$ in R^{N+2M} be the unique solutions of their linearizations (5.37) and (5.38) respectively. Then

1 For $i \to j \in E$, $S_{ij} \geq S_{ij}^{\mathrm{lin}}$ with equality if only if ℓ_{ij} and all ℓ_{kl} in T_j are zero.
2 For $j \to i \in E$, $\overline{S}_{ji} \leq \overline{S}_{ji}^{\mathrm{lin}}$ with equality if and only if all ℓ_{kl} in T_j are zero.
3 For $j \in \overline{N}$, $v_j = \overline{v}_j \leq \overline{v}_j^{\mathrm{lin}} = v_j^{\mathrm{lin}}$.

Remark 5.2. 1 *Up orientation*. While it is easy to prove $\overline{v}_j \le \overline{v}_j^{\text{lin}}$ from (5.39d) and (5.40b), it does not seem easy to prove $v_j \le v_j^{\text{lin}}$ directly, except by relating the variables (v_j, v_j^{lin}) to $(\overline{v}_j, \overline{v}_j^{\text{lin}})$ in the opposite direction. This is an advantage of the models (5.9) and (5.38) in the up orientation.

2 *Bounds for SOCP relaxation*. The bounds in Corollary 5.5 do not depend on the quadratic equalities (5.8c) and (5.9c) as long as $\ell_{jk} \ge 0$. In particular the bounds hold if the equalities are relaxed to inequalities $v_j \ell_{jk} \ge |S_{jk}|^2$. These bounds are used in Chapter **??** in a sufficient condition for exact SOCP relaxation of optimal power flow problems for radial networks.

3 *Linear approximation*. For radial networks, the linear approximation (5.34) of BFM has two advantages over the (linear) DC approximation of BIM studied in Chapter 4.6.2. First the linear models (5.37) and (5.38) with special graph orientations have a recursive structure that leads to simple bounds on power flow quantities. Second DC approximation assumes $r_{jk} = 0$, fixes voltage magnitudes, and ignores reactive power, whereas (5.34) does not. This is important for distribution systems where r_{jk} are not negligible, voltages can fluctuate significantly and reactive powers are used to regulate them. On the other hand (5.34), (5.37) and (5.38) are applicable only to radial networks whereas DC approximation applies to meshed networks as well. □

5.5 Bibliographical notes

A branch flow model, called the DistFlow equations, is proposed in [24, 25] for radial networks. Its key feature is that it does not involve phase angles of voltage and current phasors. This is extended to general meshed network in [31] by introducing a cycle condition. All of these models assume zero shunt admittances on the lines. Shunt admittances of the lines are added to the branch flow model in [32]. The main difference of the model (5.1) from the model in [24, 25, 31] is the use of undirected rather than directed graph when shunt elements are included so that line currents and power flows are defined in both directions. The equivalence of BFM and bus injection model (BIM) is proved in [33]. The equivalence of DistFlow to BFM in complex form and hence equivalent to BIM follows from [31, Theorems 2, 4]. Theorem 5.1 is from [34]. For BFM and SOCP relaxations when a radial network contain ideal transformers and multiple lines between two buses, see [35].

The linearized model (5.34) is first proposed in [25] and called the Simplified DistFlow equations. The paper also states an explicit solution for the squared voltage magnitude v_i as an affine function of the injections s_j whose coefficients ξ_{ij} are the total impedances on the common paths P_i° and P_j° from the root (bus 0) to buses i and j respectively. This is the same solution as that in Theorem 5.3. The properties in Theorem 5.3 and Corollary 5.5 of the linear model seem to have been independently observed in several papers, e.g., [36, 37, 38, 39, 40] where $v_i - v_j$ is sometimes

approximated by $2(|V_i| - |V_j|)$ since $|V_i| \approx 1$ pu. Our discussion on the local volt/var control algorithm follows [37, 38].

Backward forward sweep (edit later).

Power flow solutions for general networks are mostly based on Newton-Raphson and its variants, or more recently, interior-point methods. Another approach has been developed for radial networks, both single-phase and three-phase networks, that exploits their tree structure. The idea of backward forward sweep (BFS) is first proposed in [41] for three-phase distribution systems. Early examples of BFS algorithms for three-phase radial networks are designed in [42][43, Chapter 10.1.3]. The BFS method for single-phase networks described in Chapter 5.3.2 is from [30]. It is extended in [44] to allow PV buses by computing line power flows S_{jk} instead of currents I^s_{jk}. Both algorithms (with extensions for meshed networks) were developed for weakly meshed transmission systems as well as distribution systems. Another variant of BFS, proposed in [**?**], calculates voltages in both forward and backward iterations in linear feeders with voltage-dependent loads. The BFS algorithm in [30] is extended in [45] from single-phase to three-phase networks, and in [**?**] to four-wire neutral-grounded networks. In [**?**], three-phase voltages and line currents are calculated with generalized line models that incorporate transformers and constant impedance loads. Transformers of different configurations have been included in BFS through modified augmented nodal analysis [**?**]. Some of these works are briefly discussed in [46]. BFS algorithms tend to have better convergence properties than general algorithms such as Newton-Raphson. Simulation results in [47] suggest however that Newton-Raphson converges in a smaller number of iterations.

The solution approach in the original DistFlow paper [25] uses one-time backward sweep to express all variables in terms of the power injections at the feeder head and all branch points followed by a Newton-Raphson algorithm to solve for these injections. The existence and uniqueness of solutions are studied in [48]. By exploiting the approximate sparsity of the Jacobian matrix in [25], approximate fast decoupled methods are developed and their convergence properties analyzed in [49]. These methods are extended to three-phase radial networks in [47]. The existence and uniqueness of power flow solutions of three-phase DistFlow model is analyzed in [50].

5.6 Problems

Chapter 5.1.

Exercise 5.1 (Line angles $\beta(x)$)**.** Justify the definition of line angles in (5.10) using (5.20b)(5.20c).

Exercise 5.2 (Incidence matrix C)**.** Consider the $(N+1)\times M$ incidence matrix C of a (connected) radial network defined by:

$$C_{jl} = \begin{cases} 1 & \text{if } l = j \to k \text{ for some bus } k \\ -1 & \text{if } l = i \to j \text{ for some bus } i \\ 0 & \text{otherwise} \end{cases}$$

Show that C has rank $N = M$, the null space of C^{T} is span$(\mathbf{1})$ and its pseudo-inverse $\left(C^{\mathsf{T}}\right)^{\dagger} = C\left(C^{\mathsf{T}}C\right)^{-1}$. (See Theorem 4.10 for the inverse of the reduced incidence matrix $\hat{C}$.)

Exercise 5.3 (Line loss)**.** Consider a radial network where lines have zero shunt admittances. Show that, under assumption C5.1, (5.1) leads to:

$$S_{jk} + S_{kj} = z^s_{jk}\ell_{jk} = z^s_{jk}\ell_{kj}$$

Exercise 5.4 (DistFlow equations)**.** Suppose assumption C5.1 holds and $y^m_{jk} = y^m_{kj} = 0$ for all lines $(j,k) \in E$. Show that $\mathbb{X}_{\text{tree}} \equiv \mathbb{X}_{\text{df}}$ (these sets are defined in Chapter 5.2.2)

Exercise 5.5 (Graph orientation)**.** Prove (5.7) under assumption C5.1, i.e., x satisfies (5.3) if and only if $\hat{x} := g(x)$ satisfies (5.6)

Chapter 5.3

Exercise 5.6 (Backward forward sweep)**.** The BFS algorithm in Chapter 5.3.2, based on the branch flow model (5.20) in complex form, computes (V_j, I^s_{jk}).

1. Show that all other variables in (5.20) can be computed once (V_j, I^s_{jk}) for all $j \in \overline{N}$ and all $j \to k \in E$ are determined.
2. Design a BFS algorithm that solves the same power flow equations under the same assumptions but computes the sending-end currents I_{jk} directly, instead of I^s_{jk} over the series admittances, as well as the voltage V_j.

Exercise 5.7 (Backward forward sweep)**.** Consider a 2-bus network and prove a sufficient condition for BFS to converge under assumption C5.1.

Chapter 5.4.

Assumption C5.1 and $y^m_{jk} = y^m_{kj} = 0$ are assumed in Chapter 5.4 for linear DistFlow models and hence for problems in this section.

Exercise 5.8 (Linear solution)**.** Prove Corollary 5.4. Also show that for $j \in \overline{N}$

$$v_j^{\text{lin}} = v_0 - \sum_{l \in \mathsf{P}_j} 2\,\text{Re}\left(z_l^{\mathsf{H}} S_l^{\text{lin}}\right) \tag{5.41a}$$

$$\overline{v}_j^{\text{lin}} = v_0 + \sum_{l \in \mathsf{P}_j} 2\,\text{Re}\left(z_l^{\mathsf{H}} \overline{S}_l^{\text{lin}}\right) \tag{5.41b}$$

Hence $v_j^{\text{lin}} = \overline{v}_j^{\text{lin}}$. (Hint: Use (5.36) or induction.)

Exercise 5.9 (Nonlinear recursion)**.** Derive (5.39) from the DistFlow equations (5.8) and (5.9). (Hint: Use induction.)

Exercise 5.10 (Bounds)**.** Prove Corollary 5.5.

6 System operation: power balance

The primary function of a power system is to deliver electricity reliably, and, subject to reliable operation, economically. In this and next chapters we explain some of the operational components using the network models developed in previous chapters. This chapter focuses on a hierarchy of control mechanisms at different timescales to balance power supply and demand. Chapter 7 presents applications in state estimation, voltage control on a distribution feeder, and network identification. Our focus is on mathematical analysis of these applications; see, e.g. [1, 2, 3], for detailed description of the physical systems and operations.

After some background information in Chapter 6.1 we describe in Chapter 6.2 the problem of unit commitment and real-time dispatch to balance power on daily and 5-15 minute basis respectively. In Chapter 6.3 we explain frequency control that balances power on a second by second basis. In Chapters 6.4 we study how to price electricity to incentivize optimal real-time dispatch.

6.1 Background

6.1.1 Overview

Electricity has two important differences from most commodities such as rice and minerals. First there is not yet large-scale energy storage in our power system so that inventory control as a means to match supply and demand for most commodities is not applicable. Instead generation and load must be balanced on a second-by-second basis at all points on the network. Second electricity cannot yet be routed from generators to loads at will but must follow paths determined power flow equations. The nonlinearity of power flow equations introduces computational challenges. These differences have strong implications on how the network is operated and how markets are organized.

The central control problem is to balance supply and demand, continuously and everywhere, without violating operational constraints such as capacity limits of generators and loads, bounds on voltage magnitudes, and thermal and stability limits of transmission lines and transformers. Thermal generators such as gas, coal and nuclear

generators still generate the majority of electricity today. For example, in 2023, fossil fuels generated 60.0% and nuclear generated 18.5% of all electricity in the US [51, Table 1.1]. Hydro-generation produced 5.9% of electricity and other renewable generations 15.5%. Thermal and hydro generators are fully controllable and can produce a specified amount of electricity at a specified time and location. Traditionally a power system operator forecasts demand, which is assumed inelastic, and schedules bulk generators to meet the forecast demand. As we decarbonize our energy system by replacing fossil fuel generators by wind and solar farms, our ability to control generation decreases and we must also exploit flexibility in demand to match volatile supply. Difficulties arise from the variability and uncertainty of undispatchable demand and supply, the need to match the speed of control and that of disturbances, as well as random unscheduled outages of generators, loads, lines and transformers.

A transmission network is a high-voltage long-distance network that connects bulk power producers to power consumers. These consumers are called load centers and represent aggregate loads such as substations of a local utility company that feeds a small city. The operation of a transmission network is typically coordinated by an independent system operator that commits and dispatches generation units to meet demand at timescales ranging from hours to minutes to seconds. Control and market operations are tightly integrated in a power system in order to balance supply and demand on a second-by-second basis everywhere in the grid.

An overview of the hierarchy of control mechanisms to balance power, as well as the associated pricing of electricity and reserves is as follows:

1 *Unit commitment and real-time dispatch* (Chapter 6.2). Bulk generators such as gas, coal, and nuclear generators need nontrivial amounts of time and cost to start up and shut down, e.g., the startup time for a nuclear plant can be hours. This motivates a day-ahead market which usually closes 12–36 hours in advance of energy delivery and determines which generators will be online and their output levels for each hour or half an hour over a 24-hour horizon. This is the problem of *unit commitment* and is discussed in Chapter 6.2.1.

   The commitment decisions are determined based on forecast of loads and variable generations such as wind and solar power 12–36 hours in advance. A real-time market computes, every 5–15 minutes in advance of energy delivery, adjustments to generation and consumption levels relative to the schedules produced by the day-ahead market as uncertainty in consumption, generation, and network state is resolved. This is the problem of *real-time dispatch* and is discussed in Chapter 6.2.2.

2 *Frequency control* (Chapter 6.3). Balancing on a second-by-second basis within a real-time dispatch interval takes the form of *frequency control*, currently organized at two timescales. When there is excess supply the rotating machines in bulk generators will speed up and the system frequency will rise. When there is a shortage the rotating machines will slow down and the system frequency will drop.

Frequency deviation is used as a control signal for generators and controllable loads to adjust their power. A generating unit that participates in the *primary control* uses a governor to automatically adjusts its power in proportion to its local frequency deviation in a decentralized manner. Primary control rebalances power and stabilizes the frequency to a new equilibrium value in 30 seconds or so. This is studied in Chapter 6.3.2.

The *secondary control* adjusts generator setpoints around their dispatch values in order to restore system frequency to its nominal value and restore tie-line powers between balancing areas to their scheduled values within a few minutes (the dispatched setpoint and scheduled tie-line flows are determined by real-time dispatch studied in Chapter 6.2.2). These adjustments are determined centrally within each area based on real-time measurements of tie-line flow deviations and frequency deviations in the area. Secondary control is studied in Chapter 6.3.3.

3 *Pricing electricity* (Chapter 6.4). Chapters 6.2 and 6.3 focus on control mechanisms to balance power at timescales from subseconds to a day. The day-ahead and real-time markets determine not only generation schedules, but also electricity prices. In Chapter 6.4.2 we formulate the real-time dispatch problem for market operation and design electricity prices. In Chapter 6.4.3 we show that these prices incentivize optimal dispatch and are revenue adequate for the system operator.

The system operator needs to deal with uncertainties, both discrete uncertainties due to outages of generators, transmission or distribution lines and transformers, and continuous uncertainties due to random fluctuations of renewable generations or loads. In Chapter 6.4.4 we extend basic economic dispatch to security constrained economic dispatch that jointly optimizes energy and reserves. In Chapter 6.4.5 we show how to incorporate security constrained economic dispatch in unit commitment in day-ahead markets.

6.1.2 Basic optimization concepts

Many power system applications can be formulated as optimization problems. In this subsection we introduce some basic concepts of optimization. They provide the language in the rest of this chapter to explain control mechanisms for balancing power supply and demand

A constrained optimization problem is specified by a *primal variable* $x \in \mathbb{R}^n$, an *objective or cost function* $f : \mathbb{R}^n \to \mathbb{R}$, and constraint functions $g : \mathbb{R}^n \to \mathbb{R}^m$ and $h : \mathbb{R}^n \to \mathbb{R}^l$. It takes the form:

$$\min_{x \in \mathbb{R}^n} \; f(x) \;\text{ subject to }\; g(x) = 0,\; h(x) \le 0 \tag{6.1}$$

i.e., our objective is to choose an x^* that minimizes the cost $f(x)$ among all x that satisfies the constraints $g(x) = 0$ and $h(x) \le 0$. The set $X := \{x \in \mathbb{R}^n : g(x) \le 0, h(x) \le 0\}$ is called a *feasible set*. An $x \in X$ is called a *feasible solution* of (6.1). A feasible solution x^* that attains the minimum of f over X (i.e., $f(x^*) \le f(x)$ for all $x \in X$) is

called a (primal) *optimal solution/optimum* or a *minimizer*. Suppose f, g, h are convex and continuously differentiable functions. If there is no constraint then x^* minimizes $f(x)$ if and only if $\nabla f(x^*) = 0$. This optimality condition generalizes to constrained optimization as follows.

Associate with each constraint $g_i(x) = 0$ a variable $\lambda_i \in \mathbb{R}$ and each constraint $h_j(x) \leq 0$ a variable $\mu_j \in \mathbb{R}$. The vector $(\lambda, \mu) := (\lambda_i, i = 1, \ldots, m; \mu_j, j = 1, \ldots, l)$ is called a *Lagrange multiplier (vector)* or a *dual variable*. The problem (6.1) is called a *convex program/problem* when f, g, h are convex functions. Then a primal variable $x^* \in \mathbb{R}^n$ is an optimal solution of (6.1) if and only if there exists a dual variable $(\lambda^*, \mu^*) \in \mathbb{R}^{m+l}$ such that the following conditions are satisfied:

$$\begin{aligned}
&\text{Stationarity :} && \nabla f(x^*) + \nabla g(x^*)\lambda^* + \nabla h(x^*)\mu^* = 0 && \text{(6.2a)}\\
&\text{Primal feasibility :} && g(x^*) = 0,\ h(x^*) \leq 0 && \text{(6.2b)}\\
&\text{Dual feasibility :} && \mu^* \geq 0 && \text{(6.2c)}\\
&\text{Complementary slackness :} && \mu^{*\mathsf{T}} h(x^*) = 0 && \text{(6.2d)}
\end{aligned}$$

where ∇f is the column vector whose ith entry is $\frac{\partial f}{\partial x_i}$, ∇g is the $n \times m$ matrix whose (i, j) entry is $\frac{\partial g_j}{\partial x_i}$. and ∇h is the $n \times l$ matrix whose (i, j) entry is $\frac{\partial h_j}{\partial x_i}$. This is the KKT Theorem 8.15 studied in Chapter 8.3.2. The condition (6.2) is called the *KKT condition* associated with (6.1) and reduces to $\nabla f(x^*) = 0$ when there is no constraint. The dual variable (λ^*, μ^*) in this case is called *dual optimal*. Hence the KKT condition is necessary and sufficient for (x^*, λ^*, μ^*) to be primal and dual optimal when (6.1) is a convex program; it is necessary but generally not sufficient otherwise.

In this chapter we will formulate various control and pricing mechanisms as constrained optimization of the form

$$\min_{u,x} \ f(u,x) \quad \text{s.t.} \quad g(u,x) = 0, \ \ h(u,x) \leq 0$$

This is called an *optimal power flow (OPF) problem* and it is a basic building block that underlies numerous power system applications. The optimization variable (u, x) consists of control u and network state x and can span multiple time periods, e.g., in unit commitment problems. The optimization variable (u, x), the cost function f and the constraint functions g, h depend on the application under study. There are usually two types of constraint. The first is power flow equations in various forms studied in Chapters 4 and 5 for single-phase networks and Chapters 16 and 17 for unbalanced multiphase networks. The second type of constraint consists of operational limits such as voltage limits, capacity limits on generators and loads, and thermal and stability limits on transmission lines and transformers.

Structural properties of general OPF problems and algorithms for solving them are studied in detail in Part II.

6.2 Unit commitment and real-time dispatch

In this and next section we describe a hierarchy of control mechanisms for balancing power at timescales from daily to minutes to subseconds. In Chapter 6.4 we explain how to price electricity to incentivize optimal dispatch.

6.2.1 Unit commitment

The problem of unit commitment is typically solved by the system operator in the day-ahead market 12–36 hours in advance of energy delivery to decide which units will be turned on for each hour or half an hour over a 24-hour period. Integral to the commitment decision is also a dispatch decision that determines the output levels of those units that will be online. The commitment decision is made assuming that the dispatch decision will be optimized at delivery time. This can be formulated as a two-stage optimization problem. For most day-ahead markets, the commitment decision is binding but the dispatch decision can be binding or advisory, to be adjusted by economic dispatch in the real-time market. We will discuss in detail the problem of real-time dispatch in Chapter 6.2.2, so we will focus on formulating the commitment decision in this section.

Consider a time horizon $T := \{1,2,\dots,T\}$ and a power network represented as a graph $G := (\overline{N},E)$ as before. For example, each time t represents an hour and $T = 24$. For each period $t \in T$ let $u(t) := (u_j(t), j \in \overline{N})$ denote controllable real and reactive power injections at time t, $V(t) := (V_j(t), j \in \overline{N})$ the voltage phasor, $S(t) := \big(S_{jk}(t), S_{kj}(t), (j,k) \in E\big)$ the complex line flows. We call $u(t)$ a *dispatch* and $x(t) := (V(t), S(t))$ a *network state* at time t. Let $u := (u(t), t \in T)$ and $x := (x(t), t \in T)$. They are complex vectors of appropriate sizes. Let $\kappa_j(t) \in \{0,1\}$ be the binary variable indicating that unit j will be on at time t if $\kappa_j(t) = 1$ and off otherwise. Let $\kappa(t) := \Big(\kappa_j(t), j \in \overline{N}\Big)$ and $\kappa := (\kappa(t), t \in T)$.

Our OPF formulation includes only three features of the unit commitment problem. The first is injection bounds on a unit when it is turned on. This can be expressed as the constraint:

$$\underline{u}_j(t)\kappa_j(t) \;\le\; u_j(t) \;\le\; \overline{u}_j(t)\kappa_j(t), \qquad j \in \overline{N} \tag{6.3a}$$

where $\underline{u}_j(t)$ and $\overline{u}_(t)j$ are given bounds on the active and reactive injections respectively at bus j at time t.[1] The second feature is the startup and shut down costs incurred by a bulk unit when it is turned on or off. This can be expressed as a cost function d_t

[1] All variables are complex and, by $a \le \overline{a}$ where $a, \overline{a} \in \mathbb{C}$, we mean separate bounds on the real and imaginary parts, Re $a \le$ Re $\overline{a}$ and Im $a \le$ Im $\overline{a}$.

that is positive when the on/off status of the unit changes:

$$d_{jt}(\kappa_j(t-1),\kappa_j(t)) := \begin{cases} \text{startup cost} & \text{if} \quad \kappa_j(t)-\kappa_j(t-1)=1 \\ \text{shutdown cost} & \text{if} \quad \kappa_j(t)-\kappa_j(t-1)=-1 \\ 0 & \text{if} \quad \kappa_j(t)-\kappa_j(t-1)=0 \end{cases} \tag{6.3b}$$

Once turned on or off, if a bulk generator must stay in the same on/off state for a minimum amount of time, this imposes constraints that keep track of the time since the last on/off state change:

$$\kappa_j(t)-\kappa_j(t-1) \;\le\; \kappa_j^\tau, \qquad \forall \tau \in \{t+1, t+\text{up}_j-1\} \tag{6.3c}$$

$$\kappa_j(t-1)-\kappa_j(t) \;\le\; 1-\kappa_j^\tau, \qquad \forall \tau \in \{t+1, t+\text{down}_j-1\} \tag{6.3d}$$

where up_j and down_j are the minimum up and down time respectively once turned on or off.

We illustrate how unit commitment can be posed as an OPF using the simplest formulation that includes only the three features in (6.3). Unit commitment is then the following two-stage optimization problem:

$$\min_{\kappa\in\{0,1\}^{(N+1)T}} \quad \sum_t \sum_j d_{jt}\left(\kappa_j(t-1),\kappa_j(t)\right) \;+\; f^*(\kappa) \tag{6.4a}$$

$$\text{s.t.} \qquad (6.3c)(6.3d) \tag{6.4b}$$

where the startup/shut down costs d_{jt} are given by (6.3b). Given a commitment decision κ, $f^*(\kappa)$ is the optimal real-time dispatch cost over the entire optimization horizon:

$$f^*(\kappa) \;:=\; \min_{(u,x)} \sum_t f_t(u(t),x(t);\kappa(t)) \tag{6.4c}$$

$$\text{s.t.} \quad g_t(u(t),x(t);\kappa(t))=0, \;\; h_t(u(t),x(t);\kappa(t)) \le 0, \quad t\in T \tag{6.4d}$$

$$\tilde{g}(u,x)=0, \;\; \tilde{h}(u,x)\le 0 \tag{6.4e}$$

Here f_t is the dispatch cost, e.g., fuel cost, at time t. The constraints (6.4d) include power flow equations and capacity limits such as (6.3a) at each time t, and the constraints (6.4e) are inter-temporal constraints such as ramp rate limits of the form $|u_j(t)-u_j(t-1)| \le \rho_j$. Hence the commitment decision κ is chosen in (6.4a) in anticipation that the dispatch decisions $(u(t),x(t))$ will be optimized in the second-stage problem (6.4c)(6.4d)(6.4e).

The second-stage problem (6.4c)(6.4d)(6.4e) approximates real-time dispatch problem (6.6) explained in Chapter 6.2.2, even though real-time dispatch operates in 5-15 minute intervals (instead of hourly or half-hourly) and may not include temporal constraints (6.4e). It uses the forecast of uncontrollable injections (generations and loads) as parameters in power flow equations in (6.4d). The revised forecast of these parameters is typically much better when real-time dispatch is computed.

Remark 6.1 (Unit commitment in practice)**.** The unit commitment problem (6.4) is nonconvex and computationally challenging for large networks. Nonconvexity is due both to the binary variable κ and the nonlinear power flow equations. In practice these nonlinear power flow equations are usually replaced by their linear approximations

such as the DC power flow model (see an example in Chapter 6.4.5). This reduces the problem to a mixed integer linear program (MILP) and can often be solved within the available time using branch and bound methods (Chapter 8.5.6) or Benders decomposition (see Example 8.17 in Chapter 8.5.7). The solution (κ^*, u^*, x^*) of the MILP however may not satisfy the original nonlinear constraints. Typically the nonlinear power flow model is then used to check if the commitment and dispatch decisions (κ^*, u^*) will produce a state x that satisfies operational constraints such as voltage and line limits. This involves solving nonlinear power flow equations. If operational constraints are violated, the MILP is modified and the procedure is repeated.

Active effort is underway in the R&D community and industry to scale computation methods for mixed integer nonlinear programs to large networks, so that the OPF problem (6.4) can be applied in day-ahead markets. See Chapter 9.5 for an example. □

6.2.2 Real-time dispatch

After the on-off status of generating units and large controllable loads have been determined by a day-ahead market, a real-time market computes every 5-15 minutes optimal injection levels of those units that are online. This is the problem of optimal, or economic, dispatch. While the control, or dispatch, interval t for unit commitment is typically an hour or half an hour, the dispatch interval t for economic dispatch is 5-15 minutes. The most common, and simplest, form of the problem computes an optimal dispatch in each interval without taking into account decisions in future intervals (except for security constrained economic dispatch studied in Chapter 6.4.4). We hence fix a control interval and drop the time index t in our notation.

In this subsection we formulate the real-time dispatch problem and discuss causes for intra-interval imbalance. In the next section we describe frequency control mechanisms that balance power within a dispatch interval.

OPF formulation.

Consider a set of buses $\overline{N}$ and assume there is a generator or controllable load at each bus $j \in \overline{N}$. Let $u := (u_j, j \in \overline{N})$ denote the complex controllable injections, $V := (V_j, j \in \overline{N})$ the voltage phasors, and $S := \left(S_{jk}, S_{kj}, (j,k) \in E\right)$ the complex line flows. We call u a *dispatch* and $x := (V, S)$ a *network state*. They are complex vectors of appropriate sizes. Let $\sigma := \left(\sigma_j, j \in \overline{N}\right)$ be given complex uncontrollable injections. For real-time dispatch the objective function $f(u,x)$ may represent fuel cost which may be convex quadratic in real power generation:

$$f(u,x) = \sum_{\text{generators } j} \left(a_j \left(\text{Re}(u_j)\right)^2 + b_j \,\text{Re}(u_j)\right)$$

for some $a_j \geq 0$, $b_j \geq 0$.

The relation between the line flows $S := \left(S_{jk}, (j,k) \in E\right)$ and voltages $V := \left(V_j, j \in \overline{N}\right)$ is specified by the power flow equation

$$S = S(V) \tag{6.5a}$$

where we have abused notation to use S_{jk} to denote both a line flow and a function of voltages. For example we can write the line flow S_{jk} in terms of V in the complex form (4.25) reproduced here:

$$\begin{aligned} S_{jk}(V) &= \left(y^s_{jk}\right)^H \left(|V_j|^2 - V_j V_k^H\right) + \left(y^m_{jk}\right)^H |V_j|^2, \qquad (j,k) \in E \\ S_{kj}(V) &= \left(y^s_{jk}\right)^H \left(|V_k|^2 - V_k V_j^H\right) + \left(y^m_{kj}\right)^H |V_k|^2, \qquad (j,k) \in E \end{aligned}$$

where (y^s_{jk}, y^m_{jk}) and (y^s_{kj}, y^m_{kj}) are series and charging admittances of line (j,k), or in polar form (see (4.27)):

$$\begin{aligned} P_{jk}(V) &= \left(g^s_{jk} + g^m_{jk}\right)|V_j|^2 - |V_j||V_k|\left(g^s_{jk}\cos\theta_{jk} + b^s_{jk}\sin\theta_{jk}\right), \quad (j,k) \in E \\ Q_{jk}(V) &= \left(b^s_{jk} + b^m_{jk}\right)|V_j|^2 - |V_j||V_k|\left(g^s_{jk}\sin\theta_{jk} - b^s_{jk}\cos\theta_{jk}\right), \quad (j,k) \in E \end{aligned}$$

where (g^s_{jk}, b^s_{jk}) and (g^m_{jk}, b^m_{jk}) are series and charging admittances of line (j,k) and $\theta_{jk} := \angle V_j - \angle V_k$. Similarly for $\left(P_{kj}(V), Q_{kj}(V)\right)$ in the opposite direction on line (j,k). Different power flow equations lead to different OPF formulations with different computational properties. Then power balance is expressed as[2]

$$u_j + \sigma_j = \sum_{k:j\sim k} S_{jk}, \qquad j \in \overline{N} \tag{6.5b}$$

The most common operational constraints are:

- *Injection limits* (e.g., generator or load capacity limits):

$$\underline{u}_j \le u_j \le \overline{u}_j, \qquad j \in \overline{N} \tag{6.5c}$$

  where $\underline{u}_j$ and $\overline{u}_j$ are given bounds on the active and reactive injections respectively at buses j.[3]
- *Voltage limits*:

$$\underline{v}_j \le |V_j|^2 \le \overline{v}_j, \qquad j \in \overline{N} \tag{6.5d}$$

  where $\underline{v}_j$ and $\overline{v}_j$ are given lower and upper bounds on the squared voltage magnitudes. We assume $\underline{v}_j > 0$ to avoid triviality (in practice $v_j \approx 1$ pu).

[2] If $y^s_{jk} = y^s_{kj}$ and $y^m_{jk} = y^m_{kj} = 0$ then we can model the network by a directed graph described by a node-by-line incidence matrix C. In this case (6.5b) takes the form $u + \sigma = CS$.

[3] All variables are complex and, by $a \le \overline{a}$ where $a, \overline{a} \in \mathbb{C}$, we mean separate bounds on the real and imaginary parts, Re $a \le$ Re $\overline{a}$ and Im $a \le$ Im $\overline{a}$.

- *Line limits*: Thermal limits can be expressed as upper bounds on the magnitudes of line currents, on the magnitudes of real and reactive line power, or on the apparent line power, as:

$$|S_{jk}| \le \overline{S}_{jk}, \quad |S_{kj}| \le \overline{S}_{kj}, \qquad (j,k) \in E \tag{6.5e}$$

The real-time dispatch problem is then the following constrained optimization:

$$\min_{u,x} \; f(u,x) \quad \text{s.t.} \quad (6.5) \tag{6.6}$$

where $(u,x) := (u,V,S) \in \mathbb{C}^{2(N+1+M)}$ and $N+1, M$ are the numbers of buses and lines respectively. It is solved by the system operator for every control interval (e.g., every 5 minutes). It is what the second-stage problem (6.4c)(6.4d)(6.4e) in unit commitment approximates, although at a coarser timescale (hourly or half-hourly) and less accurate forecast of uncontrollable injections σ in (6.5b). We call u a *feasible dispatch* if $(u,x) := (u,V,S)$ satisfies (6.5) for some network state x. We call u^{opt} an *optimal dispatch* if $\left(u^{\text{opt}}, x^{\text{opt}}\right) := (u^{\text{opt}}, V^{\text{opt}}, S^{\text{opt}})$ is an optimal solution of (6.6) for some network state x^{opt}. The key parameter of (6.6) is the uncontrollable injection σ in (6.5b). We often abuse notation and write $u^{\text{opt}}(\sigma)$ for an optimal dispatch as a function of σ. We also say that the optimal dispatch $u^{\text{opt}}(\sigma)$ is *driven by* σ.

The interpretation of an optimal $(u^{\text{opt}}, x^{\text{opt}})$ is that the controllable generators and loads will produce and consume according to the dispatch command u^{opt} from the system operator. The injection u^{opt} will drive the voltage V^{opt} and line flow S^{opt} on the network to a solution of the power flow equations (6.5a)(6.5b) that satisfies the operational constraints (6.5c)(6.5d)(6.5e). In particular this should guarantee power balance at all points of the network given an uncontrollable injection σ. The reality is more complicated as we will see below.

Remark 6.2. We have assumed without loss of generality that there is at most one controllable generator or load at each bus with injection u_j. It is straightforward to extend to the case where there are multiple generators and loads at buses j. If there is no controllable injection at bus j then we can set $\underline{u}_j = \overline{u}_j = 0$ or remove u_j as an optimization variable. □

Remark 6.3 (Economic dispatch in practice)**.** The nonlinearity of power flow equations (6.5a) makes the real-time dispatch problem (6.6) nonconvex and the standard economic theory inapplicable. Most markets today adopt a linear approximation of (6.5a), e.g., the DC power flow model together with methods to determine reactive injections, to compute electricity prices together with a candidate dispatch u. This problem is usually called *DC OPF* or *economic dispatch*.[4] Given a candidate dispatch u from an economic dispatch problem a system operator may check using AC power flow equations (6.5a)(6.5b) whether the resulting network state $x := (V,S)$ satisfies the operational constraints (6.5c)(6.5d)(6.5e), i.e., whether (u,V,S) is feasible for (6.6). If

[4] In the literature, economic dispatch usually refers to the special case of DC OPF where line limits are ignored (i.e., formulation (6.22) without the constraint (6.22c)), but we do not make this distinction.

it is, then the system operator may price electricity according to a dual optimal solution of the economic dispatch problem and dispatch the injection u. Otherwise the system operator may adjust the parameters of the DC OPF problem and repeat the cycle. Even though this procedure may not produce an optimal solution of (6.6) it avoids the complication of nonconvex pricing. We will study the pricing of electricity in detail in Chapters 6.4 and 6.4.4. □

Intra-interval imbalance.

Suppose the uncontrollable injection (vector) $\sigma := (\sigma(t), t \in \mathbb{R}_+)$ is a continuous-time stochastic process with the mean process $m(t) := E\sigma(t)$. This can model wind or solar generation or inelastic demand. A realization $\sigma(\xi) := (\sigma(\xi,t), t \in \mathbb{R}_+)$ of the process is indexed by ξ associated with a probability space, though we may omit ξ and use σ or $\sigma(t)$ to refer to a realization when there is no risk for confusion. For each realization ξ and time $t \geq 0$ let $u(\sigma(\xi,t))$ denote an actual injection that can maintain power balance at all points of the network at time t. For instance $u(\sigma(\xi,t))$ is an optimal dispatch driven by the realization $\sigma(\xi,t)$, i.e., there exists a network state $x(\sigma(\xi,t))$ such that $(u(\sigma(\xi,t)), x(\sigma(\xi,t)))$ is an optimal solution of the (deterministic) problem

$$\min_{(u,x):=(s,V,S)} \quad f(x) \quad \text{s.t. (6.5a)(6.5c)(6.5d)(6.5e)} \tag{6.7a}$$

$$u_j + \sigma_j(\xi,t) = \sum_{k:j\sim k} S_{jk}, \quad \forall j \tag{6.7b}$$

It is of course impractical to compute such an optimal dispatch for each realization ξ at each time $t \geq 0$. Moreover the power flow model in (6.7) describes the steady state behavior and is not suitable for analyzing fast dynamics required for correcting intra-interval imbalances. For this we study dynamic models in Chapter 6.3.

Instead, a dispatch is computed by the real-time market in each discrete time period $n\delta$, $n = 0, 1, \ldots$, where δ is the duration of each control interval, e.g., $\delta = 5$ minutes. Suppose the system operator's dispatch for the nth control interval is an optimal solution $u^{\text{opt}}(\hat{m}(n))$ of (6.6), or its linear approximation, driven by a certain forecast $\hat{m}(n)$ of the uncontrollable injection $\sigma(\xi,t)$ over the interval. The imbalance at time t is then the difference between the injection required for power balance and the operator's dispatch:

$$\Delta u(\xi,t) := u(\sigma(\xi,t)) - u^{\text{opt}}(\hat{m}(n)), \qquad t \in [n\delta, (n+1)\delta),\ n = 0, 1, \ldots \tag{6.8}$$

In Exercise 6.1 we describe an error model in which this imbalance decomposes into three types of errors:

$$\Delta u(\xi,t) = \Delta_1(\xi,t) + \Delta_2(t) + \Delta_3(\xi,t)$$

where $\Delta_1(\xi,t)$ is a random error, $\Delta_2(t)$ a discretization error and $\Delta_3(t)$ a prediction error. In this error model, the mean random error $E\Delta_1(t) = 0$, the time average of the discretization error $\Delta_2(t)$ is zero over each control interval, and the mean prediction

error $E\Delta_3(t)$ is small if the mean process $m(t)$ is slowly time-varying. In particular if σ is stationary then $E\Delta_3(t) = 0$.

The imbalance $\Delta u(\xi, t)$ is corrected by frequency control. The operator dispatch $u^{\text{opt}}(\hat{m}(n))$ is not the actual power injection but provides setpoints for controllable generators and loads for the nth control interval. While these setpoints $\hat{u}(n)$ are updated every control interval (5-15 minutes), frequency control operates continuously to determine the actual power injection. We study frequency control in Chapter 6.3 using a dynamic model that includes the fast timescale generator and frequency dynamics and the feedback control to maintain frequency around its nominal value.

Before that, we first discuss how to handle large imbalances due contingency events.

6.2.3 Security constrained OPF

Power system security refers to the ability to withstand large disturbances. The small random imbalances are handled by real-time optimal dispatch and frequency control mechanisms discussed in Chapters 6.2.2 and 6.3 respectively. In this section we explain techniques to handle large disturbances due to contingency events such as the unanticipated loss of a bulk generator or wind or solar farm, the switching on or off of a large industrial load, or the outage of a transmission line or transformer in the transmission network.

Contingency events are rare but their potential impacts are large. North American Electric Reliability Corporation's (NERC) $N-1$ rule states that the outage of a single piece of equipment (e.g., generator, line, transformer) should not result in flow or voltage limit violations. As volatile generation from wind and solar farms continues to displace thermal generators, a large deviation of such nondispatchable generation from its predicted value may also count as a contingency event in the future. For instance the random generation can be modeled as taking one of a finite number of values, each triggering a contingency response if it differs significantly from its predicted value.

Secure operation is achieved through three main mechanisms: (i) analyze credible contingencies that may lead to voltage or line limit violations, (ii) account for these contingencies in optimal commitment and dispatch schedules, and (iii) monitor system state in real time and take corrective actions when a contingency occurs. We summarize each of these functions.

Contingency analysis

When a generator or load contingency occurs the resulting power flows might violate line limits and lead to transmission outages where transmission lines or transformers are disconnected. If reserve capacity is insufficient to re-balance generation and demand, frequency excursion will continue which can disconnect other generators to protect

them from damage, potentially leading to involuntary load shedding and even system collapse. When a transmission line or transformer is disconnected power flows in the network will redistribute and line limits can be violated, potentially leading to cascading line outages. Furthermore a transmission outage can result in reactive losses in the network which can suppress voltage magnitudes, leading to voltage violations.

The impacts of these contingency events can be assessed by solving AC power flow equations that describe the network state after each contingency. Currently this set of post-contingency equations are solved in the industry mostly using Newton-Raphson or the decoupled power flow methods because they have good speed and convergence properties. Due to the large number of contingencies that must be assessed in order to satisfy $N-k$ security for $k \geq 1$, it is a common practice to first use the DC power flow model to quickly screen contingencies and select a much smaller subset that result in voltage or line limit violations for more detailed analysis using the AC power flow model, especially for contingency scenarios where voltage magnitudes and reactive flows are important. For instance the DC power flow model can quickly estimate incremental line flow changes due to a contingency from the pre-contingency operating point determined by the AC power flow model, through the use of pre-computed quantities called the power transfer distribution factor and the line outage distribution factor. Contingency scenarios in which line or voltage limits are violated are called *credible contingencies*. (Chapter 9.5.3 presents some techniques for contingency screening for industrial-scale security constrained AC OPF.)

Security constrained dispatch and commitment

The credible contingencies that have been identified in contingency analysis are taken into account in day-ahead (e.g., 12–36 hours) unit commitment and real-time (e.g., 5–15 minutes) dispatch as well as automatic generation control (seconds to minutes). Capacities are reserved for normal operation (regulation and load-following reserves) and for contingencies (contingency reserves).

There are two approaches to account for credible contingencies in scheduling dispatch. The *preventive approach* augments the optimal dispatch problem studied in Chapter 6.2.2 with additional constraints so that the network state under the optimal dispatch will satisfy operational constraints even after contingency events. This allows the dispatch to remain unchanged until the next real-time dispatch period even if a contingency occurs in the middle of the current period. The intra-period imbalance due to contingency will be handled by the frequency control mechanisms studied in Chapter 6.3. The *corrective approach*, on the other hand, will compute in advance optimal dispatches both for normal operation and after each contingency event. This allows the system operator to dispatch a response immediately after a contingency is detected without having to wait till the next dispatch period. Both approaches can be formulated as security constrained OPF problems, as we will see below.

System monitoring

A system operator's energy management system collects and processes measurements of voltages, currents, line flows, and the status of circuit breakers and switches at all transmission substations. Other measurements such as frequencies, generator outputs, and transformer tap positions are also measured at various locations of a transmission network, e.g., using phasor measurement units. These measurements are used for state estimation (Chapter 7.1), real-time dispatch (Chapter 6.2.2), and automatic generation control (Chapter 6.3), among other applications. Based on these measurements the system can be classified as in a normal state, in an emergency state, or in a restoration state after a contingency, with default actions in each of these states.

Security constrained OPF

We contrast the preventive and the corrective approaches to handling contingencies using the real-time dispatch problem of Chapter 6.2.2 as an example. These approaches can also be applied to unit commitment; see Example **??**. Security constrained OPF are used in both control and market applications.

The real-time dispatch problem (6.6) without security constraints studied:

$$\begin{aligned} \min_{(u_0,x_0)} \quad & f_0(u_0,x_0) \\ \text{s.t.} \quad & g_0(u_0,x_0) = 0, \quad h_0(u_0,x_0) \leq 0 \end{aligned} \tag{6.9}$$

serves as the base or pre-contingency case. Here u_0 is a vector representing controls such as real power injections of controllable generators and loads, generator voltage magnitudes, transformer tap positions, x_0 is a vector representing the network state such as bus voltage magnitudes and angles at load buses, $g_0(x_0,u_0)$ represents linear or nonlinear power flow equations, and $h_0(x_0,u_0)$ represents operational constraints such as voltage and line flow limits, all in the base case.

Let credible contingencies be indexed by $k = 1,\ldots,K$. After a contingency k, the dispatch u_0 remains unchanged in the short term (e.g., 1–5 mins). The network state however changes immediately from x_0 to a new system state $\tilde{x}_k$ determined by the post-contingency network and frequency control actions. The choice of pre-contingency dispatch u_0 can take the new network state into account, in three ways.

Some operational constraints such as thermal limits may be temporarily relaxed immediately after the contingency, provided corrective actions will be implemented quickly. A preventive approach chooses u_0 so that emergency operational constraints in the short term are satisfied before corrective actions take effect. Let $\tilde{g}_k$ denote the power flow equations for the post-contingency network, and $\tilde{h}_k$ models the emergency operational constraints after contingency k. The pre-contingency control u_0 and the post-contingency network state $\tilde{x}_k$ in the short term must satisfy:

$$\tilde{g}_k(u_0,\tilde{x}_k) = 0, \quad \tilde{h}_k(u_0,\tilde{x}_k) \leq 0, \qquad k = 1,\ldots,K \tag{6.10}$$

A *preventive security-constrained OPF* (SCOPF) problem chooses an optimal control decision u_0 that will remain secure after each contingency $k = 1,\dots,K$, before corrective actions are implemented, i.e., it is of the form

$$\min_{(u_0, x_0,\ \tilde{x}_k,\ k\geq 1)} \quad f_0(u_0, x_0) \quad \text{s.t.} \quad (6.9)(6.10)$$

In the corrective approach a new dispatch u_k is applied after contingency k. In addition to changes in injections, the corrective control u_k may also include changes to network topology such as line switching or circuit breaker actions. These changes are captured in new power flow equations g_k. While $\tilde{g}_k$ in (6.10) is determined only by the contingency, e.g., a line or generator outage, g_k may include topology changes as part of the corrective control. The operational constraints, modeled by h_k, are generally different from the pre-contingency constraints h_0 and the emergency constraints $\tilde{h}_k$ immediately after contingency k. Besides constraints such as voltage and line limits under control u_k, h_k may also include constraints due to capacity reserves (see Chapter 6.4). The corrective control u_k and the resulting network state x_k therefore must satisfy

$$g_k(u_k, x_k) = 0, \quad h_k(u_k, x_k) \leq 0, \qquad k = 1,\dots,K \tag{6.11a}$$

Often the corrective control u_k is constrained to be close to the base control u_0, e.g., because of limited ramp rates ρ_k of large generators or loads:

$$\|u_k - u_0\| \leq \rho_k, \qquad k = 1,\dots,K \tag{6.11b}$$

Then a *corrective SCOPF* takes the form

$$\min_{(u_k, x_k,\ k\geq 0)} \quad \sum_{k\geq 0} w_k f_k(u_k, x_k) \quad \text{s.t.} \quad (6.9)(6.11) \tag{6.12}$$

where f_k are costs that can depend on the contingency and $w_k \geq 0$ are nonnegative weights, e.g., the probability of contingencies k.

This corrective approach ignores the emergency constraints (6.10) and assumes the system will ride through the small delay between the time a contingency occurs and when the corrective control u_k takes effect. This allows more flexibility in the base control u_0 and lowers the cost of normal operation. A more secure and potentially more costly approach will impose both the emergency constraints as well as constraints on the corrective control:

$$\min_{(u_k, x_k,\ \tilde{x}_{k+1},\ k\geq 0)} \quad \sum_{k\geq 0} w_k f_k(u_k, x_k) \quad \text{subject to} \quad (6.9)(6.10)(6.11)$$

6.3 Frequency control

The power delivered by a thermal generator is determined by the mechanical power output of a prime mover such as a steam turbine or water turbine. The output level is

controlled by opening or closing valves that regulate steam or water flow. For example if the load increases the valve of a generator must open wider to increase the generated power. When there is excess supply the rotating machines in bulk generators will speed up and the system frequency will rise. When there is a shortage the rotating machines will slow down and the system frequency will drop. If power is not re-balanced by adjusting generators or flexible loads, frequency excursion will continue which can disconnect generators to protect them from damage, potentially leading to involuntary load shedding and even system collapse. Frequency deviation from its nominal value is used as a control signal for generators and controllable loads that participate in frequency control to adjust their power.

Frequency control, also referred to as *automatic generation control*, consists of three mechanisms operating at timescales from seconds to minutes. A generating unit that participates in the *primary control*, also called *droop control*, uses a governor to automatically adjusts the mechanical power output of a turbine in proportion to its local frequency deviation. Primary frequency control is decentralized. It rebalances power and stabilizes the frequency to a new equilibrium value in 30 seconds or so. The *secondary control* adjusts generator setpoints around their dispatch values in order to restore system frequency to its nominal value within a few minutes, e.g., up to 10 minutes after a contingency event. In an interconnected power system consisting of multiple balancing areas, each managed by a single operator, the secondary control additionally restores interchanges of tie-line power between areas to their scheduled values. The adjustments are determined centrally within each area based on real-time measurements of tie-line flow deviations. The dispatched setpoint and scheduled tie-line flows are determined by the *tertiary control* that operates on a timescale of 5–15 minutes. They are chosen to attain economic efficiency as well as restoring the reserve capacities deployed in primary and secondary control so that they are available for contingency response. This is typically determined by solving a real-time dispatch problem as discussed in Chapter 6.2.2 and in Chapters 6.4 and 6.4.4 in the context of electricity pricing.

We now present a linear dynamic model of the primary and secondary control that clarifies the relation between system operator's dispatch $u^{\text{opt}}(\hat{m}(n))$ for each interval and the actual (active) power generation. A description of the physical system, including a generator, a turbine-governor system, a frequency control system, and a voltage control system, as well as their detailed models, are beyond the scope of this book. Our goal in this section is to use a simple model to connect real-time dispatch studied in Chapter 6.2.2 with its realization at a fast timescale.

6.3.1 Assumptions and notations

Consider a control interval $[n\delta, (n+1)\delta)$ for which the tertiary control has determined an optimal dispatch $u^{\text{opt}}(\hat{m}(n))$ with the associated network state $x(n)$ including sched-

uled tie-line flows. We assume that the primary and secondary control converges on a much faster timescale than δ so that the dispatch remains unchanged and serves as the operating point for our incremental model below. We fix a random realization ξ of the uncontrollable injection $\sigma(\xi,t)$. The dynamic model is deterministic with this fixed realization. We hence omit the indices n and ξ in the rest of this section.

We make several simplifying assumptions:

- There is a synchronous generator at each bus that determines the frequency dynamics at the bus. This assumption is only to simplify exposition and can be removed.
- Voltage regulation operates at a faster timescale so that voltage magnitudes $|V_j|$ are fixed for the analysis of frequency control. The effect of voltage regulation can be incorporated into the inertia constant M and damping constant D of (the rotor angle transfer function of) the generator; see below.
- The rotor angles, the internal and terminal (bus) voltage phase angles of generators swing together, i.e., the deviations of these angles from their operating points are equal at all times.
- The lines are lossless, i.e., their shunt admittances $\left(y^m_{jk}, y^m_{kj}\right)$ are zero and series admittances are inductive $y^s_{jk} = \mathbf{i}b_{jk}$ with $b_{jk} < 0$.

With these assumptions our dynamic model focuses on how active power in generating units change the voltage angles and their derivatives, i.e., frequencies. It makes similar assumptions to those in the DC power flow model. In fact the DC power flow describes the steady state of the dynamic model.

The tertiary control (i.e., the real-time dispatch in Chapter 6.2.2) determines active power dispatch u^0_j for the generators and the associated voltage angles θ^0_j and active line flows P^0_{jk} driven by estimates σ^0_j of uncontrollable real power injections. They define the operating point around which we linearize our dynamic model. In particular they satisfy power balance:

$$u^0_j + \sigma^0_j = \sum_{k:j\sim k} P^0_{jk}, \qquad j \in \overline{N}$$

Define the following variables and their perturbations around the operating point:

- $u_j(t)$ denotes the setpoint of generator j at time t. Let $\Delta u_j(t) := u_j(t) - u^0_j$ denote the adjustment to the optimal dispatch u^0_j. The adjustment will be computed by the secondary frequency control.
- $\theta_j(t)$ denotes the (terminal) voltage angle at bus j at time t, relative to a rotating frame of the operating-point frequency ω^0 (which is expected to be close but not necessarily equal to the nominal frequency), i.e., the instantaneous voltage is $v_j(t) = \sqrt{2}|V_j|\cos\left(\omega^0 t + \theta_j(t)\right)$. Define the incremental angle $\Delta\theta_j(t) := \theta_j(t) - \theta^0_j$.

- $\omega_j(t)$ denotes the voltage frequency at bus j defined to be the derivative of the phase angle $\omega^0 t + \theta_j(t)$, i.e., $\omega_j(t) = \omega^0 + \dot{\theta}_j(t)$. Hence the frequency deviation $\Delta\omega_j(t) := \omega_j(t) - \omega_j^0$ satisfies $\Delta\omega_j(t) = \Delta\dot{\theta}_j(t)$.
- $P_{jk}(t)$ denotes the line flow from bus j to bus k on line (j,k). Let $P_{kj}(t) := -P_{jk}(t)$. Let $\Delta P_{jk}(t) := P_{jk}(t) - P_{jk}^0$ and similarly for $\Delta P_{kj}(t)$.
- $p_j^M(t)$ denotes the mechanical power output of the prime mover (e.g., gas or water turbine). Let P_j^{M0} denote its value associated with the operating point $\left(u_j^0, \theta_j^0, \omega^0, P_{jk}^0, \sigma_j^0,\ j \in \overline{N},\ (j,k) \in E\right)$ and $\Delta p_j^M(t) := p_j^M(t) - P_j^{M0}$.
- $a_j(t)$ denotes the valve position of the turbine-governor at bus j. Let a_j^0 denote its value associated with the operating point $\left(u_j^0, \theta_j^0, \omega^0, P_{jk}^0, \sigma_j^0,\ j \in \overline{N},\ (j,k) \in E\right)$ and $\Delta a_j(t) := a_j(t) - a_j^0$.

We will remark on $\left(a_j^0, p_j^{M0}\right)$ below when we describe the turbine-governor model. A common model of the instantaneous line flow $P_{jk}(t)$ as a function of voltage angles $\theta(t) := \left(\theta_j(t), j \in \overline{N}\right)$ is (cf. the polar form power flow equation (4.27a)):

$$P_{jk}(t) = |V_j||V_k|\left(-b_{jk}\right)\sin\left(\theta_j(t) - \theta_k(t)\right), \qquad (j,k) \in E$$

where $\left(-b_{jk}\right) > 0$. We will adopt its linearization around the operating point as our model:

$$P_{jk}(t) = \underbrace{|V_j||V_k|\left(-b_{jk}\right)\sin\left(\theta_j^0 - \theta_k^0\right)}_{P_{jk}^0} + T_{jk}\left(\Delta\theta_j(t) - \Delta\theta_k(t)\right), \qquad (j,k) \in E$$

where $T_{jk} := |V_j||V_k|\left(-b_{jk}\right)\cos\left(\theta_j^0 - \theta_k^0\right)$ are called stiffness coefficients. Hence

$$\Delta P_{jk}(t) = T_{jk}\left(\Delta\theta_j(t) - \Delta\theta_k(t)\right), \qquad (j,k) \in E \tag{6.13}$$

The coefficient T_{jk} measures power exchange over line (j,k) with respect to changes in phase angles.

The model has three components (see Figure 6.1): (i) a turbine-governor that produces mechanical power $p_j^M(t)$ based on the setpoint $u_j(t)$; (ii) a power generator that converts $p_i^M(t)$ of the turbine-governor into electric power that serves the local load $-\sigma_j(t)$ and injects power $\sum_k P_{jk}(t)$ into the transmission system; and (iii) two feedback control mechanisms for primary and secondary frequency control. It describes the dynamics of the incremental variables $\Delta\theta_j$, $\Delta\omega_j$, etc. In the following we will describe dynamic models for the turbine-governor and the generator in Figure 6.1, leading to Figure 6.4.

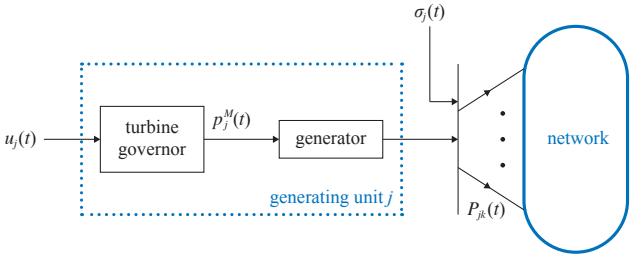


Figure 6.1 Generating unit j, its setpoint $u_j(t)$, local injection $\sigma_j(t)$, and line power $P_{jk}(t)$.

6.3.2 Primary control

Turbine-governor model

A second-order model of the turbine-governor with droop control is:

$$\begin{aligned} T_{gj}\,\dot{a}_j &= -a_j(t) + u_j(t) - \frac{\Delta\omega_j(t)}{r_j}, && j \in \overline{N} \\ T_{tj}\,\dot{p}_j^M &= -p_j^M(t) + a_j(t), && j \in \overline{N} \end{aligned}$$

where the states $a_j(t)$ and $P_j^M(t)$ are the valve position and mechanical power output of the turbine respectively. The constant r_j is called a regulation constant or a droop constant. The term $-\omega_j(t)/r_j$ increases the valve position when the frequency drops below ω^0 and decreases it otherwise. This is referred to as the droop control or the primary frequency control. This model makes several simplifying assumptions, e.g., it ignores the saturation of the valve position $a_j(t)$, but is reasonable when the frequency deviation $\Delta\omega_j(t)$ is small.

We define $\left(a_j^0, P_j^{M0}\right)$ to be the equilibrium point, defined by $\dot{a}_j = \dot{p}_j^M = 0$, when frequency deviations $\Delta\omega_j(t) = 0$ and setpoint $u_j(t) = u_j^0$ is the optimal dispatch, i.e.,

$$p^{M0} = a_j^0 = u_j^0, \qquad j \in \overline{N}$$

Then the incremental variable $\left(\Delta a_j, \Delta P_j^M\right) := \left(a_j - a_j^0,\, P_j^M - P_j^{M0}\right)$ satisfies the same equations:

$$T_{gj}\,\Delta\dot{a}_j = -\Delta a_j(t) + \Delta u_j(t) - \frac{\Delta\omega_j(t)}{r_j}, \qquad j \in \overline{N} \tag{6.14a}$$

$$T_{tj}\,\Delta\dot{p}_j^M = -\Delta p_j^M(t) + \Delta a_j(t), \qquad j \in \overline{N} \tag{6.14b}$$

This incremental model is what we will use. The block diagram representation of (6.14) is in Figure 6.2.

As we will see in Chapter 6.3.3 the setpoint adjustment $\Delta u_j(t)$ is changed by the secondary control at a much slower timescale (several minutes) than that of the primary control (approximately 30 secs). Hence a quasi steady-state of (6.14) is defined

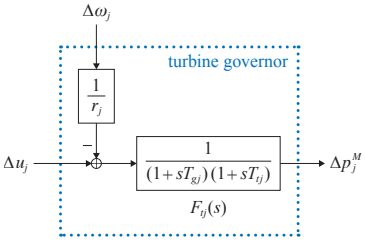


Figure 6.2 Block diagram in Laplace domain of the turbine-governor dynamic (6.14).

by a constant value of the setpoint adjustment $\Delta u_j(t) = \Delta u_j$. In this steady state, the frequency deviation $\Delta\omega_j^*$ is generally nonzero and the incremental mechanical power output Δp_j^{M*} is related to the frequency deviation by

$$\Delta p_j^{M*} \;=\; \Delta a_j^* \;=\; \Delta u_j \;-\; \frac{1}{r_j}\Delta\omega_j^*, \qquad j \in \overline{N}$$

Remark 6.4. The time constants T_{gi}, T_{ti} characterize the responsiveness of the governor and turbine respectively to a change in their input. Typical value of T_{gi} and T_{ti} are approximately 0.1 second and 0.5 second respectively. Since the governor responds much faster than the turbine the model is sometimes simplified to a first-order model

$$T_{tj}\,\Delta\dot{p}_j^M = -\Delta p_j^M(t) \;+\; \Delta u_j(t) - \frac{\Delta\omega_j(t)}{r_j}, \qquad j \in \overline{N}$$

□

Generator model.

The frequency deviation $\Delta\omega_j(t)$ is determined by the rotating speed of a generator driven by the mechanical power output $p_j^M(t)$ of the turbine. A dynamic model of the generator in terms of the incremental variables is:

$$\Delta\dot{\theta}_j \;=\; \Delta\omega_j(t), \qquad\qquad j \in \overline{N} \quad (6.15a)$$

$$M_j\Delta\dot{\omega}_j \;+\; D_j\Delta\omega_j(t) \;=\; \Delta p_j^M(t) \;+\; \Delta\sigma_j(t) \;-\; \sum_{k:j\sim k} \Delta P_{jk}(t), \quad j \in \overline{N} \quad (6.15b)$$

where $\Delta\sigma_j(t)$ is the deviation of the uncontrollable injection from its forecast σ_j^0 and $\Delta P_{jk}(t)$ are the incremental line flows given by (6.13). The block diagram representation of (6.15) is in Figure 6.3. Here M_j is the inertia constant of generator j, and D_j is the sum of damping constant of generator j and the frequency sensitivity of motor-type injection at bus j, as we now explain.

If $\sigma_j(t) < 0$ represents a load, a common model consists of both frequency sensitive load $\sigma_{1j}\left(\omega_j^0 + \omega_j(t)\right)$ such as a motor and frequency insensitive load $\sigma_{2j}(t)$ due to the switching on or off of an electrical device that draws a specified amount of power. Approximate the frequency sensitive load by its linear approximation $\sigma_{1j}\left(\omega^0\right) +$

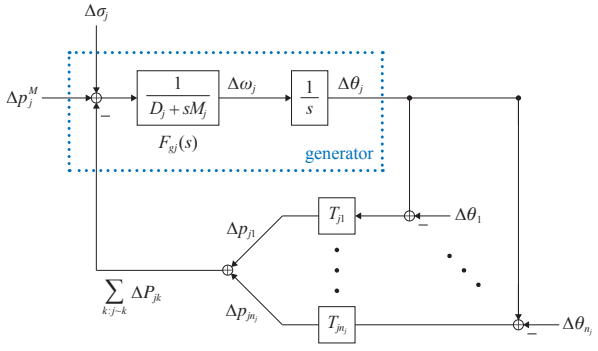


Figure 6.3 Block diagram in Laplace domain of the generator dynamic (6.15).

$\frac{\partial\sigma_{1j}}{\partial\omega_j}\left(\omega^0\right)\Delta\omega_j(t)$ and write the frequency insensitive load as $\sigma_{2j}(t) = \sigma_{2j}^0 + \Delta\sigma_{2j}(t)$. Then the deviation $\frac{\partial\sigma_{1j}}{\partial\omega_j}\left(\omega^0\right)\Delta\omega_j(t)$ of the frequency sensitive load is absorbed into $D_j\Delta\omega_j(t)$ in (6.15b). The uncontrollable load $\sigma_j(t)$ is then the sum of the remaining terms:

$$\sigma_j(t) = \underbrace{\left(\sigma_{1j}\left(\omega^0\right) + \sigma_{2j}^0\right)}_{\sigma_j^0} + \underbrace{\Delta\sigma_{2j}(t)}_{\Delta\sigma_j(t)}$$

In summary the primary frequency control is modeled by (6.13)(6.14)(6.15) reproduced here:

$$T_{gj}\Delta\dot{a}_j = -\Delta a_j(t) + \Delta u_j(t) - \frac{\Delta\omega_j(t)}{r_j}, \qquad j \in \overline{N} \quad (6.16a)$$

$$T_{tj}\Delta\dot{p}_j^M = -\Delta p_j^M(t) + \Delta a_j(t), \qquad j \in \overline{N} \quad (6.16b)$$

$$M_j\Delta\dot{\omega}_j + D_j\Delta\omega_j(t) = \Delta p_j^M(t) + \Delta\sigma_j(t) - \sum_{k:j\sim k}\Delta P_{jk}(t), \qquad j \in \overline{N} \quad (6.16c)$$

$$\Delta P_{jk}(t) = T_{jk}\left(\Delta\theta_j(t) - \Delta\theta_k(t)\right), \qquad (j,k) \in E \quad (6.16d)$$

$$\Delta\dot{\theta}_j = \Delta\omega_j(t), \qquad j \in \overline{N} \quad (6.16e)$$

This closes the droop control loop. The block diagram representation combines those in Figures 6.2 and 6.3 and is shown in Figure 6.4 (which is a detailed version of Figure 6.1). The input to the system are external disturbance $\Delta\sigma_j(t)$ at each each generating unit j and the adjustment $\Delta u_j(t)$ to the dispatch setpoint. Since the secondary control that updates the setpoint operates at a much slower timescale than the primary frequency control timescale, we can understand the behavior of the (quasi) steady state of the primary control by assuming a constant setpoint adjustment $\Delta u_j(t) = \Delta u_j$.

Consider then a step disturbance in the uncontrollable injection where $\Delta\sigma_j(t)$ changes at time $t = 0$ from 0 to a constant value $\Delta\sigma_j$. We say that $x^* := \left(\Delta\omega^*, \Delta P^*, \Delta\theta^*, \Delta a^*, \Delta p^{M*}\right)$ is an *equilibrium point* of (6.16) driven by the step change

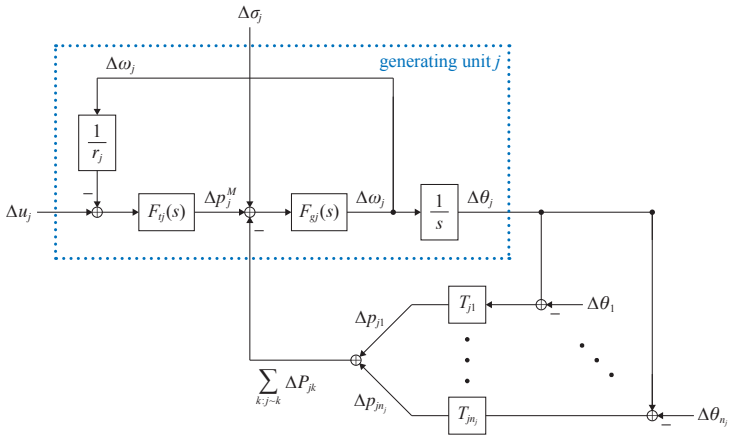


Figure 6.4 Block diagram of primary frequency control (6.16).

$\Delta\sigma$ and constant setpoint Δu_j if, at x^*,

$$\Delta\dot{\omega}_j \;=\; \Delta\dot{a}_j \;=\; \Delta\dot{p}_j^M \;=\; 0, \qquad j \in \overline{N}$$

We do not require $\Delta\dot{\theta} = 0$ in the definition of equilibrium point. Indeed $\Delta\dot{\theta}$ is generally nonzero when primary control converges. Recall the bus-by-line incidence matrix C defined by:

$$C_{jl} := \begin{cases} 1 & \text{if } l = j \to k \text{ for some bus } k \\ -1 & \text{if } l = i \to j \text{ for some bus } i \\ 0 & \text{otherwise} \end{cases}, \qquad j \in \overline{N},\ l \in E$$

The next result calculates the equilibrium frequency and line flows (its proof is left as Exercise 6.2). It motivates secondary control discussed in Chapter 6.3.3.

Theorem 6.1 (Steady state of primary control)**.** Suppose the network is connected. If x^* is an equilibrium point of (6.16) driven by a step changes $\Delta\sigma$ and constant setpoints Δu then:

1 Local frequency deviations converge to a new value equal to the total disturbance divided by the system damping:[5]

$$\Delta\omega_j^* = \Delta\omega^* \;:=\; \frac{\sum_k (\Delta u_k + \Delta\sigma_k)}{\sum_k (D_k + 1/r_k)}, \qquad j \in \overline{N}$$

2 Line flow deviations converge to

$$\Delta P^* = TC^T L^\dagger (\Delta u \,+\, \Delta\sigma \,-\, \Delta\omega^* d)$$

where $T := \mathrm{diag}(T_{jk}, (j,k) \in E)$, $L^\dagger$ is the pseudo inverse of the Laplacian matrix $L := CTC^T$, and $d := (D_j + 1/r_j, j \in \overline{N})$.

[5] We abuse notation to use $\Delta\omega^*$ to both denote a scalar and the vector whose entries are all $\Delta\omega^*$. The meaning should be clear from the context.

Remark 6.5. 1 Intuitively the larger the disturbance or the smaller the system damping, the larger will frequency deviation $\Delta\omega^*$ be. Theorem 6.1 clarifies precisely the simple relationship among them. Droop control r_j adds to the system damping and reduces frequency deviation.

2 The theorem says that frequency can be restored to the operating-point value, i.e., $\Delta\omega^* = 0$, only if we change the setpoints so that the total setpoint changes cancel out the total disturbances

$$\sum_k (\Delta u_k + \Delta\sigma_k) = 0$$

3 To restore all line flows, i.e., $\Delta P^* = 0$, requires canceling disturbances locally at each bus,

$$\Delta u_k + \Delta\sigma_k = 0, \qquad k \in \overline{N}$$

The next example illustrates a benefit of interconnecting multiple areas.

Example 6.1 (Interconnected system)**.** Consider $N+1$ balancing areas each modeled as a single bus. Suppose $\Delta u_j = 0$ for all areas j and that there is a step change of the uncontrollable injection where $\Delta\sigma_j(t)$ changes at time 0 from 0 to a value $\Delta\sigma_j$. Suppose $\Delta\sigma_j$ are independent random variables with mean $\Delta\bar{\sigma}_j$ and variance ν_j^2. We will evaluate the equilibrium frequency deviation $\Delta\omega^*$ using Theorem 6.1 when the primary frequency control converges.

Case 1: Independent operation. Suppose these buses are not connected. Then the equilibrium frequency deviation in each area j is

$$\Delta\omega_j^* = \frac{\Delta\sigma_j}{d_j}, \qquad j \in \overline{N}$$

where $d_j := D_j + 1/r_j$ with mean $\Delta\bar{\sigma}_j/d_j$ and variance ν_j^2/d_j^2.

Case 1: Interconnected system. Suppose these buses are connected. Then the equilibrium frequency deviation for the entire interconnected system is

$$\Delta\omega^* = \frac{\sum_j \Delta\sigma_j}{\sum_j d_j} = \frac{1}{N+1}\sum_j \frac{\Delta\sigma_j}{\hat{d}}$$

where $\hat{d} := \sum_j d_j/(N+1)$ is the average system damping. Define the average mean and variance of $\Delta\sigma_j$ respectively:

$$\Delta\hat{\sigma} := \frac{1}{N+1}\sum_j \Delta\bar{\sigma}_j, \qquad \hat{\nu}^2 := \frac{1}{N+1}\sum_j \nu_j^2$$

Then the mean and variance of $\Delta\omega^*$ are respectively

$$\text{mean}\,(\Delta\omega^*) = \frac{\Delta\hat{\sigma}}{\hat{d}}, \qquad \text{var}\,(\Delta\omega^*) = \frac{1}{N+1}\frac{\hat{\nu}^2}{\hat{d}^2}$$

The simple case when the random variables $\Delta\sigma_j$ are i.i.d. (independently and

identically distributed) with mean $\Delta\bar{\sigma}_1$ and variance v_1^2. Suppose also $d_j = d_1$ for all j. Then $\Delta\hat{\sigma} = \Delta\bar{\sigma}_1$, $\hat{v}^2 = v_1^2$, and $\hat{d} = d_1$. Hence the mean of the interconnected system is the same as that of each area in independent operation, but the variance is reduced by a factor of $N+1$. The bigger the interconnection, i.e., larger N, the smaller the variance in equilibrium frequency deviation $\Delta\omega^*$. □

6.3.3 Secondary control

The first objective of the secondary control is to restore system frequency, i.e., to drive $\Delta\omega(t)$ to zero. The second objective is to restore line flows to their scheduled values, i.e., to drive $\Delta P(t)$ to zero. This is less important and sometimes not pursued for an island system managed by a single operator. In an interconnected system consisting of multiple areas managed by separate operators the interchanges of tie-line power between areas have financial implications. Such a system usually operates under the principle that (i) each area absorbs its own load changes, and (ii) scheduled tie-line flows are maintained. If each bus in (6.16) models an entire area this requires driving $\Delta P(t)$ to zero.

Theorem 6.1 suggests that the objectives of the secondary control can only be achieved by adjusting the setpoints $u(t)$ of the generators to cancel the disturbances (see Remark 6.5). Suppose each bus j in (6.16) represents an area and the setpoint adjustment $\Delta u_j(t)$ represents an aggregate adjustment that will then be shared by all generators in area j that participate in the secondary control. The adjustment is based on the *area control error* (ACE) which is a weighted sum of frequency and line flow deviations:

$$\mathrm{ACE}_j(t) := \sum_{k:j\sim k} \Delta P_{jk}(t) \;+\; \beta_j \Delta\omega_j(t), \qquad j \in \overline{N}$$

where $\beta_j > 0$ is called a frequency bias setting. The setpoint adjustment $\Delta u_j(t)$ integrates ACE_j in order to drive it to zero:

$$\Delta\dot{u}_j = -\gamma_j \left(\sum_{k:j\sim k} \Delta P_{jk}(t) \;+\; \beta_j \Delta\omega_j(t) \right), \qquad j \in \overline{N} \tag{6.17}$$

The computation (6.17) requires real-time measurement of tie-line flow deviations $\Delta P_{jk}(t)$ with all neighboring areas k. This information is sent to area j's system operator which centrally computes the aggregate adjustment $\Delta u_j(t)$ for the entire area using (6.17). It then dispatches in real time setpoint adjustments $\alpha_{ji}\Delta u_j(t)$ with $\alpha_{ji} \geq 0$ and $\sum_i \alpha_{ji} = 1$ to participating generators i in area j. The weights α_{ji} are called participation factors.

In summary the primary and secondary frequency control in area j is modeled by the system (6.16)(6.17). It is driven by the uncontrollable injection $\Delta\sigma_j(t)$ and consists

of two feedback control mechanisms, the droop control with regulation parameter r_j and setpoint adjustment based on $\mathrm{ACE}_j(t)$. Its block diagram is shown in Figure 6.5.

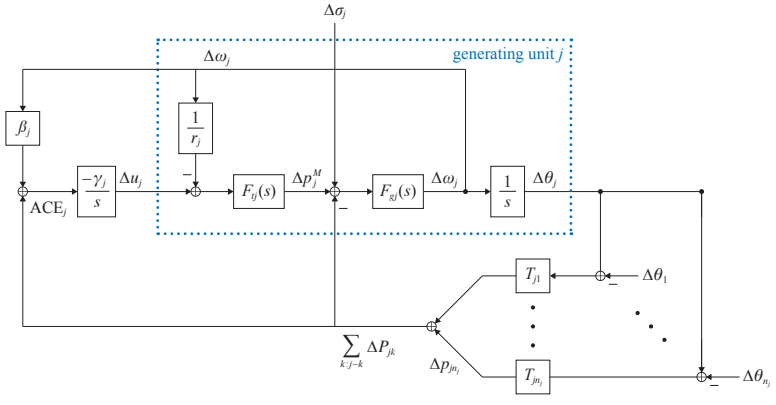


Figure 6.5 Block diagram of primary and secondary frequency control (6.16) (6.17) in area j.

To understand the behavior of the entire interconnected system it is convenient to write (6.16)(6.17) in vector form:

$$T_g\,\Delta\dot{a} = -\Delta a(t) + \Delta u(t) - R^{-1}\Delta\omega_j(t) \tag{6.18a}$$

$$T_t\,\Delta\dot{p}^M = -\Delta p^M(t) + \Delta a(t) \tag{6.18b}$$

$$M\Delta\dot{\omega} + D\Delta\omega(t) = \Delta p^M(t) + \Delta\sigma(t) - C\Delta P(t) \tag{6.18c}$$

$$\Delta P(t) = TC^T\,\Delta\theta(t) \tag{6.18d}$$

$$\Delta\dot{\theta} = \Delta\omega(t) \tag{6.18e}$$

$$\Delta\dot{u} = -\Gamma\left(C\Delta P(t) + B\Delta\omega(t)\right) \tag{6.18f}$$

where T_g, T_t, T, Γ, B are diagonal gain matrices, R is the diagonal matrix of droop parameters, M, D are diagonal matrices of generator parameters, and C is the $(N+1)\times M$ incidence matrix.

Consider a step change in uncontrollable injection where $\Delta\sigma(t)$ changes at time 0 from the 0 vector to a constant vector $\Delta\sigma$. We say that $x^* := \left(\Delta u^*, \Delta\omega^*, \Delta P^*, \Delta\theta^*, \Delta a^*, \Delta p^{M*}\right)$ is an *equilibrium point* of (6.18) driven by the step change $\Delta\sigma$ if, at x^*,

$$\Delta\dot{u} = \Delta\dot{\omega} = \Delta\dot{a} = \Delta\dot{p}^M = 0$$

Note that we do not require $\Delta\dot{\theta} = 0$ in the definition of equilibrium point. The next result proves that indeed the objectives of the secondary control are achieved (its proof is left as Exercise 6.3).Furthermore $\Delta\dot{\theta} = \Delta\omega^* = 0$ in equilibrium when frequency deviation is driven to zero.

Theorem 6.2 (Steady state of secondary control)**.** Suppose the network is connected. If x^* is an equilibrium point of (6.18) driven by a step change $\Delta\sigma$ then:

1 Frequencies are restored to ω^0 and $\Delta\omega^* = 0$.
2 Line flows are restored to their scheduled values P^0 and $\Delta P^* = 0$.
3 Disturbances are compensated for locally at each bus $\Delta u^* + \Delta\sigma = 0$.

6.4 Pricing electricity and reserves

In previous sections we focus on control mechanisms to balance power at timescales from subseconds to a day. The day-ahead and real-time markets determine not only generation schedules, but also electricity prices. In this section we derive properties of electricity prices and show that they incentivize optimal dispatch, even in contingencies.

6.4.1 DC power flow model

Consider a power network modeled by the DC power flow model summarized here (see Chapter 4.6.2 for details). The network is represented by a connected graph $G = (\overline{N}, E)$ of $N+1$ nodes and $M := |E|$ lines where $\overline{N} := \{0\} \cup N$, $N := \{1, 2, \ldots, N\}$ and $E \subseteq \overline{N} \times \overline{N}$. We assume there are no self-loops, i.e., $(j, j) \notin E$ for any $j \in N$. We endow the graph with an arbitrary orientation and we refer to a line in E by (j, k), $j \sim k$, or $j \to k$ interchangeably. With respect to this graph orientation, let C denote the $(N+1) \times M$ incidence matrix defined in (4.11) and reproduced here:

$$C_{jl} = \begin{cases} 1 & \text{if } l = j \to k \text{ for some bus } k \\ -1 & \text{if } l = i \to j \text{ for some bus } i \\ 0 & \text{otherwise} \end{cases}$$

Each line $l := (j, k) \in E$ is parametrized by its susceptance $b_l > 0$. Let $B := \text{diag}(b_l, l \in E) > 0$ be the diagonal matrix of (weighted) line susceptances. The Laplacian matrix L associated with G is defined to be

$$L := CBC^{\mathsf{T}} \tag{6.19}$$

The $(N+1) \times (N+1)$ Laplacian matrix L is real symmetric with zero row and column sums. Since the network is connected, $\text{rank}(L) = N$ and its null space is $\text{span}(\mathbf{1})$. Properties of L are studied in Chapter 4.6.1.

We assume without loss of generality that there is a single controllable unit j at each bus j (including multiple units at the same bus is straightforward). Let p_j represent the net real power injections at buses j. A unit can be a generator, a load, or a prosumer that can both generate and consume. We will sometimes call j a generator bus if $p_j > 0$ and a load bus if $p_j < 0$, even though the unit at bus j can be a prosumer. The real power flows P on the lines induced by the nodal injections p are given by

$$P = BC^{\mathsf{T}} L^{\dagger} p$$

where $L^\dagger$ is the pseudo-inverse of the Laplacian matrix L. To simplify notation we define the $(N+1)\times M$ matrix S that maps line variables to nodal variables:

$$S := L^\dagger CB \qquad \text{s.t.} \qquad P = S^\mathsf{T} p \tag{6.20}$$

The matrix S has zero row sums, $\mathbf{1}^\mathsf{T} S = \mathbf{1}^\mathsf{T} L^\dagger CB = 0$ (Theorem 4.13). It maps line congestion prices to nodal congestion prices, as we will see in Chapter 6.4.3.2. The matrix S or its transpose S^T is referred to as a *shift factor*, an *injection shift factor*, or a *power transfer distribution factor*, because $P = S^\mathsf{T} p$ describes how nodal injections impact line flows. We know from Chapter 4.6.2 that (6.20) is valid if and only if the injection p satisfies

$$\mathbf{1}^\mathsf{T} p \;=\; 0 \tag{6.21}$$

In our context this means that supply and demand must be balanced.

6.4.2 Economic dispatch and LMP

As noted in Remark 6.3 a simple OPF problem, called DC OPF or economic dispatch, is solved every 5-15 minutes using the DC power flow model. We now formulate this problem and design electricity prices called locational marginal prices.

Let $f_j(p_j)$ denote the cost function of unit j, i.e., $f_j(p_j)$ models the generation cost at a generator bus with $p_j \geq 0$ and $-f_j(p_j)$ models the utility of consuming $-p_j \geq 0$ at a load bus. We assume f_j are differentiable. For a generator j, $f_j'(p_j)$ represents the marginal cost at production level p_j whereas for a load bus, $f_j'(p_j)$ represents the marginal utility at consumption level p_j. To simplify exposition we often do not differentiate between a generator and a load in which case we will refer to $f_j'(p_j)$ as the marginal cost. Let $p_j^{\min} < p_j^{\max}$ be the generation/consumption limits. Let $p := (p_j, j \in \overline{N})$ and $(p^{\min}, p^{\max}) := (p_j^{\min}, p_j^{\max}, \in \overline{N})$.

Welfare maximization.

The problem of economic dispatch is to schedule generation and consumption levels p that minimize the total dispatch cost $\sum_j f_j(p_j)$ subject to three constraints. The power must be balanced as required in (6.21). The generation or consumption levels must respect their capacity limits:

$$p^{\min} \;\leq\; p \;\leq\; p^{\max}$$

Finally the power flow P_{jk} on each line $j \to k \in E$ is directional (i.e, $P_{jk} < 0$ means power flows from buses k to j). There are line capacity limits $P_{jk}^{\min} < 0 < P_{jk}^{\max}$ in each direction and the line flows $P = S^\mathsf{T} p$ induced by p must lie within line limits:

$$P^{\min} \;\leq\; P \;=\; S^\mathsf{T} p \;\leq\; P^{\max}$$

Economic dispatch is the following problem that chooses p to minimize the total dispatch cost subject to capacity limits, nodal power balance, and line limits:

$$\min_{P^{\min} \le p \le P^{\max}} \quad \sum_{j \in \overline{N}} f_j(p_j) \tag{6.22a}$$

$$\text{s.t.} \quad \mathbf{1}^{\mathsf{T}} p = 0 \qquad [\gamma] \tag{6.22b}$$

$$P^{\min} \le S^{\mathsf{T}} p \le P^{\max} \qquad [\kappa^-, \kappa^+] \tag{6.22c}$$

where S is defined in (6.20). This problem is also called a social welfare optimization. The dispatch variable p in (6.22) is called a primal variable. Associated with the scalar constraint (6.22b) is a scalar dual variable or Lagrange multiplier $\gamma \in \mathbb{R}$. Similarly, associated with the pair of vector constraints in (6.22c) is a pair of vector dual variables or Lagrange multipliers $(\kappa^-, \kappa^+) \in \mathbb{R}^{2M}$. We use κ to denote the difference $\kappa := \kappa^- - \kappa^+$. When there is no danger of confusion we also use κ to denote the pair $\kappa := (\kappa^-, \kappa^+)$ depending on the context.

Locational marginal price λ^*.

Given any dual variable (γ, κ) define the $(N+1)$-vector:

$$\lambda \;:=\; \lambda(\gamma, \kappa) := \gamma \mathbf{1} + S\kappa \;\in\; \mathbb{R}^{N+1} \tag{6.23}$$

where $\kappa := \kappa^- - \kappa^+$ and $S := L^{\dagger} C B$. The system operator solves (6.22) to determine an optimal dispatch p^* and an associated (dual optimal) Lagrange multiplier (γ^*, κ^*). It computes $\lambda^* := \lambda(\gamma^*, \kappa^*)$ based on the Lagrange multiplier. The vector λ^* is called a *locational marginal price* (LMP) or *nodal price* (vector), and used to price electricity: a generator that provides $p_j > 0$ amount of electricity will be paid $\lambda_j^* p_j$ by the system operator and a load that consumes $-p_j > 0$ amount of electricity will pay $-\lambda_j^* p_j$ to the system operator. Besides setting the energy prices λ^*, in many North American markets, the system operator also makes binding dispatch decisions, i.e., unit j will be required to generate/consume the amount p_j^* obtained from the socially optimal p^*. In other markets, however, units may make their own injection decisions p and pay the LMPs λ^*. As we will see these two approaches are equivalent in theory because the LMP λ^* is incentive compatible, i.e., it is in the best interest of individual units to choose the socially optimal injections by setting $p_j = p_j^*$.

KKT condition.

We will study basic optimization theory in Chapter 8. As summarized in Chapter 6.1.2, if the cost functions f_j are convex and the economic dispatch (6.22) has a finite optimal value, then there exist optimal Lagrange multipliers $(\gamma^*, \kappa^{-*}, \kappa^{+*})$ and hence an LMP λ^* such that a dispatch p^* is optimal for (6.22) if and only if p^* and $(\gamma^*, \kappa^{-*}, \kappa^{+*})$ satisfy the Karush-Kahn-Tucker (KKT) condition (the Slater Theorem 8.17 of Chapter 8.3.4):

1 *Primal feasibility*: $p^{\min} \le p^* \le p^{\max}$, $\mathbf{1}^\mathsf{T} p^* = 0$, $P^{\min} \le S^\mathsf{T} p^* \le P^{\max}$.
2 *Dual feasibility*: $\kappa^{-*} \ge 0$, $\kappa^{+*} \ge 0$.
3 *Stationarity*:

$$f_j'(p_j^*) \begin{cases} = \lambda_j^* & \text{if} \quad p_j^{\min} < p_j^* < p_j^{\max} \\ > \lambda_j^* & \text{only if} \quad p_j^* = p_j^{\min} \\ < \lambda_j^* & \text{only if} \quad p_j^* = p_j^{\max} \end{cases} \tag{6.24a}$$

4 *Complementary slackness*:

$$(\kappa^{-*})^\mathsf{T} \left(S^\mathsf{T} p^* \right) = 0, \ (\kappa^{+*})^\mathsf{T} \left(S^\mathsf{T} p^* - P^{\max} \right) = 0 \tag{6.24b}$$

As we will see in Chapter 6.4.3 all properties of optimal dispatch p^* and associated LMP λ^* are consequences of the DC power flow model represented by (6.19)(6.20) and the KKT condition (6.24).

Remark 6.6 (Reference buses)**.** The formulation here uses the pseudo-inverse $L^\dagger$ of the Laplacian matrix L in the shift factor $S := L^\dagger CB$, the line flow constraint (6.22c), and the LMP λ^* in (6.23). Alternatively one can designate a bus as a reference bus for injections and prices (slack bus) and a potentially different bus for voltage angles, obtain a submatrix $\hat{L}$ of L that is invertible, and define a reduced shift factor $\hat{S} := \hat{L}^{-1}\hat{C}B$ in terms of $\hat{L}^{-1}$. The choice of reference buses does not change the optimal dispatch p^* nor the LMP λ^* (but can change the Lagrange multiplier γ^*), and seems unnecessary; see Chapter 6.4.3.4. □

Example 6.2 (Two-bus network)**.** Consider two buses connected by a line with susceptance b so that

$$C := \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \qquad B := \begin{bmatrix} b \end{bmatrix} \tag{6.25}$$

At each bus j, $j = 1, 2$, suppose there are:

- A generator with a strictly convex increasing cost function $f_j(p_j) = \frac{1}{2} c_j p_j^2$ with $c_1 < c_2$ and $0 \le p_j \le p_j^{\max}$, i.e., generator 1 is cheaper than generator 2.
- A fixed and given load $d_j > 0$.

Let $p := (p_1, p_2)$ and $d := (d_1, d_2)$.

1 Compute the Laplacian L and its pseudo-inverse $L^\dagger$.
2 Write down the social welfare optimization (6.22) and the KKT condition (6.24).
3 Compute optimal dispatch p^*, LMP λ^*, and the resulting line flow P^*.

Solution. The Laplacian and its pseudo-inverse are respectively (Exercise 4.19):

$$L := CBC^\mathsf{T} = b \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \qquad L^\dagger = \frac{1}{4b} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

The social welfare maximization (6.22) problem is:

$$\min_{0\le p\le P^{\max}} \quad \sum_{j=1}^{2} f_j(p_j) \tag{6.26a}$$

$$\text{subject to} \quad \mathbf{1}^{\mathsf{T}}(p-d) \;=\; 0 \qquad [\gamma] \tag{6.26b}$$

$$-P^{\max} \;\le\; BC^{\mathsf{T}}L^{\dagger}(p-d) \;\le\; P^{\max} \qquad [\kappa^{-},\kappa^{+}] \tag{6.26c}$$

where $P^{\max} > 0$ is the line limit and the line flow P from buses 1 to 2 is

$$P \;:=\; BC^{\mathsf{T}}L^{\dagger}(p-d) \;=\; \frac{1}{2}\begin{bmatrix}1 & -1\end{bmatrix}(p-d) \;=\; \frac{1}{2}((p_1-d_1)-(p_2-d_2)) \tag{6.27}$$

The optimal dispatch p^* and Lagrange multipliers $(\gamma^*,\kappa^{-*},\kappa^{+*})$ are given by the KKT condition (6.24): primal feasibility, dual feasibility, and

$$f_j'(p_j^*) \;\begin{cases} = \lambda_j^* & \text{if} \quad 0 < p_j^* < p_j^{\max} \\ > \lambda_j^* & \text{only if} \quad p_j^* = 0 \\ < \lambda_j^* & \text{only if} \quad p_j^* = p_j^{\max} \end{cases}$$

$$S^{\mathsf{T}}(p^*-d^*) \;\begin{cases} = -P^{\max} & \text{if} \quad \kappa^* > 0 \\ = P^{\max} & \text{if} \quad \kappa^* < 0 \\ \in (-P^{\max},P^{\max}) & \text{only if} \quad \kappa^* = 0 \end{cases}$$

where $\kappa^* := \kappa^{-*} - \kappa^{+*}$ and $S := L^{\dagger}CB$. For simplicity, we will suppose $0 < p_j^* < p_j^{\max}$ so that $f_j'(p_j^*) = \lambda_j^*$.

Without congestion. If $S^{\mathsf{T}}(p^*-d^*) \in (-P^{\max},P^{\max})$, then $\kappa^* = \kappa^{-*} = \kappa^{+*} = 0$ and hence

$$\sum_j f_j'^{-1}(\gamma^*) \;=\; \sum_j d_j^*$$

which has a unique solution for γ^* since f_j are strictly convex. When $f_j(p_j) = \frac{1}{2}c_j p_j^2$ the optimal dispatch and LMPs are

$$\gamma^* \;=\; \left(\sum_j \frac{1}{c_j}\right)^{-1} \sum_j d_j^*, \qquad \kappa^* \;=\; \kappa^{-*} \;=\; \kappa^{+*} \;=\; 0, \qquad \lambda_j^* \;=\; \gamma^*$$

$$p_j^* \;=\; \frac{\gamma^*}{c_j} \;=\; \frac{1/c_j}{1/c_1+1/c_2}(d_1+d_2), \quad j=1,2 \tag{6.28}$$

i.e., the generators j share the total load d_1+d_2 in proportion to their $1/c_j$. Since $c_1 < c_2$ we have $p_1^* > p_2^*$ and $P^* > 0$.

With congestion $\tilde{\kappa}^* \neq 0$. Suppose $p_1^{\max} - d_1 > P^{\max}$. Then, since $c_1 < c_2$ (generator 1 is cheaper), the line congestion price (optimal Lagrange multiplier) $\tilde{\kappa}^{+*}$ must be strictly positive and $\tilde{\kappa}^* := \tilde{\kappa}^{-*} - \tilde{\kappa}^{+*} < 0$. Complementary slackness then implies that $\tilde{P}^* = S^{\mathsf{T}}(\tilde{p}^* - d^*) = P^{\max}$ where $\tilde{P}^*$ is given by (6.27) and $(\tilde{p}_1^*, \tilde{p}_2^*)$ is given by (6.28). Furthermore

$$\tilde{\lambda}_1^* \;=\; \tilde{\gamma}^* + \frac{1}{2}\tilde{\kappa}^*, \qquad \tilde{\lambda}_2^* \;=\; \tilde{\gamma}^* - \frac{1}{2}\tilde{\kappa}^* \tag{6.29a}$$

Therefore $f_1'(\tilde{p}_1^*) = \tilde{\lambda}_1^* < \tilde{\lambda}_2^* = f_2'(\tilde{p}_2^*)$ even though $\tilde{p}_1^*$ may be greater or smaller than $\tilde{p}_2^*$. Since $\tilde{\kappa}^* < 0$ and $\sum_j p_j^* = \sum_j \tilde{p}_j^* = \sum_j d_j$, we must have $\tilde{p}_1^* < p_1^*$ and $\tilde{p}_2^* > p_2^*$. Power balance means

$$f_1'^{-1}\left(\tilde{\lambda}_1^*\right) + f_2'^{-1}\left(\tilde{\lambda}_2^*\right) \;=\; d_1 + d_2 \tag{6.29b}$$

Substituting (6.27) into $P^* = S^{\mathsf{T}}(\tilde{p}^* - d^*) = P^{\max}$ we have

$$f_1'^{-1}\left(\tilde{\lambda}_1^*\right) - f_2'^{-1}\left(\tilde{\lambda}_2^*\right) \;=\; 2P^{\max} + (d_1 - d_2) \tag{6.29c}$$

When $f_j(p_j) = \frac{1}{2} c_j p_j^2$ we have from (6.29b)(6.29c) $A\tilde{\lambda}^* = b$ with

$$A \;=\; \begin{bmatrix} 1/c_1 & 1/c_2 \\ 1/c_1 & -1/c_2 \end{bmatrix}, \qquad b_1 \;:=\; d_1 + d_2, \qquad b_2 \;:=\; 2P^{\max} + (d_1 - d_2)$$

Therefore

$$\begin{bmatrix} \tilde{\lambda}_1^* \\ \tilde{\lambda}_2^* \end{bmatrix} \;=\; \begin{bmatrix} c_1(d_1 + P^{\max}) \\ c_2(d_2 - P^{\max}) \end{bmatrix}, \qquad \begin{bmatrix} \tilde{p}_1^* \\ \tilde{p}_2^* \end{bmatrix} \;=\; \begin{bmatrix} d_1 + P^{\max} \\ d_2 - P^{\max} \end{bmatrix}, \qquad \tilde{P}^* \;=\; P^{\max}$$

From (6.29a), we have

$$\begin{aligned} \tilde{\gamma}^* \;&=\; \frac{1}{2}(\tilde{\lambda}_1^* + \tilde{\lambda}_2^*) \;=\; \frac{1}{2}\left(c_1 d_1 + c_2 d_2 - (c_2 - c_1) P^{\max}\right) \\ \tilde{\kappa}^{+*} \;&=\; c_2 d_2 - c_1 d_1 - (c_1 + c_2) P^{\max}, \qquad \tilde{\kappa}^{-*} \;=\; 0 \end{aligned}$$

□

6.4.3 LMP properties

We now study properties of an optimal dispatch p^* and the associated LMP λ^*. These properties are derived from the optimality condition (6.24) for economic dispatch.

6.4.3.1 Competitive equilibrium

Consider the case where the system operator sets prices and allows generators and loads to freely choose their injections in a way that optimizes their own surpluses. An important justification for pricing electricity according to LMP is that an optimal dispatch and LMP (p^*, λ^*) satisfies the following properties:

1 *Market clearing*. The supply of equals the demand for power. This is ensured by (6.22b).
2 *Capacity limits*. The line flows respect their capacity constraints. This is ensured by (6.22c).
3 *Welfare optimization*. The pair (p^*, λ^*) solves the economic dispatch problem (6.22) that optimizes social welfare.

4 *Incentive compatibility*. Suppose the generators/loads are price takers, i.e., their bids will not alter the LMP computed by the system operator. Given any generation-price pair (p_j, λ_j) at bus j, if j is a generator it incurs a cost $f_j(p_j)$ and is paid $\lambda_j p_j$ whereas if it is a load it attains a utility $-f_j(p_j)$ and pays $-\lambda_j p_j$. When presented with the LMP λ_j^* it is rational for the unit j to choose its level of production/consumption so as to maximize its surplus, i.e., it chooses p_j to solve

$$\max_{p_j^{\min} \le p_j \le p_j^{\max}} \quad \lambda_j^* p_j - f_j(p_j)$$

The stationarity condition (6.24a) implies that the socially optimal dispatch p_j^* is a solution of individual surplus maximization given the LMP λ_j^*. If unit j's injection limits are not binding, then the LMP λ_j^* equals its marginal cost $f_j'(p_j^*)$ according to (6.24a); such a unit is called a marginal unit. If $\lambda_j^* > f_j'(p_j^*)$, then the LMP exceeds the marginal cost and therefore unit j generates at its peak $p_j^* = p_j^{\max}$. Similarly if the LMP is not sufficient to cover the marginal cost, $\lambda_j^* < f_j'(p_j^*)$, then unit j generates at its minimum $p_j^* = p_j^{\min}$.

Therefore LMP λ^* aligns individual optimality with social optimality in that, when units are paid or charged according to λ^*, their individual surplus-maximizing decisions p_j^* will coincide with the optimal dispatch the system operator would have chosen to optimize the social welfare (6.22). For this reason (p^*, λ^*) is also called a *competitive equilibrium*.

6.4.3.2 LMP λ^* and line congestion price κ^*

To simplify exposition we do not distinguish between generators and loads, and refer to $f_j(p_j)$ and $f_j'(p_j)$ as costs and marginal costs. The LMP λ_j^* defined in (6.23) consists of two components:

$$\lambda^* := \gamma^* \mathbf{1} + c^* := \gamma^* \mathbf{1} + S\kappa^*$$

where $\kappa^* := \kappa^{-*} - \kappa^{+*}$ and $S := L^\dagger CB$. We will call the first component γ^* the *energy price* (γ^* is also called the system λ), and the second component $c^* := S\kappa^*$ the *nodal congestion prices*, for the following reasons.

Energy price γ^*.

The first component γ^* is the same at every bus j and equals the LMP if none of the line constraints are binding so that $\kappa_l^{-*} = \kappa_l^{+*} = 0$. In that case $\lambda_j^* = \gamma^* = f_j'(p_j^*)$ at all marginal units j where their generation capacities are not binding. If f_j are nondecreasing, when the network is not congested, the LMP $\lambda_j^* \ge 0$ are always nonnegative and the same at every bus. In this case all marginal units j produce (consume) at their common marginal costs (marginal utilities) $f_j'(p_j^*) = \gamma^*$. More generally, $\gamma^* = (N+1)^{-1} \sum_j \lambda_j^*$ is the average LMPs across the network since $\mathbf{1}^\mathsf{T} L^\dagger = 0$ (Theorem 4.13).

Line congestion price $\kappa^* := \kappa^{-*} - \kappa^{+*}$.

To understand the second component c^* of LMP, we first interpret $\kappa_l^* := \kappa_l^{-*} - \kappa_l^{+*}$ as the *line congestion price* or *shadow price* at $l \in E$, for two reasons. First it is the marginal value of relaxing the line capacities $(P^{\min}, P^{\max})$: if we denote by $f^*(P^{\min}, P^{\max})$ the optimal value of the economic dispatch problem (6.22) as a function of $(P^{\min}, P^{\max})$ then (see Chapter 8.3.5)

$$\frac{\partial f^*}{\partial P_l^{\min}}(P^{\min}, P^{\max}) \;=\; \kappa_l^{-*}, \qquad \frac{\partial f^*}{\partial P_l^{\max}}(P^{\min}, P^{\max}) \;=\; -\kappa_l^{+*}$$

i.e., κ_l^{-*} is approximately the increase in the optimal dispatch cost f^* if the lower line limit $P_l^{\min}$ is increased (tightened) by 1 unit; and κ_l^{+*} is the reduction in f^* if $P_l^{\max}$ is increased (relaxed) by 1 unit. These prices $(\kappa_l^{-*}, \kappa_l^{+*})$ are nonnegative and at most one of them can be strictly positive due to complementary slackness. They provide a valuation for the line capacities $(P_l^{\min}, P_l^{\max})$ in the sense that each additional unit of line capacities will *reduce* the optimal cost f^* by $(\kappa_l^{-*}, \kappa_l^{+*}) \geq 0$ respectively. We therefore refer to both the pair $(\kappa^{-*}, \kappa^{+*})$ and $\kappa^* := \kappa^{-*} - \kappa^{+*}$ as line congestion prices.

Second, recall that the line flows are $P = S^{\mathsf{T}} p$. Since the summands in (6.24b) are all nonpositive we have

$$\kappa_l^{-*}\left(P_l^{\min} - P_l^*\right) \;=\; 0, \qquad \kappa_l^{+*}\left(P_l^* - P_l^{\max}\right) \;=\; 0, \qquad l \in E$$

Complementary slackness (6.24b) implies that κ_l^* is zero if line flow P_l^* is strictly within its capacity limits $(P_l^{\min}, P_l^{\max})$. If $\kappa_l^* = -\kappa_l^{+*} < 0$ then $P_l^* = P_l^{\max} > 0$ reaches the line capacity in the direction for which P_l is defined. If $\kappa_l^* = \kappa_l^{-*} > 0$ then $P_l^* = P_l^{\min} < 0$ reaches the line capacity in the opposite direction. Therefore the product $-\kappa_l^* P_l^*$ is always nonnegative at optimality. We will therefore interpret $-\kappa_l^* P_l^* \geq 0$ as the cost of carrying line flow P_l^* on line l.

Nodal congestion price $c^* := S\kappa^*$.

This leads to the following justification for treating $c^* := S\kappa^*$ as the nodal congestion prices. Since $P = S^{\mathsf{T}} p$, the shift factor $S^{\mathsf{T}} = \frac{\partial P}{\partial p}$ describes the increases in line flows for each additional units of nodal injections. Suppose the injection at bus j is increased by Δp_j. This increases the line flow at line l by $S_{jl}\Delta p_j$, and thus increases the line congestion cost at line l by $-\kappa_l^*(S_{jl}\Delta p_j)$. This means that each additional Δp_j of *injection* at j increases the congestion cost over the network by $-\sum_l S_{jl}\kappa_l^* \Delta p_j$, or equivalently, each additional Δp_j of *withdrawal* (load) at j increases the congestion cost over the network by $\left(\sum_l S_{jl}\kappa_l^*\right)\Delta p_j$. We can therefore interpret $c_j^* := \sum_l S_{jl}\kappa_l^*$ as the nodal congestion price, the price of serving an additional unit of load from bus j. It is in this sense that we say the matrix S maps the line congestion price κ^* to the nodal congestion price c^*.

Negative LMP $\lambda_j^* < 0$.

The LMP $\lambda_j^* = \gamma^* + c_j^*$ is the sum of the energy price and the nodal congestion price. Since the nodal congestion price c_j^* of serving a load at bus j can be positive or negative, the LMP at bus j may be negative in which case a load is paid to consume or a generator pays to produce at bus j. In addition to line congestion, LMP λ_j^* can also be negative due to generation limits $(p^{\min}, p^{\max})$. In practice it is not uncommon for LMP to become negative, e.g., during the day time in California when there is excess solar generation.

6.4.3.3 LMP λ^* and merchandizing surplus

The system operator collects a payment $\lambda_j^*(-p_j^*)$ from every load j and pays $\lambda_j^* p_j^*$ to every generator j. The residue

$$\text{MS} := -\sum_j \lambda_j^* p_i^* \;=\; -(\lambda^*)^{\mathsf{T}} p^* \tag{6.30}$$

is called the *merchandizing surplus*. It is left-over money with the system operator. Substitute $\lambda^* = \gamma^* \mathbf{1} + S\kappa^*$ into (6.30) one obtains (Exercise 6.7):

$$\text{MS} = \left(\kappa^{+*}\right)^{\mathsf{T}} P^{\max} \;+\; (\kappa^{-*})^{\mathsf{T}} (-P^{\min}) \tag{6.31}$$

Recall that $P_l^{\min} < 0 < P_l^{\max}$ on each line $l \in E$ and $(\kappa^{-*}, \kappa^{+*}) \geq 0$. This means that every term on the right-hand side of (6.31) is nonnegative. Therefore MS≥ 0, i.e., the system operator will not run cash negative. This is called *revenue adequacy*. Moreover MS $= 0$ if and only if $\kappa_l^{-*} = \kappa_l^{+*} = 0$, i.e., if and only if there is no congestion in the network.

The congestion price $(\kappa_l^{-*}, \kappa_l^{+*})$ induces a value $\kappa_l^{+*} P_l^{\max} + \kappa_l^{-*}(-P_l^{\min}) \geq 0$ on the line capacity $(P_l^{\min}, P_l^{\max})$, explained in Chapter 6.4.3.2. This value is called the *congestion rent of line* $l \in E$. The relation (6.31) says that MS is equal to the congestion rent over the entire network. The MS is therefore also called the *congestion rent*. Since the system operator is non-profit the MS is distributed to market participants as financial transmission rights.

Using $p^* = CP^*$ we can also express the MS in terms of optimal line flows P^* and the difference in LMP at each end of a line:

$$\text{MS} = -(\lambda^*)^T CP^* \;=\; \sum_{j \to k \in E} \left(\lambda_k^* - \lambda_j^*\right) P_{jk}^*$$

One might think that P_{jk} on line (j,k) always flows from the bus with a lower LMP towards one with a higher LMP, but this is not always the case. Recall that line flows are directional with a fixed but arbitrary direction and hence if P_{jk} is defined then P_{kj} is not a variable in our model. The summand above consists of the LMP difference that is opposite to the direction in which P_{jk} is defined. Therefore, on each line $j \to k$,

if $\left(\lambda_k^* - \lambda_j^*\right) P_{jk}^* > 0$ then power flows towards the node with a higher LMP, but if $\left(\lambda_k^* - \lambda_j^*\right) P_{jk}^* < 0$ then power flows towards the node with a lower LMP.

6.4.3.4 LMP λ^* and price reference bus

In the literature a particular bus r is sometimes designated as the *price reference bus* or a *slack bus* where it is assumed that injections p_{-r} at all other buses can be arbitrary and are always balanced by the injection $p_r := -\mathbf{1}^\mathsf{T} p_{-r}$ at the price reference bus r. This is often a bus with a large generator with many lines connecting the bus to the rest of the grid so local congestion is rare. We still assume bus 0 is the reference bus for voltage angles, i.e., $\theta_0 := 0$. The price reference bus r may or may not be bus 0 (we assume $r = 0$ in Chapter 4.6.2 on the DC power flow model). The DC power flow equations can be rewritten in terms of the injections p_{-r} at non-price reference buses. It is important to keep in mind that this set of equations depends on the choice of the price reference bus r. We show in Theorem 6.3 below, however, that the optimal dispatch and LMP (p^*, λ^*) do not.

To write DC power flow equations in terms of the injections p_{-r} at non-price reference buses, let c_0^T and c_r^T denote the rows corresponding to the angle reference bus 0 and the price reference bus r respectively, and C_{-0} and C_{-r} denote the remaining submatrices after removing c_0^T and c_r^T respectively form C. We will refer to them as row 0 and row r, but for convenience they may not appear as the first or rth row of C, i.e., we may write C as (after possibly rearranging/relabeling rows):

$$C \;=:\; \begin{bmatrix} c_0^\mathsf{T} \\ C_{-0} \end{bmatrix} \;=:\; \begin{bmatrix} C_{-r} \\ c_r^\mathsf{T} \end{bmatrix}$$

(Instead of $\hat{C}$ as in Chapter 4.6.2, we write C_{-0} here to emphasize the symmetry in angle and price reference buses.) Rewrite the DC power flow equation (4.55b) as (after possibly rearranging/relabeling rows):

$$\begin{bmatrix} p_{-r} \\ p_r \end{bmatrix} \;=\; \begin{bmatrix} C_{-r} \\ c_r^\mathsf{T} \end{bmatrix} P, \qquad\qquad P \;=\; B \begin{bmatrix} c_0 & C_{-0}^\mathsf{T} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_{-0} \end{bmatrix} \tag{6.32}$$

Since $\theta_0 := 0$ by definition, we have the DC power flow model in terms of (C_{-0}, C_{-r}):

$$p_{-r} \;=\; C_{-r} P, \qquad\qquad P \;=\; B C_{-0}^\mathsf{T} \theta_{-0}$$

yielding the relationship in terms of the $N \times N$ matrix $L_r := C_{-r} B C_{-0}^\mathsf{T}$:

$$p_{-r} \;=\; \left(C_{-r} B C_{-0}^\mathsf{T}\right) \theta_{-0} \;=:\; L_r \theta_{-0} \tag{6.33}$$

The matrix L_r can also be obtained from the Laplacian matrix $L := CBC^\mathsf{T}$ by removing the column of L corresponding to bus 0 and its row corresponding to bus r. It is not a principal submatrix of L unless $r = 0$ and hence L_0 is symmetric but L_r is generally not. While any strict principal submatrix of L is nonsingular (Theorem 4.13), L_r may be singular. This is the main disadvantage of this model.

Assuming $L_r := C_{-r}BC_{-0}^{\mathsf{T}}$ in (6.33) is nonsingular. Then, given any injections p_{-r} at non-reference buses, the line flows are given by

$$P \;=\; \left(BC_{-0}^{\mathsf{T}}L_r^{-1}\right)p_{-r} \;=:\; S_r^{\mathsf{T}}p_{-r} \tag{6.34}$$

The matrix $S_r := L_r^{-\mathsf{T}}C_{-0}B$ is also referred to as a *shift factor* and it depends on the choice of the price reference bus and the nonsingularity of L_r. The line flows P, however, do not depend on the choice of r, i.e., $P = S_r^{\mathsf{T}}p_{-r} = S^{\mathsf{T}}p$ where $S := L^{\dagger}CB$ defined in (6.20) (see Exercise 6.8). The expression (6.34) generalizes the expression $P = BC_{-0}^{\mathsf{T}}L_0^{-1}p_{-0}$ in Lemma 4.14 of Chapter 4.6.2 which assumes that $r = 0$. We now show that the economic dispatch (6.22) can be reformulated in terms of L_r^{-1} instead of $L^{\dagger}$, but that the optimal dispatch and LMP (p^*, λ^*) turn out to be independent of the choice of r.

Substituting (6.34) into (6.22), economic dispatch is equivalent to:

$$\min_{p^{\min}\le p\le p^{\max}} \quad \sum_{j\in N} f_j(p_j) \tag{6.35a}$$

$$\text{subject to} \quad \mathbf{1}^{\mathsf{T}}p \;=\; 0 \qquad [\gamma] \tag{6.35b}$$

$$P^{\min} \;\le\; S_r^{\mathsf{T}}p_{-r} := BC_{-0}^{\mathsf{T}}L_r^{-1}p_{-r} \;\le\; P^{\max} \qquad [\kappa^-, \kappa^+] \tag{6.35c}$$

with associated Lagrange multipliers $(\gamma, \kappa^-, \kappa^+) \in \mathbb{R}^{1+2M}$ with $\kappa^- \ge 0$, $\kappa^+ \ge 0$. The difference between (6.35) with a price reference bus and (6.22) without is in the line limit expression (6.35c). Since the line flow P is independent of r, we expect the optimal dispatch p^* to remain the same; the exact relation between these two formulations are clarified in Theorem 6.3. Given an optimal Lagrange multiplier vector $(\gamma^*, \kappa^{-*}, \kappa^{+*})$, the LMP is given by

$$\lambda^* \;:=\; \begin{bmatrix}\lambda^*_{-r}\\ \lambda^*_r\end{bmatrix} \;:=\; \gamma^*\mathbf{1} + \begin{bmatrix}S_r\kappa^*\\ 0\end{bmatrix} \tag{6.36}$$

where $\kappa^* := \kappa^{-*} - \kappa^{+*}$. It can be shown that a dispatch p^* and a Lagrange multiplier $(\gamma^*, \kappa^{-*}, \kappa^{+*})$ are optimal for (6.36) and its dual problem if and only if $(p^*, \gamma^*, \kappa^{-*}, \kappa^{+*})$ satisfies the KKT condition (6.24), with the line flow $S^{\mathsf{T}}p^*$ in the primal feasibility condition and the complementary slackness condition (6.24b) replaced by $S_r^{\mathsf{T}}p^*_{-r}$ (the Slater Theorem 8.17 of Chapter 8.3.4).

The choice of the reference bus r does not affect the optimal dispatch or LMP (p^*, λ^*), though it may affect the values of $(\gamma^*, \kappa^{-*}, \kappa^{+*})$. Specifically the next result implies that a dispatch p^* is optimal for (6.35) with a price reference bus r if and only if p^* is optimal for (6.22) without designating a price reference bus. Moreover their associated LMPs are equal. This is a consequence of the key fact that line flows are independent of r, $P = S_r^{\mathsf{T}}p_{-r} = S^{\mathsf{T}}p$. See Exercise 6.8 for a proof.

Theorem 6.3 (Arbitrary choice of r)**.** Suppose the cost functions f_j are convex (and hence differentiable) so that the KKT condition (6.24) is necessary and sufficient for optimality for both (6.35) and (6.22). Fix a dispatch p^*. Consider two sets of dual

variables $(\gamma^*, \kappa^{-*}, \kappa^{+*})$ and $(\tilde{\gamma}^*, \tilde{\kappa}^{-*}, \tilde{\kappa}^{+*})$ that satisfy

$$\tilde{\gamma}^* = \gamma^* - s_r^\mathsf{T}\kappa^*, \qquad \tilde{\kappa}^{-*} = \kappa^{-*}, \qquad \tilde{\kappa}^{+*} = \kappa^{+*} \tag{6.37}$$

where $\kappa^* := \kappa^{-*} - \kappa^{+*}$ and s_r^T denotes row r of $S := L^\dagger CB$.

1 Let $\tilde{\lambda}^* := \tilde{\gamma}^*\mathbf{1} + S\tilde{\kappa}^*$ and λ^* be defined in (6.36). Then $\tilde{\lambda}^* = \lambda^*$.
2 The dispatch p^* and $(\tilde{\gamma}^*, \tilde{\kappa}^{-*}, \tilde{\kappa}^{+*})$ satisfy the KKT condition (6.24) if and only if p^* and $(\gamma^*, \kappa^{-*}, \kappa^{+*})$ satisfy (6.24) with the line flow $S^\mathsf{T}p^*$ in the primal feasibility condition and the complementary slackness condition (6.24b) replaced by $S_r^\mathsf{T}p^*_{-r}$.

Theorem 6.3 is illustrated in Exercise 6.10. Its implications are collected in the following remark.

Remark 6.7 (Theorem 6.3: implications)**.** 1 The shift factor $S_r := L_r^{-T}C_{-0}B$ in (6.34) with a price reference bus r and $S := L^\dagger CB$ in (6.20) without a price reference bus are related as follows (Exercise 6.8):

$$\begin{bmatrix} S_r \\ 0 \end{bmatrix} = S - \mathbf{1}_{N+1}s_r^\mathsf{T}, \qquad S_r = [S]_{-r} - \mathbf{1}_N s_r^\mathsf{T}$$

where s_r^T denotes row r of S, $[S]_r$ denotes the submatrix of S obtained by removing row r, and $\mathbf{1}_n$ is the vector of all 1s of size n. Recall that each row j of S is the marginal increase in all line flows due to an additional injection Δp_j at bus j. By designating a price reference (slack) bus r, we renormalize the shift factor S_r so that its row j is now the marginal increase due to an additional increase at j, in excess of the marginal increase s_r^T due to an additional injection at r. This underlies the relation (6.37) between the two sets of prices.

2 The LMP $\lambda_r^* = \gamma^*$ in (6.36) at the reference bus r is generally not the energy price discussed in Chapter 6.4.3.2, but the average LMP $\mathbf{1}^\mathsf{T}\lambda^*(N+1)^{-1}$ is; see Exercise 6.9.

3 The main disadvantage of formulating the economic dispatch and LMP with a designated price reference bus r is that the submatrix L_r is not a principal submatrix of the Laplacian L and therefore may be singular (unless $r = 0$, i.e., the price reference bus is the same as the angle reference bus). The resulting DC power flow equations and the shift factor S_r will depend on the choice of r and the nonsingularity of L_r. In contrast the DC power flow model (4.55c) in terms of $L^\dagger$ and the shift factor S in (6.20) do not. Furthermore the LMP λ^* decomposes into an energy price $\tilde{\gamma}^*$ and congestion prices $c^* := S\tilde{\kappa}^*$, but not in terms of (γ^*, κ^*) in (6.36).

□

6.4.4 Security constrained economic dispatch

There are two techniques to deal with uncertainties, both discrete uncertainties due to outages of generators, transmission or distribution lines and transformers, and continuous uncertainties due to random fluctuations of renewable generations or loads. The first is to commit and dispatch generation resources or controllable loads to balance deterministic forecasts of supply and demand and deal with uncertainty through reserves requirements where a certain amount of generation capacity is set aside to handle contingencies or random fluctuations of supply and demand; e.g., the total reserve amount is greater than the capacity of the largest generator in the system or the maximum dispatch amount. The second technique is two-stage stochastic optimization with recourse where random scenarios are explicitly taken into account in dispatch decisions, in the form of a security constrained OPF discussed in Chapter 6.2.3. In this subsection we extend the economic dispatch problem of Chapter 6.4.2 to security constrained economic dispatch that jointly optimizes energy and reserves for each scenario.

6.4.4.1 Joint energy and reserve optimization

Suppose the uncontrollable generation and demand are uncertain and take one of K values $(g_k, d_k) \in \mathbb{R}_+^{2(N+1)}$ with probability $w_k > 0$ such that $\sum_{k=1}^K w_k = 1$. Suppose each unit j can decide not just a dispatch p_j before (g,d) is realized, but also an adjustment r_{kj} if (g_k, d_k) is later realized so that the actual injection at delivery time is $p_j + r_{kj}$ in scenario k. Unit j must reserve some down and up reserve capacities $(r_j^{\min}, r_j^{\max})$ in the first stage for its adjustment r_{kj} in the second stage so that

$$r_j^{\min} \le r_{kj} \le r_j^{\max}, \quad p_j^{\min} \le p_j + r_j^{\min} \le p_j + r_j^{\max} \le p_j^{\max}, \quad j \in \overline{N} \tag{6.38a}$$

The first-stage decision consists of the dispatch p and reserve capacities $r^{\min} := \left(r_j^{\min}, j \in \overline{N}\right)$ and $r^{\max} := \left(r_j^{\max}, j \in \overline{N}\right)$, but this decision must be made taking into account of the second-stage actions $r_k := (r_{kj}, j \in \overline{N})$ for each scenario $k = 1, \ldots, K$. We will formulate this as a two-stage stochastic program with recourse. In a typical application, this program is solved before (g,d) is realized for both the first-stage decision $(p, r^{\min}, r^{\max})$ and the second-stage decisions $(r_k, \forall k)$ in order to produce an optimal schedule in advance. After (g,d) is realized, the optimal action r_k can then be applied if $(g,d) = (g_k, d_k)$.

Besides (6.38a) suppose there is also a system-wide reliability requirement on the reserves $(r^{\min}, r^{\max})$ imposed by the system operator. For example, a popular reserve requirement is that the total reserve must be sufficient to cover the outage of the largest generating unit, i.e., $\sum_{j \neq j_k} r_j^{\min} \ge \max_j p_i^{\max}$ where $j_k := \arg\max_j p_j^{\max}$. We assume the reliability requirement in each scenario k depends only on $(r^{\min}, r^{\max})$, not on the

dispatch p, and is separable in j, i.e., it is of the form:

$$h_k(r^{\min}, r^{\max}) \;:=\; \sum_j h_{kj}(r_j^{\min}, r_j^{\max}) \;\geq\; 0 \tag{6.38b}$$

where $h_{kj} : \mathbb{R}^2 \to \mathbb{R}$. [6] For the example above, $h_{kj}(r_j^{\min}, r_j^{\max}) = r_j^{\min} - \alpha_j p_{j_k}^{\max}$ for $j \neq j_k$ with $\alpha_j \geq 0$ and $\sum_{j \neq j_k} \alpha_j = 1$, i.e., α_j is the fraction of the largest possible capacity lost $p_{jk}^{\max}$ that unit j can provide in scenario k. In general $h_{kj}(r_j^{\min}, r_j^{\max})$ can be positive or negative. The capacity and reserve constraints (6.38a) are decentralized across j, but the systemwide reliability requirement (6.38b) couples their reserve decisions $(r_j^{\min}, r_j^{\max})$.

Suppose the cost for unit j to provide $p_j + r_{kj}$ amount of energy is $f_{kj}(p_j + r_{kj})$ if scenario k materializes. Then the joint energy and reserve optimization, called security constrained economic dispatch, is the following two-stage optimization with recourse:

$$\min_{p, r^{\min}, r^{\max}} \quad \sum_{k=1}^{K} w_k \, Q_k(p, r^{\min}, r^{\max}) \tag{6.39a}$$

$$\text{s.t.} \quad p^{\min} \leq p + r^{\min}, \;\; p + r^{\max} \leq p^{\max} \qquad [\alpha^-, \alpha^+] \tag{6.39b}$$

$$h_k\left(r^{\min}, r^{\max}\right) := \sum_j h_{kj}\left(r_j^{\min}, r_j^{\max}\right) \;\geq\; 0 \qquad [\mu_k] \tag{6.39c}$$

where, for each $k = 1, \ldots, K$, Q_k solves the economic dispatch in scenario k:

$$Q_k(p, r^{\min}, r^{\max}) \;:=\; \min_{r_k} \; f_k(p + r_k) \;:=\; \sum_j f_{kj}(p_j + r_{kj}) \tag{6.39d}$$

$$\text{s.t.} \;\; \mathbf{1}^{\mathsf{T}}(p + r_k + g_k - d_k) \;=\; 0 \qquad [\gamma_k] \tag{6.39e}$$

$$P^{\min} \;\leq\; S^{\mathsf{T}}(p + r_k + g_k - d_k) \;\leq\; P^{\max} \qquad [\kappa_k^-, \kappa_k^+] \tag{6.39f}$$

$$r^{\min} \;\leq\; r_k \;\leq\; r^{\max} \qquad [\beta_k^-, \beta_k^+] \tag{6.39g}$$

The cost in (6.39a) is the expected optimal second-stage cost Q_k. The constraints (6.39b)(6.39c) on the first-stage decisions $(p, r^{\min}, r^{\max})$ do not involve any uncertainty. For each scenario k, the second-stage problem (6.39d)–(6.39f) optimizes the reserve decision r_k in response to the random realization of (g_k, d_k), given a first-stage decision $(p, r^{\min}, r^{\max})$. It is the same as economic dispatch (6.22) with reserve capacity constraints, power balance and line limits.

The second-stage problems $Q_k(p, r^{\min}, r^{\max})$ are separable in k. We can therefore interchange expectation and minimization over r_k and write (6.39) as a single-stage

[6] A less stringent requirement is to have enough reserve to cover the outage of the largest *dispatched* generating unit, i.e., $\sum_{j \neq j_k} r_j^{\min} \geq \max_j p_i$ where $j_k := \arg\max_j p_j$. The formulation and results here extend to the case where the dispatch decision p and the reserve decisions $(r^{\min}, r^{\max})$ are coupled.

program:

$$\min_{\substack{p,r^{\min},r^{\max}\\(r_k,k\geq 1)}} \quad \sum_k w_k f_k(p+r_k) := \sum_k w_k \sum_j f_{kj}(p_j+r_{kj}) \tag{6.40a}$$

$$\text{s.t.} \quad (6.39b)(6.39c)(6.39e)-(6.39g) \tag{6.40b}$$

Denote the primal and dual variables for (6.40) by

$$\begin{aligned} x^* &:= (p^*, r^{\min *}, r^{\max *}, r_k^*, k \geq 1) \\ \xi^* &:= (\gamma_k, \kappa_k^{-*}, \kappa_k^{+*}, \alpha^{-*}, \alpha^{+*}, \beta_k^{-*}, \beta_k^{+*}, \mu_k^*, k \geq 1) \end{aligned}$$

Let $\kappa_k^* := \kappa_k^{-*} - \kappa_k^{+*}$, $\alpha^* := \alpha^{-*} - \alpha^{+*}$, and $\beta_k^* := \beta_k^{-*} - \beta_k^{+*}$. Define the LMP λ_k^* for each scenario k (cf. (6.23)):

$$\lambda_k^* := \gamma_k^* \mathbf{1} + S\kappa_k^* \tag{6.41}$$

We assume all functions f_{kj}, h_{kj} are real-valued, convex and continuously differentiable and the parameters are appropriately chosen such that (6.40) has a finite optimal value, and the Slater condition is satisfied, e.g., $p^{\min} < p^{\max}$. Then the Slater Theorem 8.17 of Chapter 8.3.4 implies that optimal Lagrange multipliers ξ^* and hence LMPs $(\lambda_k^*, \forall k)$ always exist. Moreover a primal-dual feasible (x^*, ξ^*) is primal-dual optimal for (6.40) if and only if (x^*, ξ^*) satisfies stationarity:

$$w_k \nabla f_k(p^* + r_k^*) = \lambda_k^* + \beta_k^*, \quad \sum_k \mu_k^* \nabla h_k \left(r^{\min *}, r^{\max *}\right) = 0, \quad \alpha^* = \sum_k \beta_k^* \tag{6.42}$$

complementary slackness for decentralized constraints:

$$(\alpha^{-*})^{\mathsf{T}} \left(p^{\min} - p^* - r^{\min *}\right) = 0, \qquad (\alpha^{+*})^{\mathsf{T}} (p^* + r^{\max *} - p^{\max}) = 0 \tag{6.43a}$$

$$(\beta_k^{-*})^{\mathsf{T}} \left(r^{\min *} - r_k^*\right) = 0, \qquad (\beta_k^{+*})^{\mathsf{T}} \left(r_k^* - r^{\max *}\right) = 0 \tag{6.43b}$$

and that for coupling constraints (Exercise 6.11):

$$\mu_k^* h_k \left(r^{\min *}, r^{\max *}\right) = 0 \tag{6.43c}$$

$$(\kappa^{-*})^{\mathsf{T}} \left(P^{\min} - S^{\mathsf{T}}(p^* + r_k^* + g_k - d_k)\right) = 0 \tag{6.43d}$$

$$(\kappa^{+*})^{\mathsf{T}} \left(S^{\mathsf{T}}(p^* + r_k^* + g_k - d_k) - P^{\max}\right) = 0 \tag{6.43e}$$

The stationarity condition (6.42) has three implications. First the probability-weighted marginal cost $w_k \nabla f_k$ is the sum of LMP λ_k^* plus the "reserve capacity price" β_k^* in the second stage. Moreover the "reserve capacity price" α^* in the first stage (which is independent of scenarios k) is simply the sum of the reserve capacity prices β_k^*. Finally the total marginal reliability cost $\sum_k \mu_k^* \nabla h_k \left(r^{\min *}, r^{\max *}\right)$ is zero. Interestingly complementary slackness (6.43c) says that the total reliability cost is also zero (we will return to this point shortly).

6.4.4.2 ICRA settlement rule

For economic dispatch (6.22) without uncertainty, it is desirable to price electricity using the Lagrange multipliers (γ^*, κ^*) associated with coupling constraints (power balance and line limits) because they price the externalities caused by units j and align individual optimality with social optimality (see Chapter 6.4.3). We apply the same intuition to the two-stage problem (6.40) and design prices using the Lagrange multipliers associated only with the coupling constraints, power balance (6.39e), line limits (6.39f), as well as the systemwide reliability requirement (6.39c).

Let (x^*, ξ^*) be a primal-dual optimal solution of (6.40) and λ_k^* be the LMP defined in (6.41) for scenarios k. Consider the following settlement rule:

1 *Energy prices (scenario-dependent LMP)* λ_k^*/w_k: If the scenario k materializes at delivery time then unit j that provides energy $(p_j + r_{kj})$ is paid by the system operator the amount $\lambda_{kj}^* (p_j + r_{kj})/w_k$.

2 *Reserve payment* $\sum_k \mu_k^* h_{kj} \left(r_j^{\min}, r_j^{\max}\right)$: Regardless of scenario at delivery time, unit j that provides reserve capacities $(r_j^{\min}, r_j^{\max})$ is paid by the system operator the amount $\sum_k \mu_k^* h_{kj} \left(r_j^{\min}, r_j^{\max}\right)$.

The settlement rule enjoys three desirable properties:

- *Incentive compatible*. When unit j is faced with the scenario-dependent LMP λ_{kj}^*/w_k in scenario k and the reserve payment $\sum_k \mu_k^* h_{kj} \left(r_j^{\min}, r_j^{\max}\right)$, it would have preferred to choose $x_j^* := (p_j^*, r_j^{\min *}, r_j^{\max *}, r_{kj}^*, k \geq 1)$ that maximizes its *expected* profit, i.e., it solves:

$$\max_{x_j} \sum_k w_k \left(\lambda_{kj}^* (p_j + r_{kj})/w_k - f_{kj}(p_j + r_{kj})\right) + \sum_k \mu_k^* h_{kj} (r_j^{\min}, r_j^{\max}) \quad (6.44a)$$

$$\text{s.t. } (6.39b)(6.39g) \quad (6.44b)$$

  The settlement rule is called *incentive compatible* in expectation if a primal optimal solution x^* of (6.40) also solves the expected profit maximization (6.44) for all units under the settlement rule. Note that the individual optimization (6.44) relaxes all coupling constraints but includes all local constraints.

- *Revenue adequate*. If all units provide their energy and reserves according to a primal optimal solution x^*, then the total payment to the system operator in each scenario k, called the *merchandizing surplus*, under the settlement rule is:

$$\text{MS}_k := -\sum_j \frac{1}{w_k} \lambda_{kj}^* (p_j^* + r_{kj}^* + g_k - d_k) - \sum_i \sum_j \mu_i^* h_{ij} \left(r_j^{\min}, r_j^{\max}\right) \quad (6.45)$$

  The settlement rule is called *revenue adequate* in each scenario $k \geq 1$ if $\text{MS}_k \geq 0$.

- *Reserve payment balanced*. The reserve payments are said to be *balanced* under the settlement rule if

$$\sum_j \sum_k \mu_k^* h_{kj}\left(r_j^{\min}, r_j^{\max}\right) = 0 \tag{6.46}$$

This means that those units that need more reliability exactly compensate those that can provide more reliability.

Theorem 6.4 (ICRA)**.** Suppose f_{kj}, h_{kj} are real-valued, convex, and continuously differentiable. Let (x^*, ξ^*) be a primal-dual optimal solution of (6.40). Then the settlement rule is:

- incentive compatible in expectation: x^* solves (6.44).
- revenue adequate in each scenario $k \geq 1$: $\text{MS}_k \geq 0$ in (6.45).
- reserve payment balanced: $\sum_j \sum_k \mu_k^* h_{kj}\left(r_j^{\min}, r_j^{\max}\right) = 0$ in (6.46).

The theorem is proved in Exercise 6.12. Indeed the settlement rule is also incentive compatible in each scenarios $k \geq 1$ in the sense that, after the first-stage commitment $(p^*, r^{\min}, r^{\max})$ when the scenario k is realized, r_{kj}^* from an optimal solution of (6.40) will also maximize unit j's profit in scenario k. The formulation here can be extended to allow the network (shift factor S) and nodal injection sets to depend on contingencies k to model outages, or to include additional local constraints, ramp rates, network losses, reserve costs, and per-area reliability requirements, or to allow reserve constraints $h_{kj}(p_j, r_j^{\min}, r_j^{\max})$ to depend on both the dispatch decision and reserve decisions.

6.4.5 Security constrained unit commitment

The unit commitment problem (6.4) can be extended to include corrective reserves by replacing the real-time dispatch problem (6.4c)(6.4d)(6.4e) by a security constrained OPF similar to (6.40).

For example let the first-stage decisions be the binary commitments $u_j(t) \in \{0,1\}$ for units j in periods t. For each t let w_{tk} denote the probability of scenario k such that $w_{tk} \geq 0$ and $\sum_k w_{tk} = 1$. Let the second-stage decisions be dispatch and reserve amounts $x(t) := (p(t), r^{\min}(t), r^{\max}(t), r_k(t), k \geq 1)$ for all units in periods t. Security constrained unit commitment can be formulated as the following problem (cf. (6.4)):

$$\min_{u \in \{0,1\}^{(N+1)T}} \quad \sum_t \sum_j c_{jt}\left(u_j(t-1), u_j(t)\right) + f^*(u) \tag{6.47a}$$

$$\text{s.t.} \quad u_j(t) - u_j(t-1) \leq u_j^\tau, \qquad \forall \tau \in \{t+1, t+\text{up}_j - 1\} \tag{6.47b}$$

$$u_j(t-1) - u_j(t) \leq 1 - u_j^\tau, \qquad \forall \tau \in \{t+1, t+\text{down}_j - 1\} \tag{6.47c}$$

where c_{jt} is the commitment cost such as the startup/shut down cost defined in (6.3b)

and reproduced here

$$c_{jt}(u_j(t-1),u_j(t)) := \begin{cases} \text{startup cost} & \text{if} \quad u_j(t)-u_j(t-1)=1 \\ \text{shutdown cost} & \text{if} \quad u_j(t)-u_j(t-1)=-1 \\ 0 & \text{if} \quad u_j(t)-u_j(t-1)=0 \end{cases} \tag{6.47d}$$

and (6.47b)(6.47c) imposes minimum up/down time once unit j is turned on/off.

Given a commitment decision u, $f^*(u)$ in (6.47a) is the optimal expected security constrained real-time dispatch cost over the entire optimization horizon (cf. (6.40)):

$$f^*(u) := \min_{x(t)} \sum_t \sum_k w_{tk} f_{tk}(p(t)+r_k(t)) := \sum_t \sum_k w_{tk} \sum_j f_{tkj}(p_j(t)+r_{kj}(t))$$

$$\text{s.t.} \quad p^{\min} \odot u(t) \le p(t)+r^{\min}(t),\ p(t)+r^{\max}(t) \le p^{\max} \odot u(t) \tag{6.47e}$$

$$h_{tk}\left(r^{\min}(t),r^{\max}(t)\right) := \sum_j h_{tkj}\left(r_j^{\min}(t),r_j^{\max}(t)\right) \ge 0 \tag{6.47f}$$

$$\mathbf{1}^{\mathsf{T}}(p(t)+r_k(t)+g_k(t)-d_k(t)) = 0 \tag{6.47g}$$

$$P^{\min} \le S^{\mathsf{T}}(t)(p(t)+r_k(t)+g_k(t)-d_k(t)) \le P^{\max} \tag{6.47h}$$

$$r^{\min}(t) \le r_k(t) \le r^{\max}(t) \tag{6.47i}$$

$$|p(t)-p(t-1)| \le p^{\text{ramp}} \tag{6.47j}$$

This problem for each t would have been the same as the security constrained economic dispatch (6.40) if it were not for two features. First, for two vectors a and b, $a \odot b$ in (6.47e) denotes componentwise product, i.e., $(a \odot b)_j := a_j b_j$. If $u_j(t) = 1$ (unit j on) then (6.47e) is the same as (6.39b). If $u_j(t) = 0$ (unit j off) then (6.47e) forces $p_j(t) = r_j^{\min}(t) = r_j^{\max}(t) = r_{kj}(t) = 0$. Second the new constraint (6.47j) imposes a limit p^{ramp} on ramping of the dispatch p. Without this ramping constraint, the problem $f^*(u)$ is decoupled across t and the time-t subproblems can be solved independently of each other.

The security constrained unit commitment problem (6.47) is a mixed integer linear program. It can be solved to optimality using branch and bound methods (Chapter 8.5.6) or Benders decomposition (see Example 8.17 in Chapter 8.5.7).

6.5 Bibliography

There are many excellent texts on various aspects of power system operations in much more detail than this book, e.g., [1, 3, 2]. Automatic generation control that encompasses voltage control and load frequency control is discussed in detail in e.g. [1, Chapter 11], [52].

6.6 Problems

Chapter 6.2

Exercise 6.1 (Imbalance and error model)**.** The optimal dispatch (6.6) is a deterministic problem driven by a forecast $\hat{m}(n)$ of the random injection $\sigma(\xi,t)$ that is solved in the nth control interval.

1 Discuss three types of error: random error $\Delta_1(\xi,t) := u(\sigma(\xi,t)) - u^{\text{opt}}(m(t))$, discretization error $\Delta_2(t) := u^{\text{opt}}(m(t)) - u^{\text{opt}}(\bar{m}(n))$ where $\bar{m}(n)$ is the time average of $m(t)$ over $[n\delta,(n+1)\delta)$, and prediction error $\Delta_3(\xi,t) := u^{\text{opt}}(\bar{m}(n)) - u^{\text{opt}}(\hat{m}(n))$ at each time $t \in [n\delta,(n+1)\delta)$, where $\hat{m}(n)$ is an estimate of $\bar{m}(n)$.

2 Consider a 2-bus network described by the DC power flow model. Bus 1 has an uncontrollable load $\sigma := (\sigma(t), t \in \mathbb{R}_+)$ with mean $(m(t), t \in \mathbb{R}_+)$ and bus 2 has a controllable generator with output level $u(t)$. Suppose the generator and line capacities are high so that the injection and line limits are never active.

(a) Suppose we use the prediction

$$\hat{m}(n) \;:=\; \hat{m}(\xi,n) := \frac{1}{\delta}\int_{(n-1)\,\delta}^{n\delta} \sigma(\xi,t)\,dt, \qquad n = 0,1,\ldots \tag{6.48}$$

Show that the imbalance at time t is the difference between the actual load at time t and the time average load over the pervious interval.

(b) Suppose σ is a white Gaussian process with mean $E\sigma(t) = m(t)$ and correlation function $K(t,t') = \nu^2$ if $t = t'$ and $K(t,t') = 0$ if $t \neq t'$ for $t,t' \geq 0$. Then, under appropriate integrability assumptions, $w(\tau) := \int_0^\tau \sigma(t)dt$ is a Wiener process with the property that non-overlapping increments are independent Gaussian random variables, i.e., for any $t' < t \leq \tau' < \tau$, the random variables

$$w(t) - w(t') \;:=\; \int_{t'}^{t} \sigma(s)ds \quad \text{and} \quad w(\tau) - w(\tau') \;:=\; \int_{\tau'}^{\tau} \sigma(s)ds$$

are independent and Gaussian with means $\int_{t'}^{t} m(s)ds$ and $\int_{\tau'}^{\tau} m(s)ds$ respectively and variance $\nu^2(t-t')$ and $\nu^2(\tau-\tau')$ respectively. Derive the various errors and properties in Table 6.1.

	Expression	Random Var	Mean	Variance
Random error $\Delta_1(\xi,t)$	$-\sigma(\xi,t)+m(t)$	Gaussian	zero	ν^2
Discretiz. error $\Delta_2(t)$	$-m(t)+\bar{m}(n)$	constant	$-m(t)+\bar{m}(n)$	0
Prediction error $\Delta_3(\xi,t)$	$-\bar{m}(n)+\hat{m}(\xi,n)$	Gaussian	$-\bar{m}(n)+\bar{m}(n-1)$	ν^2/δ
Imbalance $\Delta u(\xi,t)$	$\Delta_1(t)+\Delta_2(t)+\Delta_3(t)$	Gaussian	$-m(t)+\bar{m}(n-1)$	$\nu^2\,(1+1/\delta)$

Table 6.1 Exercise 6.1: Imbalance and underlying errors.

(c) Verify the following properties: (i) The mean random error $E\Delta_1(t) = 0$. (ii) The time average of the discretization error $\Delta_2(t)$ is zero over each control

interval. (iii) The mean prediction error $E\Delta_3(t)$ is small if the mean process $m(t)$ is slowly time-varying. In particular if σ is stationary then $E\Delta_3(t) = 0$.

Chapter 6.3

Exercise 6.2 (Primary frequency control)**.** Proof Theorem 6.1.

Exercise 6.3 (Secondary frequency control)**.** Proof Theorem 6.2.

Exercise 6.4 (Optimality of primary frequency control)**.** Formulate underlying optimization problem solved by primary frequency control (c.f. Changhong2014TAC).

Exercise 6.5 (Optimality of secondary frequency control)**.** Formulate underlying optimization problem solved by secondary frequency control (c.f. LinaCZ paper).

Chapter 6.4

Exercise 6.6 (3-bus network)**.** Recall the conversion matrix Γ^{T} defined in (1.12) and reproduced here:

$$\Gamma^{\mathsf{T}} := \begin{bmatrix} 1 & 0 & -1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}$$

Treat Γ^{T} as the incidence matrix of the 3-node network in Figure 1.9(b). Assume line susceptances $b_l = 1$ for all l.

1 Show that the Laplacian matrix $L := \Gamma^{\mathsf{T}}\Gamma$ and its pseudo-inverse $L^{\dagger}$ are

$$L = \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}, \qquad L^{\dagger} = \frac{1}{9}L$$

2 Show that the shift factor $S := L^{\dagger}\Gamma^{\mathsf{T}}B = \frac{1}{3}\Gamma^{\mathsf{T}}$.
3 Show that line flows $P = \frac{1}{3}\Gamma p$ and LMP $\lambda = \gamma\mathbf{1} + \frac{1}{3}\Gamma^{\mathsf{T}}\kappa$.

Exercise 6.7 (Merchandizing surplus)**.** Prove (6.31).

Exercise 6.8 (Theorem 6.3: proof)**.** This exercise proves Theorem 6.3 step by step. Consider the DC power flow model (6.34) and the economic dispatch formulation (6.35) in terms of $S_r := L_r^{-\mathsf{T}}C_{-0}B$. Assume L_r^{-1} exists so S_r is well defined.

1 Show that $P = S_r^\mathsf{T} p_{-r} = S^\mathsf{T} p$ where $S := L^\dagger CB$ is defined in (6.20), i.e., the line flows P given in (6.34) are independent of the choice of the price reference bus r.
2 Show that S_r is related to S as:

$$\begin{bmatrix} S_r \\ 0 \end{bmatrix} = S - \mathbf{1}_{N+1} s_r^\mathsf{T}, \qquad S_r = [S]_{-r} - \mathbf{1}_N s_r^\mathsf{T}$$

where s_r^T denotes row r of S, $[S]_r$ denotes the submatrix of S obtained by removing row r, and $\mathbf{1}_n$ is the vector of all 1s of size n, i.e., S_r is obtained from the submatrix $[S]_{-r}$ of S by subtracting row r from every row in $[S]_{-r}$.
3 Prove Theorem 6.3 using parts 1 and 2.

Exercise 6.9 (Energy price)**.** Use (6.36) to show that $\sum_j \lambda_j^* = (N+1)\tilde{\gamma}^*$ where $\tilde{\gamma}^*$ is the energy price defined in (6.37). (This is what should be expected given that $\tilde{\lambda}^* = \lambda^*$ according to Theorem 6.3.)

Exercise 6.10 (Theorem 6.3: illustration)**.** Consider the two-bus network and the economic dispatch (6.26) of Example 6.2. An equivalent formulation is to replace the line flow $BC^\mathsf{T} L^\dagger (p-d)$ in the line limit (6.26c) by

$$-P^{\max} \le p_1 - d_1 \le P^{\max}$$

This is equivalent to using (6.26b) to eliminate p_2 from $BC^\mathsf{T} L^\dagger (p-d)$. This means that bus 2 is chosen as the price reference bus r in the economic dispatch formulation (6.35).

1 For the formulation (6.35), calculate L_2, L_2^{-1}, S_2 and derive expressions for LMP λ^*.
2 Compare with the corresponding quantities in Example 6.2 and verify that the LMPs are the same in both formulations, as asserted by Theorem 6.3.

Chapter 6.4.4

Exercise 6.11. Consider the two-stage economic dispatch problem (6.40) and the LMP λ_k^* defined in (6.41) for scenarios k. Show that a primal-dual feasible (x^*, ξ^*) is primal-dual optimal for (6.40) if and only if (x^*, ξ^*) satisfies (6.42)(6.43).

Exercise 6.12. Prove Theorem 6.4. (Hint: Show that $w_k \, \mathrm{MS}_k = \left(\kappa_k^{+*}\right)^\mathsf{T} P^{\max} - \left(\kappa_k^{-*}\right)^\mathsf{T} P^{\min}$. For incentive compatibility, note that (x^*, ξ^*) satisfies the complementary slackness conditions (6.43c)(6.43e) for the coupling constraints.)

7 System operation: estimation and control

In this chapter we illustrate the network models of Chapters 4 and 5 in several applications. The emphasis is on the use of structural properties of these models to attain conceptual understanding of applications or design solutions with performance guarantees, not on the scalable computation of these models. To make this chapter self-contained we summarize the models used in each application.

7.1 State estimation

Consider a power network modeled by a connected undirected graph $G = (\overline{N}, E)$ of $N+1$ nodes and M lines, where $\overline{N} := \{0\} \cup N$, $N := \{1, 2, \ldots, N\}$ and $E \subseteq \overline{N} \times \overline{N}$. The state of the network is the complex voltage $V_j \in \mathbb{C}$ at each bus $j \in \overline{N}$. We assume the voltage angle $\theta_0 := 0$ at the reference bus 0, and hence the state is a $2N+1$ dimensional real vector $x := (\theta, |V|) := (\theta_j, |V_0|, |V_j|, j \in N) \in \mathbb{R}^{2N+1}$. The problem of state estimation is to estimate the state $x := (\theta, |V|)$ from a set of noisy measurements $y \in \mathbb{R}^K$. It is a key building block for numerous power system applications, e.g., in energy management systems that dispatch controllable generators and loads in transmission systems or control voltages on distribution systems.

The measurements y may consist of voltage angles and magnitudes $(\theta_j, |V|_j)$ at a subset $N_1 \subset \overline{N}$ of the buses $j \in N_1$. These are partial and noisy state measurements. We assume the measurement noise is additive, i.e.,

$$y_{2j} = \theta_j + z_{2j}, \qquad y_{2j+1} = |V_j| + z_{2j+1}, \qquad j \in N_1 \tag{7.1}$$

where (z_{2j}, z_{2j+1}) are additive measurement noises. The measurements y may also include real and reactive power injections (p_j, q_j) at a subset $N_2 \subset \overline{N}$ of the buses $j \in N_2$, i.e.,

$$y_{2j} = p_j + z_{2j}, \qquad y_{2j+1} = q_j + z_{2j+1}, \qquad j \in N_2$$

where (z_{2j}, z_{2j+1}) are additive measurement noises. The injections satisfy power flow equations, e.g., in polar form, that relate (p_j, q_j) to the state x, i.e,

$$p_j = f_j(x), \qquad q_j = g_j(x)$$

where (from (4.27))

$$f_j(x) := \sum_{k:k\sim j}\left(g^s_{jk}+g^m_{jk}\right)|V_j|^2 - \sum_{k:k\sim j}|V_j||V_k|\left(g^s_{jk}\cos\theta_{jk}+b^s_{jk}\sin\theta_{jk}\right) \quad (7.2a)$$

$$g_j(x) := -\sum_{k:k\sim j}\left(b^s_{jk}+b^m_{jk}\right)|V_j|^2 - \sum_{k:k\sim j}|V_j||V_k|\left(g^s_{jk}\sin\theta_{jk}-b^s_{jk}\cos\theta_{jk}\right) \quad (7.2b)$$

Substituting into (y_{2j}, y_{2j+1}) we have

$$y_{2j} = f_j(x)+z_{2j}, \qquad y_{2j+1} = g_j(x)+z_{2j+1}, \qquad j\in N_2 \quad (7.2c)$$

The measurements y may also include real and reactive powers (P_{jk}, Q_{jk}) on a subset $E_1 \subseteq E$ of the lines $(j,k)\in E_1$. Then

$$y_{2l} = P_l(x)+z_{2l}, \qquad y_{2l+1} = Q_l(x)+z_{2l+1}, \qquad l\in E_1 \quad (7.3a)$$

where

$$P_{jk}(x) := \left(g^s_{jk}+g^m_{jk}\right)|V_j|^2 - |V_j||V_k|\left(g^s_{jk}\cos\theta_{jk}+b^s_{jk}\sin\theta_{jk}\right) \quad (7.3b)$$

$$Q_{jk}(x) := -\left(b^s_{jk}+b^m_{jk}\right)|V_j|^2 - |V_j||V_k|\left(g^s_{jk}\sin\theta_{jk}-b^s_{jk}\cos\theta_{jk}\right) \quad (7.3c)$$

In general we have a measurement model

$$y = f(x)+z$$

where $x := (\theta, |V|)\in\mathbb{R}^{2N+1}$ is the state of the network, $y\in\mathbb{R}^K$ is the measurement vector, $z\in\mathbb{R}^K$ is additive noise, and $f:\mathbb{R}^{2N+1}\to\mathbb{R}^K$ is a network model of the form in (7.1)(7.2)(7.3) that maps a network state to measurement.

Unconstrained formulation. The simplest formulation of state estimation is the following problem to determine an estimate $\hat{x}$ of x from y:

$$\hat{x} := \arg\min_{x\in\mathbb{R}^{2N+1}} (y-f(x))^{\mathsf{T}} R^{-1}(y-f(x)) \quad (7.4)$$

where R is a (symmetric) positive definite normalization matrix. A common normalization matrix is the covariance matrix $E(z-Ez)(z-Ez)^{\mathsf{T}}$ of the noise z, estimated from the measurement as $R := \hat{E}(yy^{\mathsf{T}}) - \hat{E}y\hat{E}y^{\mathsf{T}}$ where $\hat{E}z := (1/k)\sum_{i=1}^k z_i$ denotes the sample mean of k samples $z_1,\ldots,z_k$ of z, or its unbiased version (see discussion in Exercise 7.2). This optimization problem is called a *least square estimation* or *nonlinear regression problem*. It is a convex problem if $f(x) = Ax$ is a linear function. We will study a basic theory of and algorithms for solving convex optimization problems in Chapter 8.

A linearized version of the state estimation problem (7.4) can be obtained by linearizing $f(x)$ around an operating point x_0, i.e., we assume the measurement model $y =: y_0+\Delta y = f(x_0)+\frac{\partial f}{\partial x}(x_0)\Delta x + z$. We assume $f(x_0)$ is known and $y_0 = f(x_0)$. Hence we have

$$\Delta y = F\Delta x + z$$

where $F := \frac{\partial f}{\partial x}(x_0)$ is the $K \times (2N+1)$ Jacobian matrix of f at the operating point x_0. Then the least square estimation is the following linear regression:

$$\hat{\Delta x} \;:=\; \arg \min_{\Delta x \in \mathbb{R}^{2N+1}} (\Delta y - F\Delta x)^{\mathsf{T}} R^{-1} (\Delta y - F\Delta x) \tag{7.5a}$$

where $\Delta y := y - f(x_0)$ is obtained from the measurement y and the operating point x_0. Let the estimation error be the minimum weighted norm of the noise z:

$$\epsilon^2 \;:=\; \min_{\Delta x} z^{\mathsf{T}} R^{-1} z \;=\; \min_{\Delta x} (\Delta y - F\Delta x)^{\mathsf{T}} R^{-1} (\Delta y - F\Delta x) \tag{7.5b}$$

To simplify notation consider quantities normalized by the positive definite matrix R

$$\Delta \bar{y} \;:=\; R^{-1/2} \Delta y, \qquad \bar{F} \;:=\; R^{-1/2} F \tag{7.6a}$$

For example if $R := \mathrm{diag}(\sigma_i^2)$ is the (sample) variance of the noise z then (7.6a) normalizes the measurements Δy by its standard deviation. Then the linear regression (7.5) becomes

$$\min_{\Delta x} z^{\mathsf{T}} R^{-1} z \;=\; \min_{\Delta x} \left\| \Delta \bar{y} - \bar{F} \Delta x \right\|_2^2 \tag{7.6b}$$

An optimum $\hat{\Delta x}$ and the resulting minimum estimation error ϵ^2 can be solved in closed form. The general solution is $\hat{\Delta x} = \bar{F}^{\dagger} \Delta \bar{y}$ where $\bar{F}^{\dagger}$ is the pseudo-inverse of $\bar{F} := R^{-1/2} F$ (see Chapter A.7, particularly Remark A.2).

There are two special cases where $\bar{F}^{\dagger}$ has simple expressions in terms of $\bar{F}$ according as $\bar{F}$ has full column or row rank:

1 More measurements than state variables $K \geq 2N+1$: Redundant measurements allow us to estimate the network state by solving the linear regression (7.6). When the columns of F (and hence $\bar{F}$) are linearly independent, the unique optimal solution is (Exercise 7.1):

$$\hat{\Delta x} \;=\; \left(\bar{F}^{\mathsf{T}} \bar{F} \right)^{-1} \bar{F}^{\mathsf{T}} \Delta \bar{y} \tag{7.7a}$$

and the minimum error is

$$\epsilon^2 \;=\; \left\| \Delta \bar{y} - \bar{F} \hat{\Delta x} \right\|_2^2 \;=\; \|\Delta \bar{y}\|_2^2 - \left\| \left(\bar{F}^{\mathsf{T}} \bar{F} \right)^{-1/2} \bar{F}^{\mathsf{T}} \Delta \bar{y} \right\|_2^2 \tag{7.7b}$$

The estimated state is

$$x_0 + \hat{\Delta x} \;=\; x_0 \;+\; \left(F^{\mathsf{T}} R^{-1} F \right)^{-1} F^{\mathsf{T}} R^{-1} \Delta y \tag{7.7c}$$

where $\Delta y := y - f(x_0)$ and $F := \frac{\partial f}{\partial x}(x_0)$.

2 Fewer measurements than state variables $K < 2N+1$: When F and hence $\bar{F}$ has full row rank, then $\bar{F}$ has K linearly independent columns and the estimation error $\epsilon^2 = 0$. The unique optimal solution to (7.6) is the solution to $\bar{F} \Delta x = \Delta \bar{y}$ (see Corollary A.20 in Chapter A.7):

$$\hat{\Delta x} \;=\; \bar{F}^{\mathsf{T}} \left(\bar{F} \bar{F}^{\mathsf{T}} \right)^{-1} \Delta \bar{y}$$

Since $K < 2N+1$, there is a subspace of solutions to $\bar{F}\Delta x = \Delta\bar{y}$ and $\hat{\Delta}x$ is one with the minimum Euclidean norm. There is no reason the state of the power network is closed to such a solution and generally the lack of sufficient measurement produces poor state estimates (even though $\epsilon^2 = 0$).

Constrained formulation. State estimation can also include operational constraints such as injection limits, voltage limits and line limits. The injection limits take the form $(p_j^{\min}, q_j^{\min}) \le (f_j(x), g_j(x)) \le (p_j^{\max}, q_j^{\max})$ from (7.2), the voltage limits take the form $V_j^{\min} \le a_j^{\mathsf{T}} x \le V_j^{\max}$, and the line limits take the form $(P_{jk}^{\min}, Q_{jk}^{\min}) \le (P_{jk}(x), Q_{jk}(x)) \le (P_{jk}^{\max}, Q_{jk}^{\max})$ from (7.3). The constrained version of state estimation (7.4) is then:

$$\hat{x} \;:=\; \arg\min_{x\in\mathbb{R}^{2N+1}} \; (y - f(x))^{\mathsf{T}} R^{-1} (y - f(x)) \tag{7.8a}$$

$$\text{s.t.} \quad h(x) \;\le\; 0 \tag{7.8b}$$

where $h(x)$ represents operational constraints. There is generally no analytical solution for (7.8). We will study iterative algorithms in Chapter 8.5 for solving constrained optimization problems.

7.2 Volt/var control on radial networks

In this section we apply the linear DistFlow model (5.34) or (5.35) of Chapter 5.4 and Theorem 4.10 of Chapter 4.2.6 for voltage control on radial networks. The expression (4.24) for $\hat{Z} = \hat{Y}^{-1}$ in Theorem 4.10 is useful for various power system applications on radial networks. As explained in Remark 4.6 this structure originates from the inverse $\hat{C}^{-1}$ in (4.23) of the reduced incidence matrix $\hat{C}$ of a tree graph and is independent of the "weight matrix" D_y^s as long as D_y^s is nonsingular. In many applications, D_y^s is not only nonsingular but also positive or negative definite. In this section we apply this result to the linear DistFlow model of Chapter 5.4.2 for voltage control (or Theorem 5.3 that specializes Theorem 4.10 to linear DistFlow model). In Chapter 7.3 we apply Theorem 4.10 to a linearized polar-form power flow model for topology identification.

7.2.1 Linear DistFlow model

Consider a radial network $G := (\overline{N}, E)$ with $N+1$ buses and M lines, modeled by the linear DistFlow equations (5.34) with a given v_0, or equivalently, by (5.35) of Chapter 5.4.3.1 reproduced here:

$$\tilde{s} \;=\; \hat{C}\, S, \qquad\qquad v_0\, c_0 + \hat{C}^{\mathsf{T}} v \;=\; 2\,(D_r P \,+\, D_x Q) \tag{7.9}$$

where $(\tilde{s}, v)$ here denote the real and reactive net injections and squared voltage magnitudes at *non-reference* buses, $S := (P, Q)$ are real and reactive line flows, $C^{\mathsf{T}} := [c_0 \;\; \hat{C}^{\mathsf{T}}]$ is the transpose of the node-by-line incidence matrix C, in particular $\hat{C}$

is the $N \times N$ reduced incidence matrix corresponding to non-reference buses, and $D_r := \mathrm{diag}(r_l, l \in E)$, $D_x := \mathrm{diag}(x_l, l \in E)$ are diagonal matrices of line resistances and reactances respectively. As in Chapter 5.4.2, we assume throughout this section without stating it explicitly that the network graph G is a (connected) tree, $y^s_{jk} = y^s_{kj}$ (assumption C5.1) and $y^m_{jk} = y^m_{kj} = 0$. To simplify notation, we will use $(p,q) \in \mathbb{R}^{2N}$ and $v \in \mathbb{R}^N$ in this section to denote variables at non-reference buses (instead of $\hat{s}, \hat{v}$ as in Chapter 5.4).

We assume at each bus j there is a fixed and given active and reactive load $s^0_j := \left(p^0_j, q^0_j\right)$. In addition there is possibly an inverter on bus j with a fixed active power injection p_j and an adaptable reactive power injection q_j. For example, p_j may represent solar generation. Hence the net injections $\tilde{s}$ in (7.9) are $\tilde{s} = (p - p^0, q - q^0)$. The problem of volt/var control is to adapt the reactive outputs q_j in order to stabilize voltages on the network. To this end, since the network is radial, the reduced incidence matrix $\hat{C}$ is nonsingular and we can apply Theorem 5.3 of Chapter 5.4.3.1 to solve (7.9) and express v in terms of the net injections:

$$v \;=\; v_0 \mathbf{1} + 2\left(R(p - p^0) \;+\; X(q - q^0)\right)$$

where $R := \hat{C}^{-\mathsf{T}} D_r \hat{C}^{-1}$ and $X := \hat{C}^{-\mathsf{T}} D_x \hat{C}^{-1}$ are positive definite. We write $v := v(q)$ explicitly as a function of the control q:

$$v(q) = 2Xq \;+\; \tilde{v} \tag{7.10}$$

where $\tilde{v} := v_0 \mathbf{1} + 2R\left(p - p^0\right) - 2Xq^0$ does not depend on q.

A common model of inverters constrains the reactive power q_j to the sector $\{q_j : p_j^2 + q_j^2 \le \sigma^2\}$ with a power factor limit $-\phi_j \le \tan^{-1}(q_j/p_j) \le \phi_j \le \pi/2$. Equivalently the control q_j is constrained to the sector U_j determined by the given active power $\tilde{p}_j$:

$$U_j := \left\{\; q_j \;:\; \underline{q}_j \le q_j \le \overline{q}_j \;\right\}, \qquad j = 1, \ldots, N \tag{7.11}$$

where $\overline{q}_j := \min\left\{p_j \tan\phi_j,\ \sqrt{\sigma^2 - p_j^2}\right\}$ and $\underline{q}_j := \max\left\{-p_j \tan\phi_j,\ -\sqrt{\sigma^2 - p_j^2}\right\}$. Let $U := U_1 \times \cdots \times U_N$. If the reactive power q_j of the inverter at bus j is fixed and not controllable, this can be modeled by setting $\underline{q}_j = q_j = \overline{q}_j$. If there is no inverter at bus j, then we set $p_j = \underline{q}_j = \overline{q}_j := 0$.

7.2.2 Decentralized control: convergence and optimality

Let v^{ref} be a given vector of reference voltages at buses $j > 0$. Our goal is to choose control $q \in U$ to drive voltages towards v^{ref}. We require our control to be local, i.e., $q_j(t+1)$ depends only on voltage $v_j(t)$ at bus j, not voltages $v_k(t)$ at other buses $k \ne j$, and memoryless, i.e., $q_j(t+1)$ depends only on $v_j(t)$ but not $v_j(s), s < t$. In

particular, q_j is a function only of voltage discrepancy $v_j(t) - v_j^{\text{ref}}$, of the form

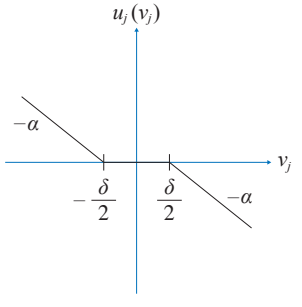
$$q_j(t+1) = \left[u_j\left(v_j(t) - v_j^{\text{ref}}\right)\right]_{U_j}, \qquad j = 1, \ldots, N$$

where $v_j(t)$ is the measured local voltage, $u_j : \mathbb{R} \to \mathbb{R}$ is a control function that maps a voltage deviation $v_j(t) - v_j^{\text{ref}}$ into a potential reactive power setting, $[a]_{U_j} := \max\left\{\underline{q}_j, \min\left\{a, \overline{q}_j\right\}\right\}$ is the projection onto U_j. Such a local memoryless control is simple to implement as it requires no communications among controllers at different buses.

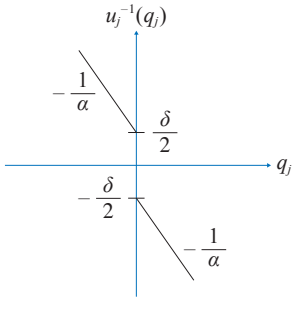
The local volt/var control problem in our formulation boils down to the design of the control function u_j. Many functions u_j have been proposed and analyzed in the literature. We now present such a control from [37, 38]. From Theorem 5.3,

$$\frac{\partial v_j}{\partial q_j} = 2X_{jj} \;=\; 2\sum_{l \in \mathsf{P}_j} x_l \;>\; 0$$

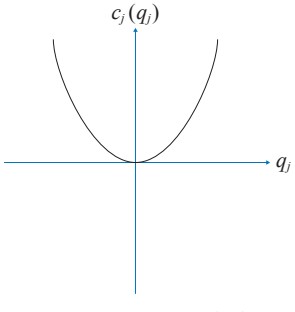
Therefore it is natural to choose a control function u_j that is nonincreasing in voltage discrepancy $v_j(t) - v_j^{\text{ref}}$. An example u_j is shown in Figure 7.1(a).



(a) Piecewise linear control $u_j(v_j)$



(b) Inverse $u_j^{-1}(q_j)$



(c) Implied cost $c_j(q_j)$

Figure 7.1 Piecewise linear control with a deadband $(-\delta/2, \delta/2)$.

Closed-loop behavior.

Consider the closed-loop system under a local control u_j. Suppose the voltages evolve according to (7.10), i.e., suppose the measured voltages at time t is $v_j(t) = v_j(q(t))$. Then the closed-loop system is a discrete-time dynamical system defined by the control function $u_j : \mathbb{R} \to \mathbb{R}$ followed by a projection onto U_j:

$$q_j(t+1) = \left[u_j\left(v_j(q(t)) - v_j^{\text{ref}}\right)\right]_{U_j}, \qquad j = 1, \ldots, N \tag{7.12}$$

where $v_j(q)$ is given by (7.10). If $q^* = \left[u\left(v(q^*) - v^{\text{ref}}\right)\right]_U$ then q^* is called a fixed point, or an equilibrium point, of (7.12).

We now analyze the convergence and optimality of the dynamical system (7.12) for a class of u_j that satisfies the following assumptions:

C5.1: The control functions u_j are differentiable on $\mathbb{R}$ and there exist α_j such that $\left|u_j'(v_j)\right| \le \alpha_j$ for all $v_j \in \mathbb{R}$.
C5.2: The control functions u_i are strictly decreasing on $\mathbb{R}$.

The differentiability assumption in C5.1 can be relaxed to allow control functions with a deadband and saturation as shown in Figure 7.1(a) (see [38]). The proof of the convergence and optimality properties in the next two theorems uses concepts in convex optimization theory that we will study in detail in Chapter 8. Let $A := \text{diag}(\alpha_j, j \in N)$.

Theorem 7.1 (Convergence)**.** Suppose assumption C5.1 holds. If the largest singular value $\sigma_{\max}(AX) < 1/2$ then there exists a unique equilibrium point $q^* \in U$ and the volt/var control (7.12) converges to q^* geometrically, i.e.,

$$\|q(t) - q^*\| \le \beta^t \|q(0) - q^*\| \quad \to \quad 0$$

for some $\beta \in [0, 1)$.

Proof Applying the mean value theorem to the control function $u_j(v_j)$ we have

$$u_j(v_j) - u_j(\hat{v}_j) = u_j'(w)(v_j - \hat{v}_j)$$

where $w = \lambda v_j + (1-\lambda)\hat{v}_j$ for some $\lambda \in [0, 1]$. Therefore

$$\|u(v) - u(\hat{v})\|_2^2 = \sum_j \left|u_j(v_j) - u_j(\hat{v}_j)\right|^2 \quad \le \quad \sum_j \left|\alpha_j(v_j - \hat{v}_j)\right|^2 \quad = \quad \|A(v - \hat{v})\|_2^2$$

where the inequality follows from the mean value theorem and assumption C5.1. Hence $\|u(v) - u(\hat{v})\|_2 \le \|A(v - \hat{v})\|_2$. Applying the chain rule to $Av = Av(q)$ as a vector-valued function of q we have

$$\frac{\partial Av}{\partial q}(q) = A\frac{\partial v}{\partial q} \quad = \quad 2AX$$

Therefore

$$\left\| u\left(v(q) - v^{\text{ref}}\right) - u\left(v(\hat{q}) - v^{\text{ref}}\right) \right\|_2 \leq \|Av(q) - Av(\hat{q})\|_2 \;\leq\; \|2AX\|_2 \, \|q - \hat{q}\|_2$$

where the first inequality follows from $\|u(v) - u(\hat{v})\|_2 \leq \|A(v - \hat{v})\|_2$. The second inequality follows from the mean value Theorem A.34 for vectored-valued functions in Appendix A.10 that says that if $f : \mathbb{R}^n \to \mathbb{R}^n$ is continuously differentiable then

$$\|f(y) - f(x)\| \;\leq\; \left\| \frac{\partial f}{\partial x}(z) \right\| \|y - x\|$$

for any induced matrix norm $\|\cdot\|$ where $z := \mu x + (1-\mu)y$ for some $\mu \in [0,1]$. Since the induced matrix norm $\|M\|_2 = \sigma_{\max}(M)$ (Exercise 7.3) we have

$$\left\| u\left(v(q) - v^{\text{ref}}\right) - u\left(v(\hat{q}) - v^{\text{ref}}\right) \right\|_2 \leq 2\sigma_{\max}(AX) \, \|q - \hat{q}\|_2$$

Therefore the control function $u\left(v(q) - v^{\text{ref}}\right)$ as a function of q is a contraction when $\sigma_{\max}(AX) < 1/2$. Since projection onto U is nonexpansive by the Projection Theorem 8.9 of Chapter 8.2.3, the function on the right-hand side of (7.12), as a function of q, is a contraction. The theorem then follows from the Contraction Mapping Theorem 8.33 of Chapter 8.6.1. □

We next show that the equilibrium point q^* guaranteed by Theorem 7.1 under assumption C5.1 implicitly optimizes a cost function implied by the control function u. Under assumption C5.2, the inverse functions u_j^{-1} exist and are strictly decreasing on $\mathbb{R}$. We hence can define $c_j : \mathbb{R} \to \mathbb{R}$ by

$$c_j(q_j) := -\int_0^{q_j} u_j^{-1}(\hat{q}_j)\, d\hat{q}_j, \qquad j \in N$$

Moreover c_j is strictly convex since $c_j''(q_j) = -1/u_j'(q_j) > 0$ under assumptions C5.1 and C5.2. Consider the optimization problem

$$\min_{q \in U} \; \sum_j c_j(q_j) \;+\; q^{\mathsf{T}} X q \;+\; q^{\mathsf{T}} \Delta\tilde{v} \tag{7.13}$$

where $\Delta\tilde{v} := \tilde{v} - v^{\text{ref}}$.

Theorem 7.2 (Optimality)**.** Suppose assumptions C5.1 and C5.2 hold. Then the unique equilibrium point $q^* \in U$ of (7.12) is the unique minimizer of (7.13).

Proof Let $C(q) := \sum_j c_j(q_j) + q^{\mathsf{T}} X q + q^{\mathsf{T}} \Delta\tilde{v}$ denote the objective function of (7.13). Since X is positive definite and c_j are strictly convex, $C(q)$ is strictly convex (and hence also continuous on $\mathbb{R}^N$). This implies, in particular, that if a minimizer of (7.13) exists (e.g., if U is bounded), then it is unique. It therefore suffices to show that q^* is an equilibrium point of (7.12) if and only if it is a minimizer of (7.13).

Since (7.13) is a convex problem, $q^* \in U$ is optimal if and only if

$$(\nabla C(q^*))^{\mathsf{T}} (q - q^*) \geq 0 \qquad \forall q \in U$$

Since each U_j in (7.11) is a box constraint, this means the optimal $q^* \in U$ is optimal if and only if (Exercise 7.4)

$$q_j^* \in (\underline{q}_j, \overline{q}_j) \quad \text{only if} \quad [\nabla C(q^*)]_j = 0 \tag{7.14a}$$

$$q_j^* = \underline{q}_j \quad \text{if} \quad [\nabla C(q^*)]_j > 0 \tag{7.14b}$$

$$q_j^* = \overline{q}_j \quad \text{if} \quad [\nabla C(q^*)]_j < 0 \tag{7.14c}$$

We have from (7.10) and (7.12)

$$\nabla C(q^*) = \nabla c(q^*) + 2Xq^* + \Delta\tilde{v} = \nabla c(q^*) + \left(v(q^*) - v^{\text{ref}}\right)$$

where $\nabla c(q^*) = (c_j'(q_j^*) = -u_j^{-1}(q_j^*), i \in N)$. Therefore

$$[\nabla C(q^*)]_j = -u_j^{-1}(q_j^*) + \left(v_j(q_j^*) - v_j^{\text{ref}}\right)$$

Since $u_j(v_j)$ is strictly decreasing in v_j we have

$$[\nabla C(q^*)]_j = 0 \iff u_j\left(v_j(q_j^*) - v_j^{\text{ref}}\right) = q_j^*$$

$$[\nabla C(q^*)]_j > 0 \iff u_j\left(v_j(q_j^*) - v_j^{\text{ref}}\right) < q_j^*$$

$$[\nabla C(q^*)]_j < 0 \iff u_j\left(v_j(q_j^*) - v_j^{\text{ref}}\right) > q_j^*$$

Substituting this into (7.14) shows that $q^* = \left[u\left(v(q^*) - v^{\text{ref}}\right)\right]_U$, i.e., q^* is the unique equilibrium point of (7.12). This shows that q^* is an equilibrium point of (7.12) if and only if it is a minimizer of (7.13). □

Remark 7.1. Theorem 7.2 shows that the control function in (7.12) implies an objective function $C(q)$ in (7.13) that an equilibrium implicitly optimizes. This is often referred to as reverse engineering. One can also start by designing an objective function $C(q)$ and deriving a control function as an iterative algorithm to solve the optimization problem (7.13). This is referred to as forward engineering; see e.g. [37, 38]. Often these algorithms require some communications among controllers at different buses but are guaranteed to converge under less stringent requirement than that in Theorem 7.1.

The formulation here imposes limits $[\underline{q}, \overline{q}]$ on the control q. It is pointed out in [53] that local memoryless control such as (7.12) may not be able to stabilize the equilibrium voltages $v(q^*)$ to within an apriori range $[\underline{v}, \overline{v}]$ (see Exercise 7.6). Alternative formulation imposes apriori limits $[\underline{v}, \overline{v}]$ on equilibrium voltages $v(q^*)$ but relaxes limits on the control q using control laws with internal state, see e.g. [53] □

7.3 Tree topology identification

In this section we illustrate the use of polar form power flow equation (4.27) of Chapter 4.3.2 and Theorem 4.10 of Chapter 4.2.6 for topology identification of radial networks from measurements of nodal voltage magnitudes. A distribution network typically consists of a meshed network with sectionalizing switches on some of the lines. At any time the switches are configured so that the operational network is a spanning tree with the substation at its root. We assume the system operator knows the topology of the meshed network, but may not know the switch configurations and hence the operational network. We first derive a linearized model using the polar form power flow equation (4.27). We then present two methods to identify the operational network, one making use of statistical properties of random voltage measurements and the other uses a graphical-model method.

7.3.1 Linearized polar-form AC model

Consider a radial network represented by a (connected) tree $G := (\overline{N}, E)$ with $N+1$ buses and $M = N$ lines and modeled by the polar-form power flow equations (4.27) reproduced here:

$$p_j = \sum_{k:k\sim j} \left(g^s_{jk} + g^m_{jk}\right)|V_j|^2 - \sum_{k:k\sim j} |V_j||V_k| \left(g^s_{jk}\cos\theta_{jk} + b^s_{jk}\sin\theta_{jk}\right), \quad j \in \overline{N} \tag{7.15a}$$

$$q_j = -\sum_{k:k\sim j} \left(b^s_{jk} + b^m_{jk}\right)|V_j|^2 - \sum_{k:k\sim j} |V_j||V_k| \left(g^s_{jk}\sin\theta_{jk} - b^s_{jk}\cos\theta_{jk}\right), \quad j \in \overline{N} \tag{7.15b}$$

We will linearize (7.15) under the following assumptions:

> C4.3: The series admittances $y^s_{jk} = y^s_{kj} = g^s_{jk} + \mathbf{i}b^s_{jk}$ (Assumption C4.1) and the shunt admittances $y^m_{jk} = y^m_{kj} = 0$ for all lines $(j,k) \in E$.
> C4.4: $g^s_{jk} > 0$ and $b^s_{jk} < 0$ for all $(j,k) \in E$.

Consider the "flat voltage profile" where $V^{\text{flat}}_j = \mu e^{\mathbf{i}\theta}$ for all $j \in \overline{N}$, so that the resulting power injection is $\left(p^{\text{flat}}, q^{\text{flat}}\right) = (0,0)$. Abuse notation and now let the variables $(\theta, |V|)$ denote *perturbations* around the flat voltage profile $V^{\text{flat}} = (\mu e^{\mathbf{i}\theta}, j \in \overline{N})$ and (p,q) denote the *perturbations* around $\left(p^{\text{flat}}, q^{\text{flat}}\right) = (0,0)$. Let $|\hat{V}| := (|V_j|, j \in N)$ and $(\hat{p}, \hat{q}) := (p_j, q_j, j \in N)$ denote the (perturbations of the) nodal voltage magnitudes and the (perturbations of the) power injections respectively at non-reference buses. Let $\hat{C}$ be the $N \times N$ reduced incidence matrix obtained from the node-by-line incidence matrix C by removing the first row of C corresponding to the reference bus 0. Partition the admittance matrix Y into the reference bus 0 and non-reference buses,

$Y = \begin{bmatrix} Y_{00} & y_0^{\mathsf{T}} \\ y_0 & \hat{Y} \end{bmatrix}$ where $\hat{Y} := \hat{C} D_y^s \hat{C}^{\mathsf{T}}$ is the reduced admittance matrix. Let $\hat{g}_0 := \mathrm{Re}(y_0)$ and $\hat{b}_0 := \mathrm{Im}(y_0)$ be the real and imaginary parts respectively of the first non-reference column of Y.

Then it is shown in Exercise 7.8 that the linearization of the polar form power flow equation (7.15) yields the following linear model for how $|\hat{V}|$ depends on the power injections $(\hat{p}, \hat{q})$ at non-reference buses:

$$|\hat{V}| \;=\; \hat{R}\hat{p} + \hat{X}\hat{q} - \hat{v}_0 \tag{7.16a}$$

where $\hat{R} := \hat{C}^{-\mathsf{T}} D_1 \hat{C}^{-1} \succ 0$, $\hat{X} := -\hat{C}^{-\mathsf{T}} D_2 \hat{C}^{-1} \succ 0$ are positive definite matrices, and $\hat{v}_0 := |V_0| \left(\hat{R}\hat{g}_0 + \hat{X}\hat{b}_0 \right)$. Here D_1 and D_2 are $N \times N$ diagonal matrices defined as:

$$\begin{aligned} D_g \;&:=\; \mathrm{diag}(g_l^s) \;\succ\; 0, & D_b \;&:=\; \mathrm{diag}(b_l^s) \;\prec\; 0 \\ D_1 \;&:=\; \left(D_g + D_b D_g^{-1} D_b \right)^{-1} \;\succ\; 0, & D_2 \;&:=\; \left(D_b + D_g D_b^{-1} D_g \right)^{-1} \;\prec\; 0 \end{aligned}$$

Let r_l and x_l denote the diagonal entries of D_1 and D_2 respectively. Then Theorem 4.10 says that $\hat{R}$ and $\hat{X}$ are given by:

$$\hat{R}_{jk} \;=\; \sum_{l \in \mathsf{P}_j \cap \mathsf{P}_k} r_l \;>\; 0, \qquad\qquad \hat{X}_{jk} \;=\; \sum_{l \in \mathsf{P}_j \cap \mathsf{P}_k} x_l \;>\; 0 \tag{7.16b}$$

where P_j denotes the unique path from bus 0 to bus j. Hence $\hat{R}_{jk}$ and $\hat{X}_{jk}$ are the sums of r_l and x_l respectively on the common segment of the unique paths from the reference bus 0 to buses j and k.

7.3.2 Covariance of voltage magnitudes and powers

The method of [40] to identify the operational network exploits statistical properties of voltage magnitudes. Define the covariance matrix $\Sigma_v := E[|\hat{V}| - E(|\hat{V}|)][(|V| - E(|V|)]^{\mathsf{T}}$ of voltage magnitudes $\hat{V}$ at non-reference buses and similarly the covariance matrices (Σ_p, Σ_q) of power injections $(\hat{p}, \hat{q})$, as well as cross-covariance matrices $\Sigma_{pq} := E(\hat{p} - E\hat{p})(\hat{q} - E\hat{q})^{\mathsf{T}}$ and $\Sigma_{qp} := E(\hat{q} - E\hat{q})(\hat{p} - E\hat{p})^{\mathsf{T}}$. Suppose the power injections at the same bus are positively correlated and those at different buses are uncorrelated, i.e. $(A[i,j]$ denotes the (i,j)th entry of matrix $A)$,

> C4.5: $\Sigma_p[j,j] > 0, \Sigma_q[j,j] > 0, \Sigma_{pq}[j,j] = \Sigma_{qp}[j,j] > 0$ for all j, and $\Sigma_p[j,k] = \Sigma_q[j,k] = \Sigma_{pq}[j,k] = \Sigma_{qp}[j,k] = 0$ for all $j \neq k$.

The key insight on which the method of [40] is based is explained in the next result. It says that the variance of voltage magnitude strictly increases as one moves away from the reference bus 0 where $|V_0|$ is fixed, and it also provides a way to identify the parent

of a bus. This is the consequence of (7.16a) that relates the covariance Σ_v of voltage magnitudes to the convariances of the power injections:

$$\Sigma_v \;=\; \hat{R}\Sigma_p\hat{R}^{\mathsf{T}} + \hat{X}\Sigma_q\hat{X}^{\mathsf{T}} + \hat{R}\Sigma_{pq}\hat{X}^{\mathsf{T}} + \hat{X}\Sigma_{qp}\hat{R}^{\mathsf{T}} \tag{7.17}$$

The jth diagonal entry $\Sigma_v[j,j] = E\left(|V_j| - E|V_j|\right)^2 =: \mathrm{var}(|V_j|)$ is the variance of voltage magnitude $|V_j|$ (deviation from its nominal value).

Recall that bus k is called a descendant of j if j is on the unique path from the reference bus 0 to bus k. Bus j is called a parent of k if $(j,k) \in E$ and k is a descendant of j. Let $\mathrm{var}(p_j)$ and $\mathrm{var}(q_j)$ denote the variance of the real and reactive power injections respectively at bus j, and $\mathrm{cov}(p_j,q_j) := E\left((p_j - Ep_j)(q_j - Eq_j)\right)$ denote the covariance of (p_j,q_j) at bus j.

Theorem 7.3 (Topology identification)**.** Suppose assumptions C4.3, C4.4 and C4.5 hold.

1 If a non-reference bus $j \in N$ is a descendant of bus i then $\mathrm{var}(|V_j|) > \mathrm{var}(|V_i|)$.
2 If bus i is a parent of bus $j \in N$ then the variance of the voltage magnitude difference $|V_i| - |V_j|$ is given by:

$$E\left((|V_i| - |V_i|) - E(|V_j| - |V_j|)\right)^2 \;=\; \sum_{k\in\mathsf{T}_j}\left(r_{ij}^2\mathrm{var}(p_k) + x_{ij}^2\mathrm{var}(q_k) + 2r_{ij}x_{ij}\mathrm{cov}(p_k,q_k)\right) \tag{7.18}$$

Proof For part 1, suppose first i is a parent of j. Theorem 4.10 and (7.16b) imply

$$\hat{R}_{jk} \;=\; \hat{R}_{ik} + r_{ij}, \qquad \hat{R}_{ik} \;=\; \sum_{l\in P_i} r_l, \qquad \text{if } k \in \mathsf{T}_j \tag{7.19a}$$

$$\hat{R}_{ik} \;=\; \hat{R}_{jk}, \qquad \text{if } k \notin \mathsf{T}_j \tag{7.19b}$$

Therefore the diagonal entry of the first matrix on the right-hand side of (7.17) yields

$$\begin{aligned}\left(\hat{R}\Sigma_p\hat{R}^{\mathsf{T}}\right)[j,j] - \left(\hat{R}\Sigma_p\hat{R}^{\mathsf{T}}\right)[i,i] \;&=\; \sum_k\sum_{k'}\Sigma_p[k',k]\left(\hat{R}_{jk'}\hat{R}_{jk} - \hat{R}_{ik'}\hat{R}_{ik}\right)\\ &=\; \sum_k \Sigma_p[k,k]\left(\hat{R}_{jk} + \hat{R}_{ik}\right)\left(\hat{R}_{jk} - \hat{R}_{ik}\right)\\ &=\; \sum_{k\in\mathsf{T}_j}\Sigma_p[k,k]\left(2\sum_{l\in\mathsf{P}_i} r_l + r_{ij}\right)r_{ij} \;>\; 0\end{aligned}$$

where the second equality follows because $\Sigma_p[k',k] = 0$ if $k' \neq k$, the last equality follows from (7.19), and the strict inequality follows because $\Sigma_p[k,k] > 0$ for all k and $r_l > 0$ for all l. Similarly $\left(\hat{X}\Sigma_q\hat{X}^{\mathsf{T}}\right)[j,j] > \left(\hat{X}\Sigma_q\hat{X}^{\mathsf{T}}\right)[i,i]$. The diagonal entry of the third matrix on the right-hand side of (7.17) yields

$$\left(\hat{R}\Sigma_{pq}\hat{X}^{\mathsf{T}}\right)[j,j] - \left(\hat{R}\Sigma_{pq}\hat{X}^{\mathsf{T}}\right)[i,i] \;=\; \sum_k \Sigma_{pq}[k,k]\left(\hat{R}_{jk}\hat{X}_{jk} - \hat{R}_{ik}\hat{X}_{ik}\right) \;>\; 0$$

where the equality follows from $\Sigma_{pq}[k',k]=0$ if $k'\neq k$ and the strict inequality uses (7.16b) and $\Sigma_{pq}[k,k]>0$. Similarly $\left(\hat{X}\Sigma_{qp}\hat{R}^{\mathsf{T}}\right)[j,j] > \left(\hat{X}\Sigma_{qp}\hat{R}^{\mathsf{T}}\right)[i,i]$. This shows that $\mathrm{var}(|V_j|)=\Sigma_v[j,j]>\Sigma_v[i,i]=\mathrm{var}(|V_i|)$ when i is a parent of j. When j is a descendant of i, the argument above applies pairwise on the path from i to j to conclude that $\Sigma_v[j,j]>\Sigma_v[i,i]$.

For part 2, suppose bus i is a parent of bus j then

$$E\left((|V_i|-E|V_i|)-(|V_j|)-|V_j|)\right)^2 \;=\; \Sigma_v[i,i]+\Sigma_v[j,j]-2\Sigma_v[i,j] \tag{7.20}$$

Consider $\Sigma_v[i,i]-\Sigma_v[i,j]$. The first matrix on the right-hand side of (7.17) yields

$$\left(\hat{R}\Sigma_p\hat{R}^{\mathsf{T}}\right)[i,i]-\left(\hat{R}\Sigma_p\hat{R}^{\mathsf{T}}\right)[i,j] \;=\; \sum_k \Sigma_p[k,k]\hat{R}_{ik}\left(\hat{R}_{ik}-\hat{R}_{jk}\right) \;=\; \sum_{k\in\mathsf{T}_j}\Sigma_p[k,k]\sum_{l\in\mathsf{P}_i} r_l(-r_{ij})$$

$$\left(\hat{R}\Sigma_p\hat{R}^{\mathsf{T}}\right)[j,j]-\left(\hat{R}\Sigma_p\hat{R}^{\mathsf{T}}\right)[j,i] \;=\; \sum_k \Sigma_p[k,k]\hat{R}_{jk}\left(\hat{R}_{jk}-\hat{R}_{ik}\right) \;=\; \sum_{k\in\mathsf{T}_j}\Sigma_p[k,k]\left(\sum_{l\in\mathsf{P}_i} r_l+r_{ij}\right)$$

where we have used $\Sigma_p[k',k]=0$ if $k'\neq k$ and (7.19). Summing these two expressions gives the part σ_1 of (7.20) due to the first matrix in (7.17):

$$\sigma_1 \;:=\; \left(\hat{R}\Sigma_p\hat{R}^{\mathsf{T}}\right)[i,i]+\left(\hat{R}\Sigma_p\hat{R}^{\mathsf{T}}\right)[j,j]-2\left(\hat{R}\Sigma_p\hat{R}^{\mathsf{T}}\right)[i,j] \;=\; r_{ij}^2\sum_{k\in\mathsf{T}_j}\Sigma_p[k,k]$$

Similarly the part σ_2 of (7.20) due to the second matrix in (7.17) is

$$\sigma_2 \;:=\; \left(\hat{X}\Sigma_q\hat{X}^{\mathsf{T}}\right)[i,i]+\left(\hat{X}\Sigma_q\hat{X}^{\mathsf{T}}\right)[j,j]-2\left(\hat{X}\Sigma_q\hat{X}^{\mathsf{T}}\right)[i,j] \;=\; x_{ij}^2\sum_{k\in\mathsf{T}_j}\Sigma_q[k,k]$$

The third matrix on the right-hand side of (7.17) yields:

$$\left(\hat{R}\Sigma_{pq}\hat{X}^{\mathsf{T}}\right)[i,i]-\left(\hat{R}\Sigma_{pq}\hat{X}^{\mathsf{T}}\right)[i,j] \;=\; \sum_k \Sigma_{pq}[k,k]\hat{R}_{ik}\left(\hat{X}_{ik}-\hat{X}_{jk}\right) \;=\; \sum_{k\in\mathsf{T}_j}\Sigma_{pq}[k,k]\sum_{l\in\mathsf{P}_i} r_l(-x_i$$

$$\left(\hat{R}\Sigma_{pq}\hat{X}^{\mathsf{T}}\right)[j,j]-\left(\hat{R}\Sigma_{pq}\hat{X}^{\mathsf{T}}\right)[j,i] \;=\; \sum_k \Sigma_{pq}[k,k]\hat{R}_{jk}\left(\hat{X}_{jk}-\hat{X}_{ik}\right) \;=\; \sum_{k\in\mathsf{T}_j}\Sigma_{pq}[k,k]\left(\sum_{l\in\mathsf{P}_i} r_l+r\right.$$

and hence

$$\sigma_3 \;:=\; \left(\hat{R}\Sigma_{pq}\hat{X}^{\mathsf{T}}\right)[i,i]+\left(\hat{R}\Sigma_{pq}\hat{X}^{\mathsf{T}}\right)[j,j]-2\left(\hat{R}\Sigma_{pq}\hat{X}^{\mathsf{T}}\right)[i,j] \;=\; r_{ij}x_{ij}\sum_{k\in\mathsf{T}_j}\Sigma_{pq}[k,k] \tag{7.21}$$

Similarly

$$\sigma_4 \;:=\; \left(\hat{X}\Sigma_{qp}\hat{R}^{\mathsf{T}}\right)[i,i]+\left(\hat{X}\Sigma_{qp}\hat{R}^{\mathsf{T}}\right)[j,j]-2\left(\hat{X}\Sigma_{qp}\hat{R}^{\mathsf{T}}\right)[i,j] \;=\; r_{ij}x_{ij}\sum_{k\in\mathsf{T}_j}\Sigma_{qp}[k,k]$$

Summing these expressions yields

$$\Sigma_v[i,i]-\Sigma_v[i,j] \;=\; \sum_{k=1}^{4}\sigma_k \;=\; \sum_{k\in\mathsf{T}_j}\left(r_{ij}^2\Sigma_p[k,k]+x_{ij}^2\Sigma_q[k,k]+2r_{ij}x_{ij}\Sigma_{pq}[k,k]\right)$$

proving (7.18). □

Part 1 of Theorem 7.3 allows us to identify a leaf node j as one that has the largest $\text{var}(p_j)$. Part 2 of the theorem allows us to identify j's parent i as one that most closely satisfies (7.18). The theorem therefore suggests the following iterative method to identify the topology of the operational network from empirical estimates of variances $\text{var}(|V_j|)$, $\text{var}(p_j)$, $\text{var}(q_j)$ and the covariance $\text{cov}(p_j, q_j)$ of voltage magnitudes and power injections at each bus j. In each iteration the algorithm identifies a leaf node j among the set of unidentified nodes (whose parents have not been identified), and then uses (7.18) to identify j's parent i. Then the algorithm removes node j from the set of unidentified nodes and the cycle repeats. The parent identification step that uses (7.18) needs the knowledge of the underlying meshed network topology and its line parameters (r_{ij}, x_{ij}).

7.3.3 Graphical-model method

7.4 Bibliographical notes

7.5 Problems

Chapter 7.1

Exercise 7.1 (State estimation)**.** Derive the optimal state estimate $\hat{\Delta}x$ in (7.7).

Exercise 7.2 (State estimation)**.** Suppose $x_i \in \mathbb{R}, i = 1, \ldots, n$, are n iid samples of a scalar random variable x with mean μ and variance σ^2. Consider the sample mean and sample variance:

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Show that (treating x_i as iid random variables with the same distribution as x):

1 For each i, $E(\bar{x}x_i) = E\bar{x}^2 = \mu^2 + \sigma^2/n$.
2 $E(\bar{\sigma}^2) = \frac{n-1}{n}\sigma^2$, i.e., $\bar{\sigma}^2$ is a biased estimator of σ^2 with mean smaller than σ^2.
3 An unbiased estimator is the scaled sample variance (i.e., $E(\hat{\sigma}^2) = \sigma^2$):

$$\hat{\sigma}^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Chapter 7.2

Exercise 7.3 (Induced matrix norm)**.** For any $n \times n$ matrix A show that the induced norm

$$\|A\|_2 := \max_{\|x\|_2=1} \|Ax\|_2 \;=\; \sigma_{\max}(A)$$

where $\sigma_{\max}(A)$ is the largest singular value of A.

Exercise 7.4. [Local volt/var control] Let $U_j := \{x_j : \underline{x}_j \le x_j \le \overline{x}_j\}$, $j = 1, \ldots, n$, and $U := U_1 \times \cdots \times U_n$. Let $f : \mathbb{R}^n \to \mathbb{R}^n$. Show that

$$x^* \in U, \qquad f^{\mathsf{T}}(x^*)\,(x - x^*) \ge 0 \quad \forall x \in U \tag{7.22}$$

if and only if

$$x_j^* \in (\underline{x}_j, \overline{x}_j) \quad \text{only if} \quad f_j(x^*) = 0 \tag{7.23a}$$

$$x_j^* = \underline{x}_j \quad \text{if} \quad f_j(x^*) > 0 \tag{7.23b}$$

$$x_j^* = \overline{x}_j \quad \text{if} \quad f_j(x^*) < 0 \tag{7.23c}$$

Exercise 7.5. [Local volt/var control] Let the control function in (7.12) be $u_j(v_j) = -\gamma_j v_j$ with $\gamma_j > 0$. Derive the condition for convergence and the resulting cost function $C(q)$.

Exercise 7.6. [Local volt/var control] Suppose it is desirable to asymptotically stabilize the voltages v to within a certain bounds $[\underline{v}, \overline{v}]$ while maintaining the limits $[\underline{q}, \overline{q}]$ on the reactive power.

1 Show that there exists $\tilde{v}$ such that no equilibrium point of (7.12) can lie in $[\underline{v}, \overline{v}]$.
2 Fix $\tilde{v}$. For each bus j, find the maximum $\underline{v}_j$ and minimum $\overline{v}_j$ for which it is possible to asymptotically stabilize v_j to within $[\underline{v}_j, \overline{v}_j]$. Note that it may not be possible for v_j to attain $\underline{v}_j$ (or $\overline{v}_j$) *smultaneously for all* j.

Exercise 7.7 (Voltage control with batteries)**.** We are given a set of battery locations on a distribution system and have to decide the optimal energy capacity for each battery and its optimal charging/discharging rate during operation. This exercise formulates the problem as a two-stage optimization with recourse.

Consider a feeder with $N+1$ buses indexed by i with $i = 0$ denoting the substation bus where voltage is fixed and real and reactive powers are variables. Candidate battery

locations are a subset N_b of the buses. Consider a typical day divided into T time periods indexed by t. The parameters are:

- v_i^{ref}: the nominal voltages at buses i;
- c_i: the unit cost of battery capacities at buses i; and
- $(p_i(t;\omega), q_i(t;\omega))$: the random real and reactive injections at non-substation buses $i \neq 0$ at each time t where the uncertainty is indexed by ω defined over a suitable probability space.
- $\underline{u}_i \leq \overline{u}_i$: the lower and upper bounds on battery charging rate $u_i(t,\omega)$ at bus i.

Our objective is to decide how much battery capacities to install at $i \in N_b$ and how to operate them so as to minimize a weighted sum of capital cost (due to c_i) for installation and voltage deviations from their nominal values during operation. The design decisions are:

- $B_i \geq 0$: battery capacity to be installed at bus i (battery will be installed at bus i if $B_i > 0$).
- $u_i(t;\omega)$: the charging (when $u_i(t;\omega) \geq 0$) or discharging (when $u_i(t;\omega) \leq 0$) rate of the battery at bus i at time t so that the net real injection at bus i is $p_i(t;\omega) - u_i(t;\omega)$ at time t.

The battery capacities B_i must be determined before the realization of ω, but the charging rates $u_i(t;\omega)$ can be chosen in response after ω is realized (and batteries installed). Formulate this problem as a two-stage optimization where B_i are the first-stage decisions and $u_i(t;\omega)$ are the second-stage decisions adapted to ω. Incorporate appropriate power flow models and voltage constraints from Chapters 4 or 5.

Chapter 7.3

Exercise 7.8 (Linearized polar form)**.** Consider a radial network for which G is a (connected) tree. Suppose assumptions C4.3 and C4.4 hold.

1 Show that linearization of the polar form of the power flow equation (7.15) around $(V^{\text{flat}}, p^{\text{flat}}, q^{\text{flat}})$ is given by (7.16) where $\hat{R}$ and $\hat{X}$ are positive definite matrices. Assume without loss of generality that $\mu = 1$.
2 Show that if $g^s_{jk} = 0$ for all (j,k), the linearized model reduces to the DC power flow model (4.55a) with $|V_j| = \mu$ for all $j \in \overline{N}$.

Exercise 7.9 (Topology identification)**.** Consider the network in Exercise 7.8. Denote the variance of the voltage magnitude difference at buses j and k by $v(j,k) :=$

$E\left((|V_j| - |V_k|) - E(|V_j| - |V_k|)\right)^2$. Consider any non-reference bus $j \in N$. Show that among buses k that are not descendants of j, j's parent uniquely minimizes $v(j,k)$, i.e., if i_j is the (unique) parent of j then

$$i_j = \arg\min_{k \notin \mathsf{T}_j} v(j,k)$$

Part II

Power flow optimization

8 Smooth convex optimization

In this chapter we study the following questions:

1 How to specify an optimization problem (Ch 8.1 and 8.2)?
2 How to characterize its optimal solutions and determine if one exists (Ch 8.3 and 8.4)?
3 How to compute an optimal solution iteratively when one exists (Ch 8.5)?
4 How to ensure the correctness of the computation (Ch 8.6)?

Specifically we formulate convex optimization problems (Chapter 8.1) and introduce some of the most useful tools for convex analysis (Chapter 8.2). We develop a general theory to characterize optimal solutions and provide sufficient conditions for their existence (Chapter 8.3). We then apply the general theory to special classes of convex optimization problems widely used in applications (Chapter 8.4). We describe iterative algorithms based on optimality conditions of Chapter 8.3 for solving these problems (Chapter 8.5) and explain basic techniques for analyzing their convergence (Chapter 8.6).

Convexity is a simplifying structure that enables a rich theory on algorithm design and analysis for convex optimization. Even though optimal power flow problems are nonconvex, convex optimization theory is useful for two reasons. First, iterative algorithms that have been designed and analyzed for convex problems are often used also for solving nonconvex problems. Unlike for convex problems, there is typically no guarantee on optimality or convergence for nonconvex problems, but they often perform well nonetheless. Second, an important method to deal with a nonconvex problem is solving its convex relaxation where an approximate convex problem is solved instead. We will study optimal power flow problems in Chapter 9 and their convex relaxations in Chapters 10 and 11.

8.1 Convex optimization

A convex program is defined by a convex set and a convex function. We start by defining some basic concepts that are used both in this chapter on smooth convex optimization and in Chapter 12.1 on nonsmooth convex optimization.

8.1.1 Affine hull and relative interior

Consider a nonempty set $X \in \mathbb{R}^n$. A point $x \in \mathbb{R}^n$ is called *a closure point or a limit point of* X if there is a sequence $\{x_k \in X\} \subseteq X$ that converges to x. The *closure* of X, denoted by $\mathrm{cl}(X)$, is the set of all closure points of X. We say that X is *closed* if $\mathrm{cl}(X) = X$, i.e., X contains all its limit points. The closure of X is the smallest closed set that contains X. The set X is called *open* if its complement is closed, i.e., $\{x \in \mathbb{R}^n : x \notin X\}$ is closed. It is called *bounded* if there exists a finite b such that $\|x\| \leq b$ for all $x \in X$.[1] It is called *compact* if it is closed and bounded.

An alternative approach to defining open and closed sets is to define a topological space by specifying all subsets of an ambient set Y that are open in that topological space. In this approach the empty set $\emptyset$ and the ambient set Y are always defined to be open sets in any topology. When the ambient set $Y := \mathbb{R}^n := (-\infty, \infty)^n$, $\mathbb{R}^n$ is both open and closed in the topological space regardless of topology. This is consistent with the definition above in terms of limit points (under the usual topology induced by a norm) because, e.g., the sequence $x_k := (k, \ldots, k)$ does not converge as $k \to \infty$ since it tends to $(\infty, \ldots, \infty)$ which is not a point in $Y := \mathbb{R}^n$. If $Y := \mathbb{R}^n \cup \{-\infty, \infty\}^n$ is an extended space under the usual topology induced by a norm, however, $\mathbb{R}^n$ is open but not closed.

A point x is called an *interior point* of X if there exists an open neighborhood of x that is contained in X, i.e., there is $\epsilon > 0$ such that $B_\epsilon(x) := \{y : \|y - x\| < \epsilon\} \subseteq X$. The *interior* of X, denoted by $\mathrm{int}(X)$, is the set of all interior points of X. A point $x \in \mathrm{cl}(X)$ that is not an interior point of X is called a *boundary point* of X. A boundary point may or may not be in X. The set of all boundary points is called the *boundary* of X.

A concept that is important in convex optimization theory is relative interior of a set X, which we now define. A set Y is called an *affine set* if Y contains all the lines that pass through pairs of distinct points $x, y \in Y$. The *affine hull* of X, denoted by $\mathrm{aff}(X)$, is the intersection of all affine sets containing X. The affine hull $\mathrm{aff}(X)$ is itself an affine set. A point $x \in X \subseteq \mathbb{R}^n$ is called a *relative interior point* of X if there exists an open neighborhood $B_\epsilon(x) \subseteq \mathbb{R}^n$ such that $B_\epsilon(x) \cap \mathrm{aff}(X) \subseteq X$, i.e., x is an interior point of X relative to $\mathrm{aff}(X)$. The set of all relative interior points of X is called the *relative interior* of X, denoted by $\mathrm{ri}(X)$. The set X is called *relatively open* if $\mathrm{ri}(X) = X$. A

[1] The norm $\|\cdot\|$ defines the usual topology. Since all norms are equivalent on a finite dimensional space, these concepts remain the same regardless of topology.

point $x \in \mathrm{cl}(X)$ that is not a relative interior point is called a *relative boundary point* of X. The set of all relative boundary points of X is called the *relative boundary* of X.

Example 8.1 ($\mathrm{ri}(X)$)**.** Consider the set $X := \{x \in \mathbb{R}^3 : x_1 = x_2, x_1 \in (0,1), x_2 \in (0,1)\}$. It is not an affine set since x_1, x_2 are bounded. Its affine hull is $\mathrm{aff}(X) = \{x \in \mathbb{R}^3 : x_1 = x_2\}$. The set X is not open in $\mathbb{R}^3$ as it has no interior point relative to $\mathbb{R}^3$. It is relatively open because every point $x \in X$ is an interior point relative to $\mathrm{aff}(X)$ and hence $\mathrm{ri}(X) = X$. The closure of X is $\mathrm{cl}(X) = \{x \in \mathbb{R}^3 : x_1 = x_2, x_1 \in [0,1], x_2 \in [0,1]\}$. □

8.1.2 Convex set

A set is called convex if, given any two points in the set, every point in between lies in the set.

Definition 8.1 (Convex set)**.** A set $D \subseteq \mathbb{R}^n$ is *convex* if, given any $x, y \in D$,

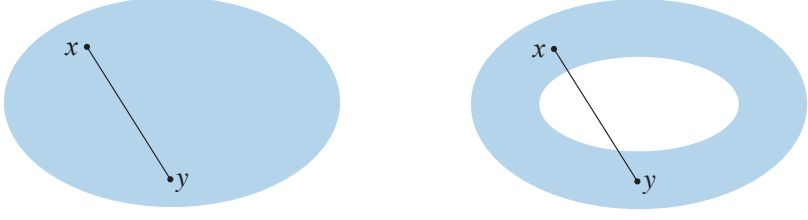
$$\alpha x + (1-\alpha) y \ \in \ D, \qquad \forall \alpha \in [0,1]$$

□

For instance, if D is an open set, then for any $x_0 \in D$ there exists $r > 0$ such that the r-ball around x_0,

$$B_r(x_0) := \{\ x \in D \mid \|x - x_0\|_2 \leq r\ \}$$

is contained in D, where $\|x\|_2 := \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$ is the Euclidean norm. Moreover $B_r(x_0)$ is convex for any $r > 0$, $x_0 \in D$. The definition is illustrated in Figure 8.1.



(a) Convex set. (b) Nonconvex set.

Figure 8.1 Definition of a convex set: every point in between two points in the set lies in the set.

Definition 8.2 (Convex hull)**.** Let $X \subseteq \mathbb{R}^n$ be a nonempty set. The *convex hull* of X, denoted $\mathrm{conv}(X)$, is the intersection of all convex sets containing X. □

The convex hull $\mathrm{conv}(X)$ of any set $X \subseteq \mathbb{R}^n$ is contained in its affine hull $\mathrm{aff}(X)$. When X is a convex set, the *dimension* of X is defined to be the dimension of $\mathrm{aff}(X)$.

Three types of convex sets are the most useful in engineering applications (the proof that these sets are convex is left as an exercise):

1 *Polyhedral set* $H \subseteq \mathbb{R}^n$. A polyhedral set is specified by affine equalities or inequalities. A *hyperplane* is a set $H_1 := \{x \in \mathbb{R}^n : c^\mathsf{T} x = b\}$ specified by an affine equality with $c \in \mathbb{R}^n$ and $b \in \mathbb{R}$. A *polyhedral set*, or a *polyhedron*, is a set $H_2 := \{x \in \mathbb{R}^n : Ax \leq b\}$ specified by a finite number of affine inequalities. We may call the intersection $H_1 := \{x \in \mathbb{R}^n : Ax = b\}$ of hyperplanes with $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ a hyperplane or a polyhedron.

2 *Second-order cone (SOC)* $K_{soc} \subseteq \mathbb{R}^n$. A second-order cone (SOC) is defined as:

$$K_{\text{soc}} := \{x \in \mathbb{R}^n \mid \|x^{n-1}\|_2 \leq x_n\}, \quad n \geq 2 \tag{8.1}$$

where $x =: (x^{n-1}, x_n)$, i.e., x^{n-1} denotes the subvector of x consisting of its first $n-1$ entries. A ball $B_{x_n}(0) := \{x^{n-1} : \|x^{n-1}\|_2 \leq x_n\}$ in $\mathbb{R}^{n-1}$ centered at the origin for a fixed radius x_n is a cross section of the second-order cone. SOC K_{soc} is a special type of convex set called a convex cone. We will study in more detail cones, convex cones, and second-order cones in Chapter 8.2.1.

3 *Semidefinite cones* $K_{psd}, K_{nsd} \subset \mathbb{S}^n$. A real matrix $X \in \mathbb{R}^{n \times n}$ is *symmetric* if $X = X^\mathsf{T}$, i.e., $X_{ij} = X_{ji}$ for all $i, j = 1, \ldots, n$. Let $\mathbb{S}^n \subset \mathbb{R}^{n \times n}$ denote the set of all real symmetric matrices. It is a vector space (or linear space) over the field $\mathbb{R}$ (see Appendix A.1.1 for definitions of vector space and subspace). A real matrix X is *positive semidefinite* (psd) if X is symmetric and $x^\mathsf{T} X x = \sum_{i,j} X_{ij} x_i x_j \geq 0$ for all $x \in \mathbb{R}^n$. Given a symmetric matrix $X \in \mathbb{R}^{n \times n}$ the following are equivalent:
   1 X is positive semidefinite.
   2 All eigenvalues of X are nonnegative.
   3 $X = BB^\mathsf{T}$ for some matrix $B \in \mathbb{R}^{n \times m}$ and some natural number m.

   A real matrix X is *negative semidefinite* (nsd) if $-X$ is psd. We denote the set of all positive semidefinite matrices by K_{psd} and the set of all negative semidefinite matrices by K_{nsd}. We write $X \in K_{\text{psd}}$ or $X \succeq 0$ to denote that X is positive semidefinite. Similarly $X \in K_{\text{nsd}}$ or $X \preceq 0$ denotes that X is negative semidefinite. These sets are special convex sets called *semidefinite cones* in the vector space $\mathbb{S}^n \subset \mathbb{R}^{n \times n}$ over the field $\mathbb{R}$. In Chapter 8.2.2 we extend these notions to the complex domain and treat the set $\mathbb{S}^n \subset \mathbb{C}^{n \times n}$ of complex Hermitian matrices as a vector space over the field $\mathbb{R}$ (not $\mathbb{C}$), define the inner product in $\mathbb{S}^n$, and the semidefinite cones of complex matrices in the vector space $\mathbb{S}^n$.

Given these three basic convex sets we can create other convex sets through simple *convexity-preserving* operations. Let $\mathbb{X}$ and $\mathbb{Y}$ be linear subspaces. For example $\mathbb{X} := \mathbb{R}^n$ and $\mathbb{Y} := \mathbb{R}^m$.

1 *Linear transformation:* Let $f : \mathbb{X} \to \mathbb{Y}$ be linear.
   1 If $A \subseteq \mathbb{X}$ is convex then $f(A) := \{f(x) : x \in A\} \subseteq \mathbb{Y}$ is convex.
   2 If $B \subseteq \mathbb{Y}$ is convex then $f^{-1}(B) = \{x : f(x) \in B\} \subseteq \mathbb{X}$ is convex.

2 *Arbitrary direct product:* Let $A \subseteq \mathbb{X}$, $B \subseteq \mathbb{Y}$ be convex. Then $A \times B := \{(x,y) : x \in A, y \in B\}$ is convex. In fact the direct product of an arbitrary collection of (e.g., uncountably many) convex sets is convex.
3 *Finite sum:* Let $A, B \subseteq \mathbb{X}$ be convex. Then $A + B := \{a + b : a \in A, b \in B\}$ is convex. Therefore the sum of any finite number of convex sets is convex.
4 *Arbitrary intersection:* Let $A, B \subseteq \mathbb{X}$ be convex. Then the intersection $A \cap B$ is convex, even if the intersection is empty. In fact the intersection of an arbitrary collection of (e.g., uncountably many) convex sets is convex.

The proof that these set operations preserve convexity is left as an exercise. If A, B are convex, then $A \cap B$ is convex, but the converse may not hold; e.g., $A := \{x : x \geq 0\} \cup \{x : x \leq 0\} \subseteq \mathbb{R}^n$ and $B := \{x : x \geq 0\} \subseteq \mathbb{R}^n$. In contrast to intersection the union of two convex sets can be nonconvex.

Example 8.2. Consider the ellipsoid

$$E := \{x \in \mathbb{R}^n \mid x^\mathsf{T} A x \leq c\}$$

where $A \in \mathbb{R}^{n \times n}$ is a psd matrix. It is easy to show that E is convex by verifying Definition 8.1. In this example we show that E is convex by deriving it from the application of convexity-preserving operations on a convex set. Since A is psd it can be expressed as $A := BB^\mathsf{T}$ for some $B \in \mathbb{R}^{n \times m}$. Hence $x^\mathsf{T} A x = x^\mathsf{T} B B^\mathsf{T} x = \|B^\mathsf{T} x\|_2^2$.

Let $y = B^\mathsf{T} x$. Then the set $C := \{(y,t) \in \mathbb{R}^{m+1} \mid \|y\|_2 \leq t\}$ is a (convex) SOC. Hence the set $D := \{y \in \mathbb{R}^m \mid \|y\|_2 \leq c\}$ is convex since it is the intersection of two convex sets:

$$D = C \cap (\mathbb{R}^m \times \{t = c\})$$

Then $E = f^{-1}(D)$ where $f(x) := B^\mathsf{T} x$ is a linear function from $\mathbb{R}^n$ to $\mathbb{R}^m$. Hence E is convex as desired. Note that if $c < 0$ then $E = \emptyset$ which is convex. □

8.1.3 Derivative, directional derivative and partial derivative

In this subsection we review different notions of derivatives of real-valued functions f on $\mathbb{R}^n$ (see Chapter A.9 for more details).

Consider a real-valued function $f : X \to \mathbb{R}$ where $X \subseteq \mathbb{R}^n$ is an open set. At each $x \in X$ and for each $v \in \mathbb{R}^n$ the one-sided *directional derivative of f at x in the direction v* is defined as

$$df(x;v) := \lim_{\substack{t \in \mathbb{R}_+ \\ t \downarrow 0}} \frac{f(x+tv) - f(x)}{t}$$

provided the limit exists, possibly $\pm\infty$. Since X is open and f is real-valued, $df(x;v)$ if exists is always real valued for any $v \in \mathbb{R}^n$. The function f is said to be *differentiable*

at $x \in X$ if the directional derivative $df(x;v)$ exists at x for all directions $v \in \mathbb{R}^n$ and is a linear function of v, i.e., if there exists a vector $m_x \in \mathbb{R}^n$ such that, for all $v \in \mathbb{R}^n$,

$$df(x;v) \;:=\; \lim_{\substack{t \in \mathbb{R}_+ \\ t \downarrow 0}} \frac{f(x+tv) - f(x)}{t} \;=\; m_x^\top v$$

In this case the column vector $m_x \in \mathbb{R}^n$ is called the *gradient or derivative of* f *at* $x \in X$ and denoted by $\nabla f(x)$. If f is differentiable at every $x \in X$ then f is called *differentiable* on X.

At each $x \in X$ and for the unit vector $e_j \in \{0,1\}^n$ that has a single 1 in its jth position, if the directional derivatives $df(x;e_j)$ and $df(x;-e_j)$ exist in both directions and are equal, then they are called the *partial derivative of* f *at* $x \in X$ with respect to x_j and denoted by $\frac{\partial f}{\partial x_j}(x)$:

$$\frac{\partial f}{\partial x_j}(x) \;:=\; \lim_{\substack{t \in \mathbb{R} \\ t \to 0}} \frac{f(x+te_j) - f(x)}{t}$$

In this case f is called *partially differentiable at* $x \in X$ with respect to x_j. The row vector of partial derivatives of f at $x \in X$ is

$$\frac{\partial f}{\partial x}(x) \;:=\; \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) & \cdots & \frac{\partial f}{\partial x_n}(x) \end{bmatrix}$$

If f is partially differentiable at all $x \in X$ then it is called *partially differentiable on* X. The partial derivative $\frac{\partial f}{\partial x}(x)$ describes the behavior of f at x only along the coordinate axes whereas the derivative $\nabla f(x)$ describes its behavior in all directions. If f is differentiable then it is partially differentiable, but the converse does not generally hold. If f is not only partially differentiable but $\frac{\partial f}{\partial x}(x)$ is also continuous at x, then the converse holds at $x \in X$. Such an f is called *continuously differentiable at* x. If f is continuously differentiable at all $x \in X$ then it is *continuously differentiable on* X.

Lemma 8.1 (Differentiability and partial differentiability)**.** Consider a real-valued function $f : X \to \mathbb{R}$ where $X \subseteq \mathbb{R}^n$ is an open set.

1 If f is differentiable at $x \in X$ then it is partially differentiable at x. Moreover its gradient $\nabla f(x)$ is given by

$$\nabla f(x) \;=\; \left[\frac{\partial f}{\partial x}(x)\right]^\top$$

2 If f is continuously differentiable at $x \in X$ then it is differentiable at x.

The following example shows that a partially differentiable function may not be differentiable when the partial derivative $\frac{\partial f}{\partial x}(x)$ is discontinuous at x. Indeed a partially differentiable function may not even be continuous at all $x \in X$. A continuously differentiable function is always continuous.

Example 8.3. 1 Consider $f : \mathbb{R}^2 \to \mathbb{R}$ defined by:

$$f(x,y) := \begin{cases} 0 & \text{if} \quad xy = 0 \\ 1 & \text{if} \quad x \neq 0,\ y \neq 0 \end{cases}$$

Its partial derivative on the axes exists only at the origin where $\frac{\partial f}{\partial (x,y)}(0,0) = [0\ 0]$. The function f is however not differentiable at $(0,0)$ as it is discontinuous at every point on the axes. Clearly $\frac{\partial f}{\partial (x,y)}$ is discontinuous at the origin.

2 Consider $f : \mathbb{R}^2 \to \mathbb{R}$ defined by:

$$f(x,y) := \begin{cases} \frac{x^a y^a}{x^{2a}+y^{2a}} & \text{if} \quad (x,y) \neq (0,0) \\ 0 & \text{if} \quad (x,y) = (0,0) \end{cases}$$

It is discontinuous at the origin along the line $x = y$ (Exercise 8.3). Therefore the directional derivative of f along $x = y$ does not exist. □

Hence f is differentiable at $x \in X$ if and only if $df(x;v) = v^\mathsf{T} \nabla f(x) = \frac{\partial f}{\partial x}(x)\, v$ for all $v \in \mathbb{R}^n$. For convex but non-differentiable functions, derivatives are generalized in Chapter 12.3.2 to subdifferentials.

For a vector-valued function $f : \mathbb{R}^n \to \mathbb{R}^m$ that maps an $x \in \mathbb{R}^n$ to a vector $f(x) \in \mathbb{R}^m$, the *Jacobian* $J(x) := \left[\frac{\partial f}{\partial x}(x)\right]$ *of* f *at* x is the $m \times n$ matrix whose ijth entry $J_{ij}(x) := \frac{\partial f_i}{\partial x_j}(x)$ is the partial derivative of f_i with respect to x_j evaluated at x. The *gradient or derivative of* f *at* x is $\nabla f(x) := J^\mathsf{T}(x)$.

8.1.4 Convex function

Definition 8.3 (Convex function)**.** A function $f : X \to \mathbb{R}$ defined over a convex set $X \subseteq \mathbb{R}^n$ is *convex* if, for all $x, y \in X$ and all $\alpha \in [0,1]$,

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$$

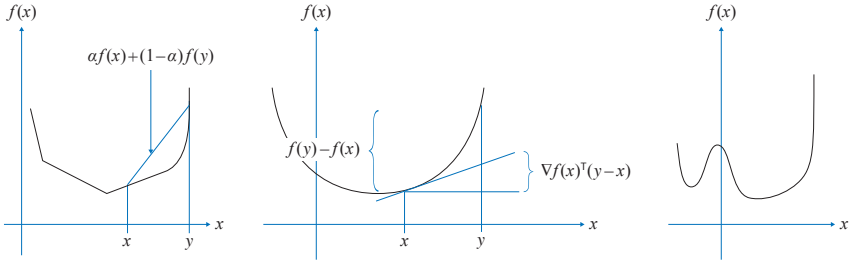
It is *strictly convex* if the inequality is strict for $x \neq y$ and $\alpha \in (0,1)$. A vector-valued function $f : X \to \mathbb{R}^m$ is *convex* or *strictly convex* if every component $f_i : X \to \mathbb{R}$ is convex or strictly convex respectively. A function f is *concave* (*strictly concave*) if $-f$ is convex (strictly convex). □

The definition says that, for a scalar-valued function f, the straight line connecting $f(x)$ and $f(y)$ lies above f between x and y, or equivalent, the linear approximation of f is always an underestimate, as illustrated in Figure 8.2(a).

Example 8.4. If $f(x) = x^2$ then for any x, y and $\alpha \in [0,1]$

$$\alpha f(x) + (1-\alpha) f(y) - f(\alpha x + (1-\alpha)y) = \alpha(1-\alpha)\,(x-y)^2 \ > \ 0$$

for $x \neq y$ and $\alpha \in (0,1)$. Hence f is strictly convex. □



(a) Convex function. (b) Convex function. (c) Nonconvex function.

Figure 8.2 Convex function: The straight line connection $f(x)$ and $f(y)$ lies above f. The linear approximation of a differentiable convex function f lies below f.

Checking if a function is convex by verifying the convexity definition is often difficult. The following theorem provides three different ways to check the convexity of a function. Consider $f : X \to \mathbb{R}$ over a convex domain $X \subseteq \mathbb{R}^n$. Let $\nabla f(x)$ denote the *column* vector of partial derivatives of f (whereas $\frac{\partial f}{\partial x}$ denotes the row vector of partial derivatives). Let

$$\nabla^2 f(x) := \frac{\partial^2 f}{\partial x^2} := \left[\frac{\partial^2 f}{\partial x_i \partial x_j}\right]$$

denote the $n \times n$ symmetric Hessian matrix.

Theorem 8.2 (Convex function)**.** Consider a function f defined on a convex open set $X \subseteq \mathbb{R}^n$. The function f is convex if and only if any one of the following holds:

1 For $x \in X$ and all $v \in \mathbb{R}^n$ the function

$$g(t) := f(x + tv) \tag{8.2}$$

is convex on $\{t \in \mathbb{R} \mid x + tv \in X\}$.

2 For a differentiable function f,

$$f(y) - f(x) \geq \nabla f(x)^\mathsf{T}(y - x), \qquad \forall x, y \in X \tag{8.3}$$

3 For a twice differentiable function f,

$$\nabla^2 f(x) \geq 0, \qquad \forall x \in X$$

i.e., the Hessian matrix is positive semidefinite (all eigenvalues are nonnegative).

The condition in Theorem 8.2.1 does not require differentiability of f and says that, if we take any cross section of the surface f defined by (x, v), i.e., from x in the direction of v or $-v$, the corresponding *scalar* function $g(t)$ is convex. The first-order condition in Theorem 8.2.2 says that the function f always lies above its linear approximation, i.e., $f(y)$ is always greater than or equal to the tangent plane to f at any point x. This

is illustrated in Figure 8.2(b). The second-order condition in Theorem 8.2.3 roughly says that the gradient at any point x is increasing around x.

Proof of Theorem 8.2 1 Suppose f is convex. Fix any $x \in X$ and any $v \in \mathbb{R}^n$. We will show that $g(t) := f(x+tv)$ is convex in t, i.e., for $s < u$ such that $x+sv$ and $x+uv$ are both in X, we have, for any $t := \alpha s + (1-\alpha)u$ with $\alpha \in [0,1]$,

$$g(t) \;\le\; \alpha g(s) + (1-\alpha) g(u)$$

From Figure 8.3 we have

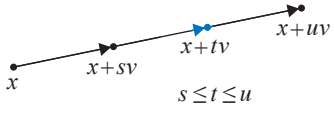


Figure 8.3 Proof of Theorem 8.2.1.

$$x+tv \;=\; \alpha(x+sv) + (1-\alpha)(x+uv)$$

Hence, since f is convex,

$$g(t) \;=\; f(x+tv) \;=\; f\left(\alpha(x+sv) + (1-\alpha)(x+uv)\right) \;\le\; \alpha g(s) + (1-\alpha) g(u)$$

i.e., g is convex. Conversely suppose g is convex but f is not, i.e., there exists two points $x, y \in X$ and a point $z := (1-\alpha)x + \alpha y$, $\alpha \in [0,1]$, in between such that

$$f(z) \;>\; (1-\alpha) f(x) + \alpha f(y)$$

Define $g(t) := x + tv$ where $v := y - x$. Then $z = x + \alpha v$ and, since g is convex,

$$f(z) \;=\; g(\alpha) \;\le\; (1-\alpha) g(0) + \alpha g(1) = (1-\alpha) f(x) + \alpha f(y)$$

contradicting that f is not convex.

2 We first prove the result for a scalar differentiable function $g : \mathbb{R} \to \mathbb{R}$. Then we use the result to prove the theorem for a differentiable function $f : X \to \mathbb{R}$ where $X \subseteq \mathbb{R}^n$ with $n \ge 1$.

Consider first $g : \mathbb{R} \to \mathbb{R}$. We prove that the following are equivalent:

(a) g is convex.

(b) $g(t) - g(s) \ge g'(s)(t-s)$ for any $s \ne t \in \mathbb{R}$.

(c) $g'(t) \ge g'(s)$ for any $t \ge s$ in $\mathbb{R}$, i.e. g has nondecreasing slope.

Suppose (a): g is convex. Fix any $s, t \in X$. For any $\alpha \in [0,1]$ we have $g(s+\alpha(t-s)) \le (1-\alpha) g(s) + \alpha g(t)$ and hence

$$g(t) - g(s) \ge \frac{g(s+\alpha(t-s)) - g(s)}{\alpha}$$

Taking limit

$$\lim_{\alpha \downarrow 0} \frac{g(s+\alpha(t-s)) - g(s)}{\alpha(t-s)} (t-s) = g'(s)\,(t-s)$$

we have (b). Conversely suppose (b) and we want to prove (a), i.e.

$$\alpha g(t) + (1-\alpha)g(s) \; - \; g(z) \geq 0 \tag{8.4}$$

for any $z := s + \alpha(t-s)$, $\alpha \in [0,1]$. Compare the difference $g(t) - g(z)$ and $g(s) - g(z)$ in terms of gradient at the common point z:

$$g(t) - g(z) \; \geq \; g'(z)(t-z) \;\text{ and }\; g(s) - g(z) \; \geq \; g'(z)(s-z)$$

To obtain (8.4), multiply the first inequality by α and the second inequality by $1-\alpha$ and sum, noting that $t - z = (1-\alpha)(t-s)$ and $s - z = -\alpha(t-s)$ so that the right-hand sides of these two inequalities sum to zero. This proves (a) $\Leftrightarrow$ (b).

Now suppose (b). Fix any $t \geq s$ and compare $g(t) - g(s)$ in terms of slope at s and at t:

$$g'(s)(t-s) \leq \; g(t) - g(s) \; \leq g'(t)(t-s)$$

yielding (c). Conversely suppose (c) and fix any $t \geq s$. By the mean value theorem we have, for some $z \in [s,t]$, $g(t) - g(s) \; = \; g'(z)(t-s) \; \geq \; g'(s)(t-s)$, which is (b). This proves (b) $\Leftrightarrow$ (c).

Now consider $f : X \to \mathbb{R}$ where $X \subseteq \mathbb{R}^n$ with $n \geq 1$. We use the result above on scalar functions to prove the theorem. Suppose f is convex and fix any $x, y \in X$. Define the scalar function $g : \mathbb{R} \to \mathbb{R}$ by

$$g(s) := f(x + sy), \qquad \text{for } s \in \mathbb{R} \text{ such that } x + sy \in X \tag{8.5}$$

It is easy to show that $g(s)$ is convex. By the mean value theorem there exists an $s \in [0,1]$ such that

$$f(x+y) - f(x) = g(1) - g(0) \;\; = \;\; g'(s)$$

By (c) above we have $g'(s) \geq g'(0) = (\nabla f(x))^{\mathsf{T}} y$ and hence

$$f(x+y) - f(x) \geq (\nabla f(x))^{\mathsf{T}} y$$

establishing (8.3). Moreover if f is strictly convex then the inequalities above are strict.

Conversely suppose (8.3) holds. To prove the convexity of f, use the same proof above for (b) $\Rightarrow$ (a). Take $z := x + \alpha(y-x)$ for any $\alpha \in [0,1]$. We have

$$f(y) - f(z) \; \geq \; (\nabla f(z))^{\mathsf{T}}(y-z) \;\text{ and }\; f(x) - f(z) \; \geq \; (\nabla f(z))^{\mathsf{T}}(x-z)$$

Multiply the first inequality by α and the second inequality by $1-\alpha$ and sum to obtain:

$$\alpha f(y) + (1-\alpha) f(x) \; - \; f(z) \geq (\nabla f(z))^{\mathsf{T}}(\alpha(y-z) - (1-\alpha)(z-x)) \;\; = \;\; 0$$

proving the convexity of f. Moreover if the inequalities above are strict then f is strictly convex.

3 To prove the second-order condition, fix any $x, y \in X$, and define the scalar function $g(s) := f(x+s(y-x))$. Applying the second-order Taylor expansion to g:

$$\begin{aligned} f(y) - f(x) &= g(1) - g(0) = g'(0) + \frac{1}{2} g''(s) \\ &= (\nabla f(x))^{\mathsf{T}}(y-x) + \frac{1}{2}(y-x)^{\mathsf{T}} \nabla^2 f(x+s(y-x))(y-x) \end{aligned}$$

for some $s \in [0,1]$. If $\nabla^2 f(z) \succeq 0$ for all $z \in X$, then $f(y) - f(x) \geq (\nabla f(x))^{\mathsf{T}}(y-x)$ which is equivalent to the convexity of f from part 2.

Conversely, suppose f is convex but $\nabla^2 f(x) \prec 0$ for some $x \in X$. Then there exists a vector $v \in \mathbb{R}^n$ such that $v^{\mathsf{T}} \nabla^2 f(x) v < 0$. Since f is convex, part 1 shows that the scalar function $g(t) := f(x+tv)$ is convex in t. Then the proof of part 2(c) shows that, when g is twice differentiable, $g''(t) \geq 0$ for all $t \in \mathbb{R}$ such that $x+tv \in X$. But $g''(t) = v^{\mathsf{T}} \nabla^2 f(x+tv) v$ and hence $v^{\mathsf{T}} \nabla^2 f(x) v < 0$ means $g''(0) < 0$, contradicting that g is convex.

□

Theorem 8.2 provides an exact characterization for convexity. For *strict* convexity, the second-order characterization is sufficient but not necessary: e.g., $f(x) = x^4$ is strictly convex but $f''(x) = 0$ at $x = 0$. The following result can be proved following the argument for Theorem 8.2 (Exercise 8.7).

Corollary 8.3 (Strictly convex function)**.** Consider a function f defined on a convex open set $X \subseteq \mathbb{R}^n$.

1 The function f is strictly convex if and only if the function $g(t)$ in (8.2) is strictly convex on $\{t \in \mathbb{R} \mid x+tv \in X\}$.
2 For a differentiable function f, f is strictly convex if and only if strict inequality holds in (8.3) for $x \neq y$.
3 For a twice differentiable function f, f is strictly convex if $\nabla^2 f(x) \succ 0$ for all $x \in X$.

A common mistake is to confuse the second-order condition in Theorem 8.2.3 that $\nabla^2 f(x)$ is positive semidefinite with the condition that

$$x^{\mathsf{T}} \nabla^2 f(x) x \geq 0 \qquad \text{for all } x \in X$$

For any $x \in X$, $\nabla^2 f(x) \succeq 0$ if and only if

$$y^{\mathsf{T}} \nabla^2 f(x) y \geq 0 \qquad \text{for all } y \in \mathbb{R}^n$$

i.e., regardless of what X is, the test on $\nabla^2 f(x)$ is for *all* $y \in \mathbb{R}^n$. This is illustrated in the next example.

Example 8.5. Consider the function

$$f(x_1, x_2) = x_1 x_2$$

over the domain $X := \{(x_1, x_2) \in \mathbb{R}^2 : x_1 > 0, x_2 > 0\}$ with

$$\nabla^2 f(x) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

We have $x^\mathsf{T} \nabla^2 f(x) x = 2x_1 x_2 > 0$ for all $x \in X$, but $\nabla^2 f(x)$ is not positive semidefinite. Indeed its eigenvalues are 1 and -1 and hence f is convex along the eigenvector corresponding to eigenvalue 1, but concave along that corresponding to eigenvalue -1. Specifically the function value along the direction $x_1 = x_2$ corresponding to the eigenvalue-eigenvector pair $(1, [1\ 1]^\mathsf{T})$ is given by

$$g(t) := f\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + t \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix}\right) = (x_1 + t)(x_2 + t), \qquad t > -\min\{x_1, x_2\}$$

Hence $g(t)$ is convex in t, i.e. f is convex along $x_1 = x_2$. Along the direction $x_1 = -x_2$ corresponding to the eigenvalue-eigenvector pair $(-1, [1\ -1]^\mathsf{T})$ the function value is

$$g(t) := f\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + t \cdot \begin{bmatrix} 1 \\ -1 \end{bmatrix}\right) = (x_1 + t)(x_2 - t), \qquad -x_1 \le t \le x_2$$

Therefore $g(t)$ is concave in t, i.e., f is concave along $x_1 = -x_2$. □

Example 8.6. We illustrate Theorem 8.2 using $f(x) = \log x$ for $x > 0$.

1 We have $f'(x) = x^{-1}$ and for $x \neq y > 0$ (such that $\frac{y}{x} \neq 1$)

$$f(y) - f(x) = \log \frac{y}{x} < \frac{y}{x} - 1 = \frac{1}{x}(y - x) = f'(x)(y - x)$$

where the inequality follows from $\log z < z - 1$ for $z > 0$ and $z \neq 1$. Hence f is strictly concave by Theorem 8.2.2.

2 To use Theorem 8.2.3 we have

$$f''(x) = -\frac{1}{x^2} < 0$$

implying strict concavity of f.

□

The addition, multiplication by a positive constant, and supremum operations preserve convexity. Specifically suppose f_1 and f_2 are two convex functions on the same domain. Then (Exercise 8.8):

1. $f := \alpha f_1 + \beta f_2$, $\alpha, \beta \ge 0$, is convex.
2. $f := \max\{f_1, f_2\}$ is convex. In fact $f(x) := \sup_{y \in Y} f(x; y)$ is convex in x for arbitrary set Y, provided that, for every $y \in Y$ fixed, $f(x; y)$ is convex in x.
3. $f(x, y) := |x| + |y|$ defined on $\mathbb{R}^2$ is convex as it can be expressed in terms of the supremum and addition operations ($f(x, y) = \max\{x, -x\} + \max\{y, -y\}$).

4. $f(g(x))$ is convex if $g : \mathbb{R}^n \to \mathbb{R}^m$ is convex (componentwise) and $f : \mathbb{R}^m \to \mathbb{R}$ is convex and nondecreasing (componentwise), i.e., $f(y_1) \le f(y_2)$ for $y_1, y_2 \in \mathbb{R}^m$ with $y_1 \le y_2$.

Convex functions define another important class of convex sets. Let $f : X \to \mathbb{R}$ where $X \subseteq \mathbb{R}^n$. If X is a convex set and f a convex function then for each $a \in \mathbb{R}$ the *level set* $X_a := \{x \in X \mid f(x) \le a\}$ is convex. For a vector-valued function $f : X \to \mathbb{R}^m$ where $f := (f_1, \ldots, f_m)$ with $f_i : X \to \mathbb{R}$. Then the set specified by:

$$X_b := \{x \in X \mid f(x) \le b\}, \qquad b \in \mathbb{R}^m$$

is convex if f is convex, i.e, if each f_i is convex. This is because the level sets $X_{b_i} := \{x \in X \mid f_i(x) \le b_i\}$ are convex for all $i = 1, \ldots, m$, and hence their intersection $X_b = \cap_{i=1}^m X_{b_i}$ is convex. Note that the converse may not hold, i.e., a level set that is convex may be specified by nonconvex functions. For example the second-order cone K_{soc} may be specified as $K_{\text{soc}} = \{x \in \mathbb{R}^n : f(x) \le 0, x_n \ge 0\}$ where $f(x) := \|x^{n-1}\|_2^2 - x_n^2$ is nonconvex (see the discussion after (8.16)).

An important property of a real-valued convex function is that it is continuous on the interior of its domain, as the following lemma from [54, Proposition 1.3.11] shows. See Lemma 12.15 for generalization to proper extended real-valued convex functions. Lemma 12.15 also implies that a real-valued convex function over a compact set X is Lipschitz continuous on X.

Lemma 8.4 (Continuity of convex functions)**.** Let $f : X \to \mathbb{R}$ where $X \subseteq \mathbb{R}^n$. If f is convex on X then it is continuous on $\text{int}(X)$.

Proof Fix any point $\tilde{y} \in \text{int}(X)$ and consider any sequence $\{y_k\}$ such that $y_k \ne \tilde{y}$ and $\lim_k y_k = \tilde{y}$. We will show that

$$\limsup_k f(y_k) \;\le\; f(\tilde{y}) \;\le\; \liminf_k f(y_k) \tag{8.6}$$

implying $\lim_k f(y_k) = f(\tilde{y})$. Since $\tilde{y} \in \text{int}(X)$ there exists $\delta > 0$ such that the compact set $B_\delta(\tilde{y}) := \{x : \|x - \tilde{y}\|_\infty \le \delta\} \subseteq X$. We will consider sufficiently large integers k such that $y_k \in B_\delta(\tilde{y})$ for all (such) k.

Since any $x \in B_\delta(\tilde{y})$ is within distance 1 from $\tilde{y}$ along each coordinate, we can write x in terms of a convex combination of the unit bases e_j, i.e., $x = \tilde{y} + \sum_{j=1}^n \alpha_j e_j$ for some $\alpha_j \ge 0$ with $\sum_j \alpha_j = 1$. Then

$$f(x) = f\left(\sum_j \alpha_j(\tilde{y} + e_j)\right) \le \sum_i \alpha_i f(\tilde{y} + e_j) =: A \tag{8.7}$$

We now establish the first inequality in (8.6). For each k, since $y_k \in B_\delta(\tilde{y})$, we can find a $z_k \in B_\delta(\tilde{y})$ that is δ distance from $\tilde{y}$ such that y_k is a convex combination of $\tilde{y}$

and z_k (see Figure 8.4):

$$y_k \;=\; \tilde{y} + \frac{\|\Delta y_k\|}{\delta}\,(z_k - \tilde{y})$$

The convexity of f then implies

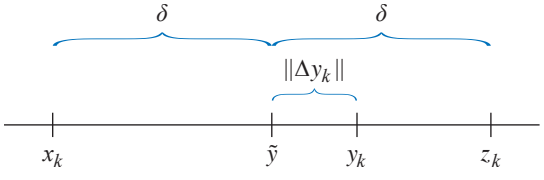


Figure 8.4 Proof of Lemma 8.4: Construction of (x_k, z_k) from $\tilde{y}$ and y_k.

$$f(y_k) \;\le\; \left(1 - \frac{\|\Delta y_k\|}{\delta}\right) f(\tilde{y}) + \frac{\|\Delta y_k\|}{\delta} f(z_k) \;\le\; \left(1 - \frac{\|\Delta y_k\|}{\delta}\right) f(\tilde{y}) + \frac{\|\Delta y_k\|}{\delta} A$$

where the last inequality follows from (8.7). Taking $\limsup_k$ therefore yields the first inequality in (8.6) since $\Delta y_k \to 0$.

The second inequality in (8.6) follows a similar argument. For each k, let $x_k \in B_\delta(\tilde{y})$ be the vector that is δ distance from $\tilde{y}$ such that $\tilde{y}$ is a convex combination of x_k and y_k (see Figure 8.4):

$$\tilde{y} \;=\; x_k + \frac{\delta}{\delta + \|\Delta y_k\|}\,(y_k - x_k)$$

Hence

$$f(\tilde{y}) \;\le\; \frac{\|\Delta y_k\|}{\delta + \|\Delta y_k\|} f(x_k) + \frac{\delta}{\delta + \|\Delta y_k\|} f(y_k) \;\le\; \frac{\|\Delta y_k\|}{\delta + \|\Delta y_k\|} A + \frac{\delta}{\delta + \|\Delta y_k\|} f(y_k)$$

Taking $\liminf_k$ therefore yields the second inequality in (8.6) since $\Delta y_k \to 0$. □

Strong convexity.

A function f is strictly convex if $\nabla^2 f(x) > 0$ for all $x \in X$ (Corollary 8.3). Its curvature however may be arbitrarily flat, i.e., $y^\mathsf{T} \nabla^2 f(x) y > 0$ can be arbitrarily close to zero. A stronger form of convexity bounds this away from zero uniformly in x, i.e., for some $\alpha > 0$, $\nabla^2 f(x) \succeq \alpha I$ for all $x \in \mathbb{R}^n$.

Definition 8.4 (Strong convexity)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable on $\mathbb{R}^n$. It is called *strongly convex* with parameter α on a set $X \subseteq \mathbb{R}^n$ if there exists $\alpha > 0$ such that

$$(\nabla f(y) - \nabla f(x))^\mathsf{T} (y - x) \ge \alpha \|y - x\|_2^2 \qquad \forall x, y \in X \subseteq \mathbb{R}^n \tag{8.8}$$

Definition 8.4 implies that $\|\nabla f(y) - \nabla f(x)\|_2 \ge \alpha \|y - x\|_2$ for all $x, y \in X$. Strong convexity is stronger than strict convexity.

Lemma 8.5 (Strong convexity)**.** Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable on $\mathbb{R}^n$. If f is strongly convex on $X \subseteq \mathbb{R}^n$ then it is strictly convex on X.

Proof As in the proof of Lemma 8.32, fix any $x, y \in X$ and consider the (scalar) function along the path from x to y in a straight line:

$$g(s) := f(x+sy) \qquad \text{for } s \in [0,1]$$

with $g'(s) = y^\mathsf{T} \nabla f(x+sy)$ as the directional derivative of f at $x+sy$ in the direction y. Then

$$\begin{aligned} f(x+y) - f(x) = \int_0^1 g'(s)ds &= \int_0^1 y^\mathsf{T} \nabla f(x+sy)ds \\ &= \int_0^1 \left(y^\mathsf{T} \nabla f(x) + y^\mathsf{T} (\nabla f(x+sy) - \nabla f(x)) \right) ds \\ &\geq y^\mathsf{T} \nabla f(x) + \int_0^1 \frac{1}{s} \alpha \|sy\|_2^2 \, ds \\ &= y^\mathsf{T} \nabla f(x) + \frac{\alpha}{2} \|y\|_2^2 \end{aligned} \tag{8.9}$$

where the inequality follows from (8.8). Corollary 8.3 then implies the strict convexity of f. □

Definition 8.4 does not require f to be twice continuously differentiable. For a twice continuously differentiable function f, if it is strongly convex and f is finite on X, then the Hessian $\nabla^2 f(x)$ is both lower and upper bounded uniformly on X, as explained in the next result.

Theorem 8.6 (Strong convexity)**.** Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be twice continuously differentiable on $\mathbb{R}^n$.

1 (8.8) is equivalent to $\nabla^2 f(x) \succeq \alpha \mathbb{I}$ for all $x \in X$ where $\mathbb{I}$ is the identity matrix of size n.
2 Suppose f is strongly convex and $\sup_{x \in X} f(x) < \infty$. Then
   - $\nabla^2 f(x) \preceq \beta \mathbb{I}$ for all $x \in X$ where β is a finite upper bound on the maximum eigenvalue $\lambda_{\max}(x)$ on X.
   - The gradient ∇f is Lipschitz continuous with Lipschitz constant β:

$$\|\nabla f(y) - \nabla f(x)\|_2 \leq \beta \|y - x\|_2, \qquad x, y \in X \tag{8.10}$$

Proof For part 1, suppose $\nabla^2 f(x) \succeq \alpha \mathbb{I}$ for all $x \in X$. We will show that f is strongly convex, i.e., f satisfies (8.8). Fix any $x, y \in X$ and let

$$h(s) := \nabla f(x + s(y-x))^\mathsf{T} (y-x)$$

Then

$$h'(s) = (y-x)^\mathsf{T} \nabla^2 f(x + s(y-x)) \, (y-x)$$

and

$$(\nabla f(y)-\nabla f(x))^{\mathsf{T}}(y-x) = h(1)-h(0) \;=\; \int_0^1 h'(s)ds$$
$$= \int_0^1 (y-x)^{\mathsf{T}}\nabla^2 f(x+s(y-x))\,(y-x)\,ds \;\geq\; \alpha\,\|y-x\|_2^2$$

where the inequality follows from $\nabla^2 f(x) \succeq \alpha\mathbb{I}$. Hence $f(x)$ is strongly convex. Conversely suppose f is strongly convex. To estimate $\nabla^2 f(x)$ we have for any $x \in X$, $y \in \mathbb{R}^n$,

$$y^{\mathsf{T}}\nabla^2 f(x)\,y = \lim_{\lambda\to 0}\frac{1}{\lambda}\left(\frac{\partial f}{\partial x}(x+\lambda y)-\frac{\partial f}{\partial x}(x)\right)y \;\geq\; \lim_{\lambda\to 0}\frac{1}{\lambda^2}\left(\alpha\,\|\lambda y\|_2^2\right) \;=\; \alpha\,\|y\|_2^2$$

where the inequality follows from the strong convexity of f. Hence $\nabla^2 f(x) \succeq \alpha I$ as desired. This shows the equivalence of (8.8) and $\nabla^2 f(x) \succeq \alpha\mathbb{I}$ for all $x \in X$.

For Part 2 we will show that if f is strongly convex, i.e., $\nabla^2 f(x) \succeq \alpha I$ on X, then it is also upper bounded, i.e., $\nabla^2 f(x) \preceq \beta\mathbb{I}$ for a finite β, provided $\sup_{x\in X} f(x) < \infty$. Since $\nabla^2 f(x)$ is symmetric and positive definite, its eigenvalues are positive for all $x \in X$ and

$$y^{\mathsf{T}}\nabla^2 f(x)y \;\leq\; \max_{y'\in\mathbb{R}^n}\frac{(y')^{\mathsf{T}}\nabla^2 f(x)y'}{\|y'\|_2^2}\|y\|_2^2 \;=\; \lambda_{\max}(x)\|y\|_2^2, \quad x \in X,\ y \in \mathbb{R}^n$$

where $\lambda_{\max}(x) > 0$ is a largest eigenvalue of $\nabla^2 f(x)$ and the last equality follows from the variational inequalities for eigenvalues of symmetric matrices (see (A.15) in Chapter A.6.2). This is equivalent to

$$\nabla^2 f(x) \;\leq\; \lambda_{\max}(x)\mathbb{I}, \qquad x \in X$$

It thus suffices to show that $\lambda_{\max}(x)$ is finite over X.

For all $x, y \in X$, we have

$$\begin{aligned} f(y) \;&=\; f(x)+\frac{\partial f}{\partial x}(x)(y-x)+\frac{1}{2}(y-x)^{\mathsf{T}}\nabla^2 f(z)(y-x) \\ &\geq\; f(x)+\frac{\partial f}{\partial x}(x)(y-x)+\frac{\alpha}{2}\|y-x\|_2^2 \end{aligned} \tag{8.11}$$

for some z between x and y. (That (8.9) and (8.11) are the same is why (8.8) is equivalent to $\nabla^2 f(x) \succeq \alpha I$.) If $f^{\max} := \sup_{x\in X} f(x) < \infty$, then fix an $x \in X$ and we have, for all $y \in X$,

$$f^{\max} \;\geq\; f(y) \;\geq\; f(x)+\frac{\partial f}{\partial x}(x)(y-x)+\frac{\alpha}{2}\|y-x\|_2^2$$

This implies that $y \in X$ must be bounded. Since this holds for all $y \in X$, X must be a bounded set and therefore its closure $\mathrm{cl}(X)$ is a compact set. Eigenvalues are continuous functions of their matrix entries and f is twice continuously differentiable, and hence $\beta := \max_{x\in\mathrm{cl}(X)}\lambda_{\max}(x)$ is finite (Theorem 8.16).

Finally for (8.10), we have (Lemma A.34 in Chapter A.10)

$$\|\nabla f(y) - \nabla f(x)\| \;\le\; \left\|\nabla^2 f(z)\right\| \|y - x\|, \forall x, y$$

for any vector norm and any induced matrix norm, and for some z between x and y. For the spectral norm (induced by the l_2 vector norm; see Theorem A.25 of Chapter A.8.3), $\nabla^2 f(x) \le \beta \mathbb{I}$ implies that $\|\nabla^2 f(x)\|_2 \le \beta$ because

$$\|\nabla^2 f(x)\|_2 \;=\; \max_{\|y\|_2=1} \|\nabla^2 f(x) y\|_2 \;=\; \max_{\|y\|_2=1} \sqrt{y^\mathsf{T} \left(\|\nabla^2 f(x)\right)^2 y} \;=\; \lambda_{\max}(x) \;\le\; \beta$$

This proves (8.10) and completes the proof of Theorem 8.6. □

Theorem 8.6 is critical in the convergence analysis of gradient algorithms. We explain its implications in the next remark.

Remark 8.1 (Strong convexity and convergence analysis)**.** 1 The condition $f^{\max} := \sup_{x \in X} f(x) < \infty$ is not restrictive even if $f(x) \to \infty$ as x recedes in X along a certain direction (e.g., $f(x) = x^2$). For instance if a feasible point $x_0 \in X$ is known then the feasible set X can be replaced by $X' := \{x \in X : f(x) \le f(x_0)\}$ without affecting minimization, and $f^{\max}$ in the proof can be replaced by $f(x_0)$.

2 Strong convexity in terms of the gradient $\nabla f(x)$ in Definition 8.4 is equivalent to $\nabla^2 f(x) \ge \alpha \mathbb{I}$ for all $x \in X$. The variational inequality for eigenvalues of symmetric matrices says that $\min_{x:\|x\|_2=1} x^\mathsf{T} \nabla^2 f(x) x = \lambda_{\min}(x) > 0$, a minimum eigenvalue of $\nabla^2 f(x)$ (see (A.15) in Chapter A.6.2). Hence the strong convexity parameter α can be any finite lower bound, e.g., $\alpha := \min_{x \in \mathrm{cl}(X)} \lambda_{\min}(x) > 0$.

3 Suppose $f^{\max} := \sup_{x \in X} f(x) < \infty$. Then Theorem 8.6 implies

$$\alpha \mathbb{I} \;\le\; \nabla^2 f(x) \;\le\; \beta \mathbb{I}, \qquad x \in X \subset \mathbb{R}^n \tag{8.12a}$$

where $\alpha > 0, \beta < \infty$ are lower and upper bounds on the minimum and maximum eigenvalues of $\nabla^2 f(x)$ on X respectively. This implies the following bounds on the gradient $\nabla f(x)$

$$\alpha \|y - x\|_2 \;\le\; \|\nabla f(y) - \nabla f(x)\|_2 \;\le\; \beta \|y - x\|_2, \qquad x, y \in X \tag{8.12b}$$

4 As explained in Theorem 8.36 of Chapter 8.6.3 on the convergence of the steepest descent algorithm, the upper bound in (8.12b) guarantees strict descent while the lower bound in Definition 8.4 guarantees geometric convergence. Hence the steepest descent algorithm converges to a unique optimal point geometrically (exponentially fast).

8.1.5 Convex program

Consider an optimization problem of the form:

$$\min_{x \in \mathbb{R}^n} \; f(x) \;\text{ subject to }\; x \in X \tag{8.13}$$

$X \subseteq \mathbb{R}^n$ is called the *feasible set* and $f : \mathbb{R}^n \to \mathbb{R}$ the *objective function*. An $x \in X$ is called a *feasible solution* of (8.13). A feasible solution x^* that attains the minimum of f over X (i.e., $f(x^*) \leq f(x)$ for all $x \in X$) is called a (global) *optimal solution/optimum* or a (global) *minimizer*. A feasible solution x^* that attains the minimum of f over a neighborhood of x^* (i.e., $f(x^*) \leq f(x)$ for all $x \in B_r(x^*) \cap X$ for some $r > 0$) is called a *local optimal solution/optimum* or a *local minimizer*.

The problem (8.13) is called a *convex program/problem* if f is a convex function and X is a convex set. It is tractable if X can be efficiently represented. For instance

$$X := \{x \in \mathbb{R}^n \mid g(x) \leq b\}$$

for a vector-valued convex function $g : \mathbb{R}^n \to \mathbb{R}^m$ and a $b \in \mathbb{R}^m$. By setting $u(x) = -f(x)$, the following maximization problem is also called a convex program if $u(x)$ is a concave function and X is a convex set:

$$\max_{x \in \mathbb{R}^n} \; u(x) \;\text{ subject to }\; x \in X$$

Importance of convexity.

As we will see in Chapter 8.3 the existence of optimal solutions and their characterization may not require the cost function f to be a convex function or the feasible set X to be a convex set. Convexity of f and X is important for efficient computation of an optimal solution. This is because for a convex objective function, local optimality implies global optimality. Moreover only the first-order condition is required to guarantee local optimality. Specifically, for an unconstrained minimization problem

$$\min_{x \in \mathbb{R}^n} \; f(x)$$

a necessary condition for a point x^* to be a *local* minimizer is (assuming f is differentiable)

$$\nabla f(x^*) = 0$$

If f is convex then this is also sufficient for x^* to be *globally* optimal, as illustrated in Figure 8.2. For constrained minimization problem (8.13) where X is nonempty, closed and convex, the first-order necessary condition for $x^* \in X$ to be a local minimizer becomes: there is a neighborhood $B_r(x^*)$ for some $r > 0$ such that

$$(\nabla f(x^*))^\mathsf{T} (x - x^*) \geq 0 \qquad \forall x \in B_r(x^*) \cap X \tag{8.14}$$

i.e., moving away from x^* to any other feasible point x in $B_r(x^*)$ can only locally increase the function value f (see Figure 8.16). If f is convex then this is both necessary and sufficient for x^* to be *globally* optimal. To see this, suppose (8.14) holds but there is another $\hat{x} \in X$ such that $f(\hat{x}) < f(x^*)$. Consider $z(\alpha) := \alpha\hat{x} + (1-\alpha)x^*$. Since X is convex $z(\alpha)$ is feasible for $\alpha \in [0,1]$. Since f is convex we have, for *any* $\alpha \in (0,1]$,

$$f(z(\alpha)) \leq \alpha f(\hat{x}) + (1-\alpha) f(x^*) \;<\; f(x^*)$$

But, for small enough $\alpha > 0$ so that $z(\alpha) \in B_r(x^*)$, this contradicts

$$f(z(\alpha)) \geq f(x^*) + \nabla^{\mathsf{T}} f(x^*)(z(\alpha) - x^*) \quad \geq \quad f(x^*)$$

where the first inequality follows from Theorem 8.2.2 and the second inequality from (8.14). Hence x^* is globally optimal in X.

Example 8.7 (Optimality condition for constrained optimization)**.** Consider

$$\min_{x \in \mathbb{R}} \; f(x) := x^2 \quad \text{subject to} \quad x \geq a$$

See Figure 8.5. It is clear from the figure that the unique minimizer is 0 where

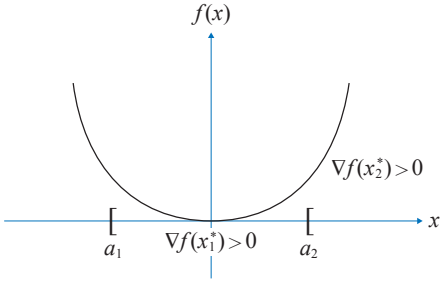


Figure 8.5 Example 8.7: $\min_{x \geq a} x^2$. If $a \leq 0$ then the unique minimizer is $x_1^* = 0$ where $f'(x^*) = 0$. If $a > 0$ then the unique minimizer is $x_2^* = a$ where $f'(x^*) > 0$.

$f'(0) = 0$ if $a \leq 0$ and a where $f'(a) > 0$ if $a > 0$. We will derive this conclusion from the optimality condition (8.14) which is

$$f'(x^*)(x - x^*) \geq 0, \qquad \forall x \geq a \tag{8.15}$$

First suppose $a \leq 0$. If $a \leq x^* < 0$ then $f'(x^*) < 0$ and there exists a feasible $x > x^*$ where (8.15) cannot be satisfied. Similarly if $x^* > 0 \geq a$ then $f'(x^*) > 0$ and there exists a feasible $a \leq x < x^*$ where (8.15) cannot be satisfied. Hence the unique optimal is $x^* = 0$ where $f'(x^*) = 0$. Suppose next $a > 0$. Then $f'(x) > 0$ for any feasible $x \geq a$. Then the only way (8.15) can be satisfied is if $x^* = a$.

Therefore the optimality condition reduces for this example (for any $a \in \mathbb{R}$) to: x^* is optimal if and only if there exists a p^* such that

$$x^* \geq a, \; p^* \geq 0, \; f'(x^*) = p^*, \; p^*(x^* - a) = 0$$

This is an example of the Karush-Kuhn-Tucker (KKT) condition (see Chapter 8.3.2).

□

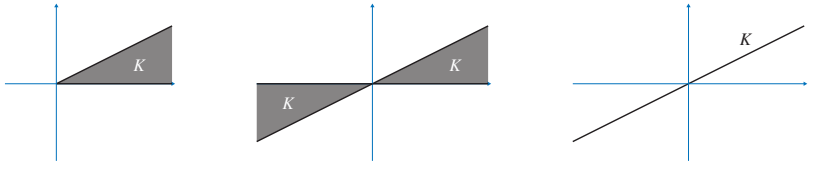
8.2 Properties of convex sets and convex cones

In this section we study some of the most useful properties of convex sets and cones. For example the Projection Theorem 8.9 is used to prove the separating hyperplane Theorems 8.10 and 8.11 which are used to prove the Farkas Lemma (Theorem 8.12). We will also use the Projection Theorem 8.9 to prove in Chapter 8.6 some convergence properties of optimization algorithms, use the Farkas Lemma (Theorem 8.12) to prove in Chapter 8.4.2 linear program duality, and use the separating hyperplane theorems to prove convex duality in Chapters 12.7.2 and 12.7.3.

8.2.1 Second-order cone K_{soc} in $\mathbb{R}^n$

Cones in $\mathbb{R}^n$.

A set $K \subseteq \mathbb{R}^n$ is called a *cone* if $x \in K$ implies that $\gamma x \in K$ for all $\gamma > 0$. A cone K may not contain the origin though the closure of a nonempty cone always contains the origin. A cone is not necessarily convex. For example $K := \{\gamma_1 a_1 : \gamma_1 \geq 0\} \cup \{\gamma_2 a_2 : \gamma_2 \geq 0\}$ for some $a_1, a_2 \in \mathbb{R}^n$ is a cone consisting of two rays from the origin and is nonconvex unless $a_1 = \gamma a_2$ for some $\gamma \in \mathbb{R}$. A cone K is called *pointed* if $x \in K$ and $-x \in K$ implies that $x = 0$. Figure 8.6 shows pointed and non-pointed cones that may be convex or not, a subspace or not. A cone K is called *proper* if (i) K is closed and convex; (ii)



(a) Pointed convex cone

(b) Non-pointed nonconvex cone

(c) Non-pointed convex cone

Figure 8.6 Cones and their affine hulls. (a) A pointed convex cone K. It is not a subspace; its affine hull $\text{aff}(K) = \mathbb{R}^2$. (b) A non-pointed nonconvex cone K. It is not a subspace; its affine hull $\text{aff}(K) = \mathbb{R}^2$. (c) A non-pointed convex cone K which is a subspace. Hence $\text{aff}(K) = K$.

has a nonempty interior; and (iii) is pointed.[2] Common examples are the nonnegative quadrant $\mathbb{R}^n_+ := \{x \in \mathbb{R}^n : x \geq 0\}$, the second-order cone $K_{\text{soc}} := \{x \in \mathbb{R}^n : \|x^{n-1}\|_2 \leq x_n\}$,

[2] A proper cone K can be used to define a partial ordering on $\mathbb{R}^n$ through a *generalized inequality* $\preceq_K$:

$$x \preceq_K y \Leftrightarrow y - x \in K$$

It also defines a strict partial ordering on $\mathbb{R}^n$:

$$x \prec_K y \Leftrightarrow y - x \in \text{int}\, K$$

where $\text{int}(K)$ is the interior of K. We also write $x \succeq_K y$ for $y \preceq_K x$ and $x \succ_K y$ for $y \prec_K x$. We will usually write directly $y - x \in K$ instead of $x \preceq_K y$.

and the set $K_{\text{psd}} \subset \mathbb{S}^n$ of positive semidefinite matrices in the linear space $\mathbb{S}^n$ of Hermitian matrices.

Definition 8.5 (cone(X))**.** Let $X \subseteq \mathbb{R}^n$ be a nonempty set. The *cone generated by* X, denoted cone(X), is the set of all nonnegative combination of vectors in X, i.e.,

$$\text{cone}(X) := \left\{ \sum_{i=1}^{m} \alpha_i x_i : x_i \in X, \alpha_i \geq 0, \text{ integers } m > 0 \right\}$$

If $\{a_1, \ldots, a_n\}$ are the column vectors of $A \in \mathbb{R}^{m \times n}$ then cone($\{a_1, \ldots, a_n\}$) $\subseteq \mathbb{R}^m$ is abbreviated as cone(A). □

The set cone(X) is always a convex cone that contains the origin for arbitrary nonempty X. See Figure 8.7 for examples. It therefore contains the set $\{\gamma x : \gamma \geq 0, x \in$

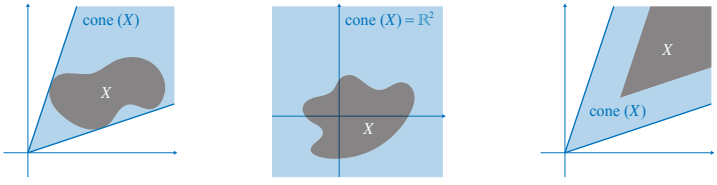


Figure 8.7 Cones cone(X) generated by $X \subseteq \mathbb{R}^n$.

$X\}$ which may not be convex, e.g., $X := \{a_1, a_2\}$ with $a_1 \neq \gamma a_2$. It is not necessarily closed even if X is compact (see [54, Figure 1.2.2, p.21] for an example). We will mostly be dealing with closed convex cones in this book.

Recall from Definition 8.2 that conv(X) of an arbitrary set X is the intersection of all convex sets containing X. A *convex combination* of $x_1, \ldots, x_m$ in X is the vector $x := \sum_{i=1}^m \alpha_i x_i$ with $\alpha_i \geq 0$ and $\sum_{i=1}^m \alpha_i = 1$. Any convex combination of vectors in X is in conv(X). The next fundamental result implies the converse, e.g., [54, Proposition 1.2.1, p.20].

Theorem 8.7 (Carathéodory Theorem)**.** Let $X \subseteq \mathbb{R}^n$ be a nonempty set.

1 If $x \in \text{conv}(X)$ is nonzero, then $x = \sum_{i=1}^m \alpha_i x_i$ for some $m \leq n+1$, $\alpha_i > 0$ with $\sum_{i=1}^m \alpha_i = 1$, and $x_i \in X$.
2 If $x \in \text{cone}(X)$ is nonzero, then $x = \sum_{i=1}^m \alpha_i x_i$ for some $m \leq n$, $\alpha_i > 0$ and linearly independent $x_i \in X$. □

The convex hull conv(X) of an arbitrary set X is not necessarily closed, e.g., $X = (0,1) = \text{conv}(X)$. A consequence of the Carathéodory theorem is that conv(X) is compact if X is compact. Suppose $x \in \text{conv}(X)$ is given by $x = \sum_{i=1}^m \beta_i y_i$ for some $m > n$, $\beta_i > 0$ with $\sum_{i=1}^m \beta_i = 1$, and $y_i \in X$. At most n of $y_i \in X$ can be linearly independent, say, $y_1, \ldots, y_k$ are linearly independent with $k \leq n$. Therefore other y_i for $i > k$ can be written as linear combinations of $y_1, \ldots, y_k$, and we can write $x = \sum_{i=1}^k \lambda_i y_i$

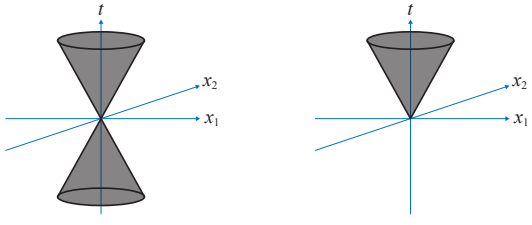
with $k \le n$. The coefficients λ_i, however, may not form a *convex* combination of y_i, unlike in the Carathéodory theorem. In other words, any $x \in \text{conv}(X)$ can be written as a linear combination of $k \le n$ vectors $y_i \in X$ (these y_i depend on x) and as a convex combination of $m \le n+1$ vectors $x_i \in X$ (these x_i depend on x). An example application of the Carathéodory theorem is in Exercises 12.9, 13.12 and 13.13.

Second-order cone.

A particularly useful convex cone is the *second-order cone*, defined by

$$K_{\text{soc}} := \left\{ (x,t) \in \mathbb{R}^n \times \mathbb{R} : \sqrt{x_1^2 + \cdots + x_n^2} \le t \right\} \tag{8.16}$$

It is also called the Lorentz cone or ice-cream cone. It has several equivalent specifications. It is equivalent to $K_{\text{soc}} = \{(x,t) : \|x\|_2^2 \le t^2,\ t \ge 0\}$ or the intersection $K_{\text{soc}} = \tilde{K} \cap H$ where $\tilde{K} := \{(x,t) : \|x\|_2^2 \le t^2\}$ and $H := \{(x,t) : t \ge 0\}$ is a halfspace. While K_{soc} is a convex cone, $\tilde{K}$ is a nonconvex cone; see Figure 8.8 and Exercise 8.11 (see Theorem 12.10 in Chapter 12.1.4 for more properties of K_{soc}). The second-order cone K_{soc} can



(a) Nonconvex cone $\tilde{X}$ (b) Convex cone K

Figure 8.8 (a) Nonconvex cone $\tilde{K} := \{(x,t) \in \mathbb{R}^{n+1} : \|x\|_2^2 \le t^2\}$. (b) Second-order cone $K_{\text{soc}} = \tilde{K} \cap H$.

also be specified as a level set of a constraint function, $K_{\text{soc}} := \{(x,t) : h_1(x,t) \le 0\}$ where $h_1(x,t) := \|x\|_2 - t$ is convex. Equivalently $K_{\text{soc}} := \{(x,t) : h_2(x,t) \le 0, t \ge 0\}$ where $h_2(x,t) := \|x\|_2^2 - t^2$ is nonconvex (Exercise 8.11). Hence a convex set can be specified by constraint functions that may not all be convex functions. This has important implications on structural and computational properties of equivalent representations of a constrained optimization; see Chapter 8.3.7.

A *rotated second-order cone* is the set

$$K_{\text{rsoc}} := \{(x,y,z) \in \mathbb{R}^n \times \mathbb{R}^2 : \|x\|_2^2 \le yz,\ y \ge 0,\ z \ge 0\} \tag{8.17}$$

It can be represented as a linear transformation (a rotation) of the standard second-order cone K_{soc} defined in (8.16) using the equivalence:

$$\|x\|_2^2 \le yz,\ y \ge 0,\ z \ge 0 \iff \left\| \begin{bmatrix} 2x \\ y - z \end{bmatrix} \right\|_2 \le y + z$$

i.e., $(w,t) = A(x,y,z) \in K_{\text{soc}} \subseteq \mathbb{R}^{n+2}$ if and only if $(x,y,z) \in K_{\text{rsoc}}$ for a $(n+2)\times(n+2)$ nonsingular matrix A. Indeed (Exercise 8.12)

$$K_{\text{soc}} = AK_{\text{rsoc}}, \qquad A = \begin{bmatrix} 2\mathbb{I}_n & 0_n & 0_n \\ 0_n^T & 1 & -1 \\ 0_n^T & 1 & 1 \end{bmatrix} \tag{8.18a}$$

$$K_{\text{rsoc}} = A^{-1}K_{\text{soc}}, \qquad A^{-1} = \frac{1}{2}\begin{bmatrix} \mathbb{I}_n & 0_n & 0_n \\ 0_n^T & 1 & 1 \\ 0_n^T & -1 & 1 \end{bmatrix} \tag{8.18b}$$

See Corollary 12.11 in Chapter 12.1.4 for more properties of K_{rsoc}.

SOC constraint.

A convex set specified in terms of a second-order cone $K_{\text{soc}} \subseteq \mathbb{R}^{n+1}$ in (8.16) is

$$C := \{x \in \mathbb{R}^n : (Ax+b, c^Tx+d) \in K_{\text{soc}}\} = \{x \in \mathbb{R}^n : \|Ax+b\|_2 \le c^Tx+d\} \tag{8.19}$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$, and $d \in \mathbb{R}$. It is a convex set because C is the pre-image of a convex set K_{soc} under an affine function (see also Exercise 8.13). The constraint in (8.19) is called a *second-order cone (SOC) constraint*, even though C in general may not be a cone itself. For example

- If $A = 0$ then C is a halfspace, generally not a cone.
- If $c = 0$ then C is an ellipsoid ($d > 0$), generally not a cone.

The set defined in (8.16) is a special case of (8.19) with $b = 0, d = 0$, $c = e_n$ the unit vector with a single 1 as its nth entry, and $A = \begin{bmatrix} \mathbb{I}_{n-1} & 0_{n-1} \end{bmatrix}$ where $\mathbb{I}_{n-1}$ and 0_{n-1} are the identity matrix and 0 vector respectively of size $n-1$.

Example 8.8 (SOC constraint)**.** Consider C defined in (8.19) where

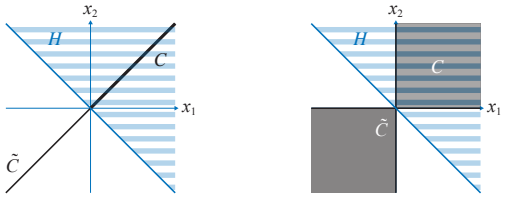
$$A := \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad c := \alpha\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad b := 0, \qquad d := 0$$

$C = \tilde{C} \cap H$ where $\tilde{C} := \{x : \|Ax+b\|_2^2 \le (c^Tx+d)^2\}$ and $H := \{x : c^Tx+d \ge 0\}$ is a halfspace. Then $\tilde{C} = \left\{x \in \mathbb{R}^2 : x^T\tilde{A}x \le 0\right\}$ where

$$\tilde{A} := A^TA - cc^T = \begin{bmatrix} 1-\alpha & -\alpha \\ -\alpha & 1-\alpha \end{bmatrix}$$

whose eigenvalues are 1 and $1-2\alpha$. Therefore if $\alpha \le 1/2$ then $\tilde{A}$ is positive semidefinite and $\tilde{C}$ is convex. Otherwise $\tilde{C}$ is nonconvex. In both cases $C = \tilde{C} \cap H$ is convex.

For example when $\alpha = 1/2$, $\tilde{C} = \{x : \frac{1}{2}(x_1-x_2)^2 \le 0\} = \{x : x_1 = x_2\}$. When $\alpha = 1$, $\tilde{C} = \{x : x_1x_2 \ge 0\} = \{x : x \ge 0\} \cup \{x : x \le 0\}$. These sets and their intersections with the halfspace $H := \{x : x_1+x_2 \ge 0\}$ are shown in Figure 8.9. □



(a) Convex $\tilde{C}$ (b) Nonconvex $\tilde{C}$

Figure 8.9 Exampel 8.8. (a) When $\alpha = 1/2$, $\tilde{C} = \{x : x_1 = x_2\}$ is convex. (b) When $\alpha = 1$, $\tilde{C} = \{x : x \geq 0\} \cup \{x : x \leq 0\}$ is nonconvex. In both cases $C = \tilde{C} \cap H$ is convex.

Similarly a convex set can be specified in terms of a rotated second-order cone $K_{\text{rsoc}} \subseteq \mathbb{R}^{m+2}$ in (8.17):

$$\begin{aligned} C_r &:= \{x \in \mathbb{R}^n \ : \ (Ax+b, c_1^\mathsf{T}x+d_1, c_2^\mathsf{T}x+d_2) \in K_{\text{rsoc}}\} \\ &= \{x \in \mathbb{R}^n \ : \ \|Ax+b\|_2^2 \leq (c_1^\mathsf{T}x+d_1)(c_2^\mathsf{T}x+d_2),\ c_1^\mathsf{T}x+d_1 \geq 0,\ c_2^\mathsf{T}x+d_2 \geq 0\} \end{aligned}$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$, and $d \in \mathbb{R}$. It is a convex set because C_r is the pre-image of a convex set K_{rsoc} under an affine function. The constraints that define C_r are also called *second-order cone constraints*, even though C_r in general may not be a cone itself. This form of constraint is used in Chapter 11 to relax the nonconvex quadratic constraint $v_j \ell_{jk} = |S_{jk}|^2$ into a second-order cone constraint $v_j \ell_{jk} \geq |S_{jk}|^2$.

We study properties of SOC constraints in Chapter 12.1.4.

8.2.2 Semidefinite cone K_{psd} in $\mathbb{S}^n$

Numerous power system applications can be formulated as a constrained optimization problem often using complex variables in the phasor domain. Moreover some solution methods for solving these problems give rise to constraints or variables involving matrices (see e.g. Chapter 10). Even though any optimization problem in the complex domain can be converted into one in the real domain, it is sometimes more convenient to use complex variables. In this subsection we define inner product on complex matrices and dual cones in the linear space of Hermitian matrices (all these concepts apply directly to the vector space of real symmetric matrices). We will use these concepts in Chapter 8.4.5 to define an important class of convex optimization problems called semidefinite program and study its duality and optimality properties.

Inner product, polar cone and dual cone.

For two complex matrices $x \in \mathbb{C}^{m\times n}$ and $y \in \mathbb{C}^{m\times n}$ (not necessarily square), the (Frobenius) inner product is $x \cdot y := \text{tr}\left(y^\mathsf{H}x\right) = \sum_{j,k} x_{jk}\bar{y}_{jk}$ where $y^\mathsf{H} = (\bar{y})^\mathsf{T}$ is the Hermitian transpose of matrix y, $\bar{y}_{jk}$ is the complex conjugate of the scalar y_{jk} and $\bar{y}$ is the entry-wise complex conjugate of matrix y. If $x, y \in \mathbb{C}^n$ are complex vectors, then $x \cdot y = y^\mathsf{H}x$

reduces to the normal inner product on $\mathbb{C}^n$. It can be checked that $x \cdot y$ satisfies the three properties that are sometimes used to define inner product:

1 *Conjugate symmetry*: $x \cdot y = \overline{y \cdot x}$.
2 *Linearity in the first argument*: For any $a, b \in \mathbb{C}$ and any fixed $y \in \mathbb{C}^{m \times n}$, $(a_1 x_1 + a_2 x_2) \cdot y = a_1 (x_1 \cdot y) + a_2 (x_2 \cdot y)$.
3 *Positive-definiteness*: $x \cdot x \geq 0$ with equality if and only if $x = 0$.

Let $x \in \mathbb{C}^{n \times n}$ be a square matrix. It is called a *Hermitian matrix* if $x_{jk} = \bar{x}_{kj}$ for all j, k. If x is Hermitian its diagonal entries x_{jj} are necessarily real. Let $\mathbb{S}^n \subset C^{n \times n}$ denote the set of all $n \times n$ Hermitian matrices. If $x, y \in \mathbb{S}^n$ then

$$x \cdot y \;=\; \sum_j x_{jj} \bar{y}_{jj} + \sum_{j<k} \left(x_{jk} \bar{y}_{jk} + x_{kj} \bar{y}_{kj} \right) \;=\; \sum_j x_{jj} y_{jj} + \sum_{j<k} \left(x_{jk} \bar{y}_{jk} + \bar{x}_{jk} y_{jk} \right)$$

i.e., $x \cdot y$ is a real number. This means that if $x, y \in \mathbb{S}^n$ are Hermitian matrices then

$$x \cdot y \;=\; y \cdot x \;\in\; \mathbb{R} \tag{8.20}$$

This implies that, for Hermitian matrices, the order of inner product in Definition 8.6 does not matter. We will consider $\mathbb{S}^n$ as a vector (or linear) space over the field $\mathbb{R}$ of real numbers, not over $\mathbb{C}$ (see Appendix A.1.1 for definitions of vector space and subspace). We can then call a set $K \subseteq \mathbb{S}^n$ of Hermitian matrices a *cone* in the vector space $\mathbb{S}^n$ if $x \in K$ implies that $\gamma x \in K$ for any $\gamma > 0$ in the field $\mathbb{R}$. As for a cone K of vectors in $\mathbb{R}^n$, a cone in $\mathbb{S}^n$ is not necessarily convex, e.g., $K := \{\gamma_1 x_1 : \gamma_1 \geq 0\} \cup \{\gamma_2 x_2 : \gamma_2 \geq 0\}$ is a nonconvex set unless $x_1 = \gamma x_2$ for some $\gamma \in \mathbb{R}$. We define the notion of dual cone in $\mathbb{S}^n$

Definition 8.6 (Cones in $\mathbb{S}^n$)**.** Consider the vector space $\mathbb{S}^n \subset \mathbb{C}^{n \times n}$ of Hermitian matrices. Let $X \subseteq \mathbb{S}^n$ be a nonempty set.

1 The *polar cone of* X is $X^\circ := \{y \in \mathbb{S}^n : y \cdot x \leq 0 \;\; \forall x \in X\}$
2 The *dual cone of* X is $X^* := -X^\circ = \{y \in \mathbb{S}^n \;:\; y \cdot x \geq 0 \;\; \forall x \in X\}$.
3 A cone K is called *self-dual* if $K^* = K$. □

The nonnegativity cone $\mathbb{R}^n_+ \subset \mathbb{R}^n$, the second-order cone $K_{\text{soc}} \subset \mathbb{R}^n$, and the positive semidefinite cone $K_{\text{psd}} \subset \mathbb{S}^n$ of positive semidefinite matrices are all self-dual proper cones (recall a proper cone is closed, convex, pointed and has nonempty interior).

Polar and dual cones in $\mathbb{R}^n$ are defined in exactly the same way in Chapter 12.1.1. Their properties are given in Proposition 12.1 and extend directly to cones in the vector space $\mathbb{S}^n$. For example for an arbitrary nonempty set $X \subseteq \mathbb{S}^n$ of matrices, its polar cone X° and dual cone X^* are closed convex cones. If X is itself a closed convex cone then $(X^\circ)^\circ = X$. The following property of the dual cone underlies the definition of dual problem and duality. Consider a cone K in an underlying vector space K^+, e.g., $K^+ := \mathbb{R}^n$ or $K^+ := \mathbb{S}^n$. Then the minimum value over K of the inner product

with another vector y is 0 if $y \in K^*$ and $-\infty$ if otherwise. It follows directly from the definition of dual cone and therefore applies to cones in both vector spaces $\mathbb{R}^n$ and $\mathbb{S}^n$.

Lemma 8.8 (Duality over cone)**.** Let K^+ be a vector space with an inner product $x \cdot y = y \cdot x$ which is in $\mathbb{R}$. Let $K \subseteq K^+$ be a nonempty cone. Then

$$\min_{x \in K} y \cdot x = \min_{x \in K} x \cdot y \quad = \begin{cases} 0 & \text{if } y \in K^* \\ -\infty & \text{if } y \in K^+ \setminus K^* \end{cases}$$

□

Lemma 8.8 holds whether or not the cone K is self dual or not; if $K^* = K$ then we can replace K^* by K in the lemma. The minimization over $x \in K$ arises, e.g., in partial dualization of a constrained optimization (see Remark 8.4).

Remark 8.2 (Semidefinite cones in $\mathbb{S}^n$)**.** The vector space $\mathbb{S}^n$ can be partitioned into the cone K_{psd} of positive semidefinite matrices, the cone K_{nsd} of negative semidefinite with $K_{\text{psd}} \cap K_{\text{nsd}} = \{0\}$ the zero matrix 0, and the set of indefinite Hermitian matrices (those with both positive and negative eigenvalues). Both K_{psd} and K_{nsd} are self-dual proper cones. They are also polar cones of each other, i.e., $K_{\text{psd}} = K_{\text{nsd}}^\circ$ and $K_{\text{nsd}} = K_{\text{psd}}^\circ$. □

8.2.3 Projection theorem

Given a set $X \subseteq \mathbb{R}^n$ the *projection of* $x \in \mathbb{R}^n$ *onto* X is defined to be:

$$[x]_X := \arg\min_{y \in X} \|x - y\|_2 \tag{8.21}$$

where $\|\cdot\|_2$ is the *Euclidean* norm. Hence $[x]_X$ is the unique point in X that is closest to $x \in \mathbb{R}^n$ in the Euclidean norm. They are illustrated in Figure 8.10.

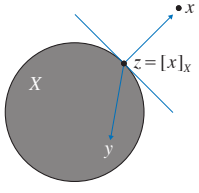


Figure 8.10 The point $z := [x]_X$ is the unique closest point to x in the convex set X under the Euclidean norm. For all other points $y \in X$, the inner product of $y - z$ and $x - z$ is nonpositive. (April 19, 2025: Arrow $z \to y$ is too obscure.)

Theorem 8.9 (Projection theorem)**.** Suppose $X \subseteq \mathbb{R}^n$ is a nonempty, closed and convex set.

1 For every $x \in \mathbb{R}^n$ there exists a unique $[x]_X$ defined by (8.21).

2 For every $x \in \mathbb{R}^n$, $z = [x]_X$ if and only if $z \in X$ and $(y-z)^\mathsf{T}(x-z) \leq 0$ for all $y \in X$.
3 The projection mapping $T : \mathbb{R}^n \to \mathbb{X}$ defined by $T(x) := [x]_X$ is continuous and *nonexpansive* under the Euclidean norm, i.e.,

$$\|[y]_X - [x]_X\|_2 \leq \|y - x\|_2 \qquad \forall x, y \in \mathbb{R}^n$$

Note that Theorem 8.9 does not require X to be bounded (compact), only closed. This is because since X is nonempty there is an $w \in X$. Hence the minimization in the projection (8.21) can be equivalently restricted to the compact set $\{y \in X \mid \|x-y\|_2 \leq \|x-w\|_2\}$.

8.2.4 Separating hyperplanes

Recall that for any set $X \subseteq \mathbb{R}^n$, $\mathrm{cl}(X)$ denotes the closure of X, $\mathrm{int}(X)$ denotes the interior of X, $\mathrm{ri}(X)$ denotes the relative interior of X, and $\mathrm{cl}(X) \setminus \mathrm{int}(X)$ is the boundary of $\mathrm{cl}(X)$.

Definition 8.7 (Separating hyperplane)**.** 1 A *hyperplane* is a set $H := \{x \in \mathbb{R}^n : a^\mathsf{T}x = b\}$ for some $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$.
2 Two sets $X, Y \subseteq \mathbb{R}^n$ are *separated by a hyperplane* $H = \{x \in \mathbb{R}^n : a^\mathsf{T}x = b\}$ if each lies in a different closed halfspace associated with H, i.e., ether

$$a^\mathsf{T}x \leq b \leq a^\mathsf{T}y, \quad x \in X,\ y \in Y \qquad \text{or} \qquad a^\mathsf{T}x \geq b \geq a^\mathsf{T}y, \quad x \in X,\ y \in Y$$

H is called a *separating hyperplane*.
3 If x^* is in the boundary $\mathrm{cl}(X) \setminus \mathrm{int}(X)$ of $X \subseteq \mathbb{R}^n$, the hyperplane $H := \{x \in \mathbb{R}^n : a^\mathsf{T}x = a^\mathsf{T}x^*\}$ that separates $\mathrm{cl}(X)$ (or X) and $\{x^*\}$ is called a *supporting hyperplane of cl(X) (or X) at* x^*. □

If point x^* is not in the interior of a set X then either x is on the boundary of X or x^* is not in the closure of X. The next result says that such a point x^* can always be separated from X by a hyperplane if X is convex. The hyperplane is a supporting hyperplane of X at x^* if and only if x^* is on the boundary of X. It is a straightforward consequence of the Projection Theorem 8.9.

Theorem 8.10 (A point x^* and a convex set X)**.** Suppose $X \subseteq \mathbb{R}^n$ is nonempty convex and $x^* \in \mathbb{R}^n \setminus \mathrm{int}(X)$.

1 There exists a hyperplane that passes through x^* that contains X in one of its halfspaces, i.e., there exists a nonzero $a \in \mathbb{R}^n$ such that

$$a^\mathsf{T}x \leq a^\mathsf{T}x^*, \qquad x \in \mathrm{cl}(X) \tag{8.22a}$$

A separating hyperplane is $H := \{x \in \mathbb{R}^n : a^\mathsf{T}x = a^\mathsf{T}x^*\}$.

2 If $x^* \notin \mathrm{cl}(X)$ then the inequality in (8.22a) is strict. Hence there exists $b \in (a^\mathsf{T}\hat{x}^*, a^\mathsf{T}x^*)$ such that the hyperplane $H := \{x \in \mathbb{R}^n : a^\mathsf{T}x = b\}$ strictly separates $\mathrm{cl}(X)$ and x^*, i.e.,

$$a^\mathsf{T}x \;<\; b \;<\; a^\mathsf{T}x^*, \qquad\qquad x \in \mathrm{cl}(X) \tag{8.22b}$$

where $\hat{x}^*$ is the projection of x^* onto the convex set $\mathrm{cl}(X)$.

Proof We prove part 2 first and then part 1.

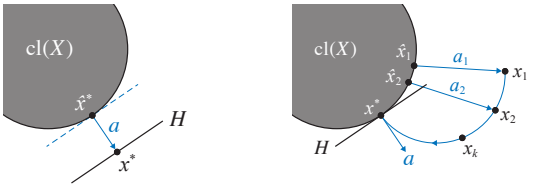
Part 2: $x^* \notin cl(X)$. Let $\hat{x}^* \neq x^*$ be the projection of x^* onto $\mathrm{cl}(X)$, i.e., $\hat{x}^* := \arg\min_{x \in \mathrm{cl}(X)} \|x - x^*\|_2$. Then $(x^* - \hat{x}^*)^\mathsf{T}(x - \hat{x}^*) \leq 0$ for all $x \in \mathrm{cl}(X)$ by the Projection Theorem 8.9. Define the normalized (error) vector

$$a \;:=\; \frac{x^* - \hat{x}^*}{\|x^* - \hat{x}^*\|_2} \;\neq\; 0 \tag{8.23a}$$

Therefore

$$a^\mathsf{T}x \;\leq\; a^\mathsf{T}\hat{x}^* \;=\; a^\mathsf{T}x^* \;-\; a^\mathsf{T}(x^* - \hat{x}^*) \;<\; a^\mathsf{T}x^*, \qquad\qquad x \in \mathrm{cl}(X) \tag{8.23b}$$

where the last inequality follows because $a^\mathsf{T}(x^* - \hat{x}^*) = \|x^* - \hat{x}^*\|_2 > 0$. By definition, (8.23) says that $\mathrm{cl}(X)$ is in a halfspace associated with the hyperplane $H := \{x \in \mathbb{R}^n : a^\mathsf{T}x = a^\mathsf{T}x^*\}$, as shown in Figure 8.11(a). Another separating hyperplane is the supporting hyperplane $H := \{x \in \mathbb{R}^n : a^\mathsf{T}x = a^\mathsf{T}\hat{x}^*\}$ of $\mathrm{cl}(X)$ at $\hat{x}^*$ (the dashed line in 8.11(a)).



(a) Case 1: $x^* \notin \mathrm{cl}(X)$ (b) $x^* \in \mathrm{cl}(X) \setminus \mathrm{int}(X)$

Figure 8.11 Proof of Theorem 8.10. The vectors a, a_i are in the normal cones of $\mathrm{cl}(X)$ at $\hat{x}^*$ and $\hat{x}_i$ respectively and $H := \{x \in \mathbb{R}^n : a^\mathsf{T}x = a^\mathsf{T}x^*\}$ is a hyperplane separating $\mathrm{cl}(X)$ and x^*. In Case 1 the separating hyperplane is nonunique: even with the same a, b can take any value in $(a^\mathsf{T}\hat{x}^*, a^\mathsf{T}x^*)$ and $\{x : a^\mathsf{T}x = b\}$ will be a separating hyperplane.

We now show (8.22b) by explicitly constructing a $b \in (a^\mathsf{T}\hat{x}^*, a^\mathsf{T}x^*)$ so that $H := \{x \in \mathbb{R}^n : a^\mathsf{T}x = b\}$ is a separating hyperplane (see Figure 8.11(a)). We claim that we can choose any $z = \beta\hat{x}^* + (1-\beta)x^*$ between $\hat{x}^*$ and x^* for some $\beta \in (0,1)$ and let $b := a^\mathsf{T}z$. To see this we have from (8.23b)

$$a^\mathsf{T}x \;\leq\; a^\mathsf{T}\hat{x}^* = a^\mathsf{T}z - a^\mathsf{T}(z - \hat{x}^*) \;<\; a^\mathsf{T}z, \qquad\qquad x \in \mathrm{cl}(X)$$

proving the first half of (8.22b), where the last inequality follows because

$$a^\mathsf{T}(z - \hat{x}^*) \;=\; (1-\beta)\,a^\mathsf{T}(x^* - \hat{x}^*) \;=\; (1-\beta)\|x^* - \hat{x}^*\|_2 > 0$$

For the second half of (8.22b) we have

$$a^{\mathsf{T}}(x^* - z) \;=\; \beta a^{\mathsf{T}}(x^* - \hat{x}^*) \;>\; \beta\|x^* - \hat{x}^*\|_2 \;>\; 0$$

as desired.

Part 1. In view of part 1 we only need to consider $x^* \in \mathrm{cl}(X) \setminus \mathrm{int}(X)$. In this case $\hat{x}^* = x^*$ and hence we cannot define a by (8.23). Take a sequence $\{x_i\}$ not in $\mathrm{cl}(X)$ such that $\lim_i x_i = x^*$. Let $\hat{x}_i$ be the projection of x_i onto the convex set $\mathrm{cl}(X)$, i.e., $\hat{x}_i := \arg\min_{x\in\mathrm{cl}(X)} \|x - x_i\|_2$. Then $(x_i - \hat{x}_i)^{\mathsf{T}}(x - \hat{x}_i) \le 0$ for all $x \in \mathrm{cl}(X)$ by the Projection Theorem 8.9. Define the normalized (error) vectors

$$a_i := \frac{x_i - \hat{x}_i}{\|x_i - \hat{x}_i\|_2}, \qquad i = 1, 2, \ldots$$

Therefore

$$a_i^{\mathsf{T}} x \le a_i^{\mathsf{T}} \hat{x}_i \;=\; a_i^{\mathsf{T}} x_i - a_i^{\mathsf{T}}(x_i - \hat{x}_i) \;\le\; a_i^{\mathsf{T}} x_i, \qquad x \in \mathrm{cl}(X) \tag{8.24}$$

where the second inequality follows because $a_i^{\mathsf{T}}(x_i - \hat{x}_i) = \|x_i - \hat{x}_i\|_2$. Since $\|a_i\| = 1$ the sequence $\{a_i, i = 1, 2, \ldots\}$ has a subsequence $\{a_{i_k}, k = 1, 2, \ldots\}$ that converges to a nonzero vector a. Taking limit as $k \to \infty$ in (8.24) yields $a^{\mathsf{T}} x \le a^{\mathsf{T}} \lim_k x_{i_k} = a^{\mathsf{T}} x^*$ for all $x \in \mathrm{cl}(X)$ as desired. This completes the proof of (8.22a). □

Theorem 8.11 (Two convex sets X and Y)**.** Suppose two disjoint sets $X, Y \in \mathbb{R}^n$ are nonempty convex.

1 There exists a nonzero $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$ such that

$$a^{\mathsf{T}} x \;\le\; b \;\le\; a^{\mathsf{T}} y, \qquad x \in \mathrm{cl}(X),\; y \in \mathrm{cl}(Y)$$

i.e. X and Y are contained in different halfspaces of the hyperplane $H := \{x \in \mathbb{R}^n : a^{\mathsf{T}} x = b\}$.

2 If $\mathrm{cl}(X) \cap \mathrm{cl}(Y) = \emptyset$, i.e., $\min_{x\in\mathrm{cl}(X)} \min_{y\in\mathrm{cl}(Y)} \|x - y\|_2 > 0$, then there exists $b \in \mathbb{R}$ such that the hyperplane $H := \{x \in \mathbb{R}^n : a^{\mathsf{T}} x = b\}$ strictly separates X and Y:

$$a^{\mathsf{T}} x \;<\; b \;<\; a^{\mathsf{T}} y, \qquad x \in \mathrm{cl}(X),\; y \in \mathrm{cl}(Y)$$

Proof Consider the set $W := \{x - y : x \in X, y \in Y\}$. W is nonempty convex. Moreover the origin $0 \notin W$. Apply Theorem 8.10 to W and $x^* = 0$. Then there exists a nonzero a such that $a^{\mathsf{T}}(x - y) \le 0$ for all $x - y \in \mathrm{cl}(W)$, or $a^{\mathsf{T}} x \le a^{\mathsf{T}} y$ for all $x \in \mathrm{cl}(X)$, $y \in \mathrm{cl}(Y)$.

When $\mathrm{cl}(X) \cap \mathrm{cl}(Y) = \emptyset$, then $x^* \notin \mathrm{cl}(W)$ and hence Theorem 8.10 guarantees a $b \in (a^{\mathsf{T}}\hat{x}^*, a^{\mathsf{T}} x^*)$ such that the inequalities are strict, where $\hat{x}^*$ is the projection of x^* onto W. □

8.2.5 Farkas Lemma

A very useful result is the following theorem which, e.g., underlies the strong duality of linear programming. It is a simple consequence of the separating hyperplane

Theorem 8.10. Recall that if $\{a_1,\ldots,a_n\}$ are the column vectors of $A \in \mathbb{R}^{m\times n}$ then $\text{cone}(A) := \text{cone}(\{a_1,\ldots,a_n\}) \subseteq \mathbb{R}^m$.

Theorem 8.12 (Farkas Lemma)**.** Let $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. Then

1 Exactly one of the following holds:
   1 $b \in cone(A)$: There exists an $x \in \mathbb{R}^n$ such that $Ax = b$ and $x \geq 0$.
   2 $b \notin cone(A)$: There exists an $y \in \mathbb{R}^m$ such that $y^\mathsf{T}A \geq 0$ and $y^\mathsf{T}b < 0$.
2 Exactly one of the following holds:
   1 $b \in range(A)$: There exists an $x \in \mathbb{R}^n$ such that $Ax = b$.
   2 $b \notin range(A)$: There exists an $y \in \mathbb{R}^m$ such that $y^\mathsf{T}A = 0$ and $y^\mathsf{T}b \neq 0$.

A variant of Theorem 8.12.1 is: Exactly one of the following holds:

1 There exists an $x \geq 0$ such that $Ax \leq b$.
2 There exists an $y \geq 0$ such that $y^\mathsf{T}A \geq 0$ and $y^\mathsf{T}b < 0$.

Its proof is similar to that for Theorem 8.12 but considers $Y := \{y \in \mathbb{R}^m : Ax \leq y, x \geq 0\} = \{Ax + s : x \geq 0, s \geq 0\}$ instead of $\text{cone}(A)$. (Exercise 8.14).

Proof of Theorem 8.12 For part 1, we clearly cannot have both because otherwise, $y^\mathsf{T}b = y^\mathsf{T}Ax \geq 0$ contradicting $y^\mathsf{T}b < 0$. According to the Carathéodory Theorem 8.7, any $b \in \text{cone}(A)$ can be expressed as $b = \sum_{i=1}^{k} \alpha_i a_i$ for some $k \leq m$, $\alpha_i > 0$, and k linearly independent column vectors a_i of A. Therefore $Ax = b$ for some $x \geq 0$ if and only if $b \in \text{cone}(A) \subseteq \mathbb{R}^m$. Suppose there exists no such x. We now prove that there must exist $y \in \mathbb{R}^m$ such that $y^\mathsf{T}A \geq 0$ and $y^\mathsf{T}b < 0$, by applying Theorem 8.10 to the closed convex cone $\text{cone}(A)$ and the point b. Since $b \notin \text{cone}(A)$ there exists $y \in \mathbb{R}^m$ such that $y^\mathsf{T}b < y^\mathsf{T}z$ for all $z \in \text{cone}(A)$.[3] Since $0 \in \text{cone}(A)$ we have $y^\mathsf{T}b < 0$. Moreover $y^\mathsf{T}A \geq 0$ because otherwise, if $\epsilon := y^\mathsf{T}a_i < 0$ for any column vector a_i of A, then $ta_i \in \text{cone}(A)$ for any $t \geq 0$ and $y^\mathsf{T}(ta_i) = t\epsilon \to -\infty$ as $t \to \infty$, contradicting $y^\mathsf{T}b < y^\mathsf{T}z$ for all $z \in \text{cone}(A)$.

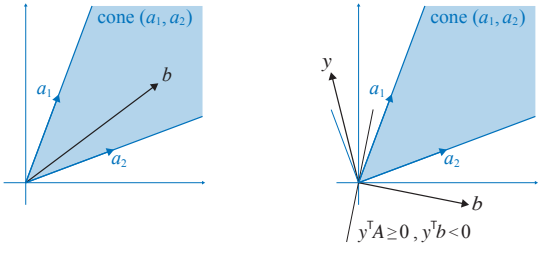
Part 2 of the theorem is a consequence the rank-nullity theorem which says that $\mathbb{R}^m$ can be decomposed into two orthogonal subspaces, $\text{null}(A^\mathsf{T})$ and $\text{range}(A)$ (see (A.1) in Chapter A.1.2). Decompose $b \in \mathbb{R}^m$ into two orthogonal components $b =: b_1 + b_2$ with $b_1 \in \text{null}(A^\mathsf{T})$ and $b_2 \in \text{range}(A)$, i.e., $A^\mathsf{T}b_1 = 0$ and $b_2 = Ax$ for some $x \in \mathbb{R}^n$. Either b is in $\text{range}(A)$ (i.e., $b_1 = 0$ and $Ax = b$) or there exists a nonzero $y := b_1 \in \text{null}(A^\mathsf{T})$ such that $A^\mathsf{T}y = 0$ and

$$y^\mathsf{T}b = y^\mathsf{T}b_1 + y^\mathsf{T}b_2 = \|b_1\|^2 > 0$$

where the last equality follows because b_1 and b_2 are orthogonal. □

Part 1 of Theorem 8.12 is illustrated in Figures 8.12. Either b is in $\text{cone}(A)$ or b is not.

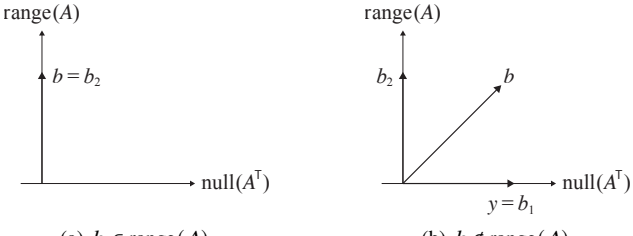
[3] The argument here that concludes $y^\mathsf{T}b < 0 \leq y^\mathsf{T}A$ is typical and is used in many proofs in this chapter and in Chapter 12. It originates from the Projection Theorem 8.9.



(a) $b \in \text{cone}(a_1, a_2)$ (b) $b \notin \text{cone}(a_1, a_2)$

Figure 8.12 Theorem 8.12.1: Farkas Lemma

In the first case, $b = Ax$ for some $x \geq 0$ according to the Carathéodory Theorem 8.7, as shown in Figures 8.12(a). Otherwise, let $\text{cone}^*(A) := \{y \in \mathbb{R}^m : y^T z \geq 0 \, \forall z \in \text{cone}(A)\}$; see Figures 8.12(b). This is called the dual cone of $\text{cone}(A)$ and studied in Chapter 12.1.1. Since b is outside $\text{cone}(A)$, there must exist an y in the intersection of $\text{cone}^*(A)$ and the set $\{b\}^\circ := \{y \in \mathbb{R}^m : y^T b \leq 0\}$ (called the polar cone of $\{b\}$ in Chapter 12.1.1) such that $y^T A \geq 0$ and $y^T b < 0$. Part 2 of Theorem 8.12 is illustrated in Figure 8.13.



(a) $b \in \text{range}(A)$ (b) $b \notin \text{range}(A)$

Figure 8.13 Theorem 8.12.2: Decomposition of $\mathbb{R}^m$ into range(A) and null(A^T).

See Exercise 12.6 for an application of the Farkas Lemma to derive the polar cone of a pre-image of the nonpositive quadrant under a linear transformation.

We next study various characterizations of optimal solutions, including the KKT condition, on which many optimization algorithms are based.

8.3 General theory: optimality conditions

Consider the optimization problem (8.13) reproduced here:

$$\min_{x \in \mathbb{R}^n} \; f(x) \;\; \text{subject to} \;\; x \in X$$

In this section we develop a basic theory to answer the following questions:

Q1 How to characterize optimal solutions?
Q2 When will optimal solutions exist and when will it be unique?

Associated with (8.13) is a dual problem $\max_\mu d(\mu)$. Q1 is important because many algorithms to compute an optimal solution (studied in Chapter 8.5) are based on necessary conditions for optimality; these conditions are often sufficient for convex programs. To answer Q1 we show in Chapter 8.3.1 that a saddle point (x^*, μ^*) is optimal for both the primal and the dual problems and closes the duality gap (Saddle-point Theorem 8.14). This characterization does not require the cost function f to be smooth (e.g. continuous or differentiable) or convex or the feasible set X to be convex. In Chapter 8.3.2 we show that (x^*, μ^*) is a saddle point if and only if it satisfies the KKT condition (KKT Theorem 8.15). This characterization requires the cost function f and constraint functions to be continuously differentiable and convex (with affine equality constraints). These results characterize the primal and dual optimal solutions but do not ensure their existence.

For Q2 we show in Chapter 8.3.3 that continuity of the cost function f and compactness of the feasible set X is sufficient for the existence of primal solutions x^* (Theorem 8.16). Strict convexity of f ensures the uniqueness of x^*. We show in Chapter 8.3.4 that if the primal optimal value is finite and a kind of feasibility condition called constraint qualification is satisfied then the duality gap is zero and dual optimal solutions exist (Slater Theorem 8.17). These results are summarized in Table 8.1.

	Primal-dual characterization	Assumptions
Th 8.14	saddle point = p-d optimality + strong duality	arbitrary f, g, h
Th 8.15	KKT point = saddle point	diff. conv. f and h, affine g
	Existence	
Th 8.16	primal optimal x^*	cont. f, compact X
Th 8.17	dual optimal λ^* & strong duality	conv. f and h, affine g, finite f^*, Slater cond.
Co 8.18	combination of Ths 8.14, 8.15, 8.16, 8.17	intersection

Table 8.1 Summary of characterization and existence of primal and dual optimal solutions.

As summarized in Table 8.1 smoothness is required for the KKT Theorem (continuously differentiable cost and constraint functions) and the existence of primal optimal solutions (continuous cost function). Neither the Saddle-point Theorem 8.14 nor the Slater Theorem 8.17 requires smoothness. These results are generalized to a nonsmooth setting in Chapter 12 when the feasible set is convex.

8.3.1 Characterization: saddle point = p-d optimality + strong duality

Primal problem.

We now study the case where the feasible set $X \subseteq \mathbb{R}^n$ is specified by a set of equality and inequality constraints. Consider

$$f^* := \min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t.} \quad g(x) = 0,\ h(x) \leq 0 \tag{8.25}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^m$ and $h : \mathbb{R}^n \to \mathbb{R}^l$ are arbitrary real-valued functions. In particular f, g, h are not necessarily convex or differentiable or even continuous. We will call this problem the *primal problem*.

Associated with every constrained optimization problem (8.25) (at least partially) specified by equality and inequality constraints is a dual problem, defined as follows.

Dual problem.

Associated with the equality constraint is the *dual variable* $\lambda \in \mathbb{R}^m$ and associated with the inequality cosntraint is the dual variable $\mu \in \mathbb{R}^l_+$. Define the *Lagrangian function* or the *Lagrangian* associated with (8.25) as the function $L : \mathbb{R}^{n+m+l} \to \mathbb{R}$:

$$L(x, \lambda, \mu) := f(x) + \lambda^\mathsf{T} g(x) + \mu^\mathsf{T} h(x), \qquad x \in \mathbb{R}^n,\ \lambda \in \mathbb{R}^m,\ \mu \in \mathbb{R}^l \tag{8.26a}$$

For any (λ, μ) define the *dual function* by the unconstrained minimization of the Lagrangian over the primal variable x:

$$d(\lambda, \mu) := \min_{x \in \mathbb{R}^n} L(x, \lambda, \mu) \tag{8.26b}$$

The *dual problem* of (8.25) is defined to be:

$$d^* := \max_{\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^l} d(\lambda, \mu) \quad \text{s.t} \quad \mu \geq 0 \tag{8.26c}$$

Let $X := \{x \in \mathbb{R}^n : g(x) = 0,\ h(x) \leq 0\}$ denote the primal feasible set and $Y := \{(\lambda, \mu) \in \mathbb{R}^{m+l} : \mu \geq 0\}$ the dual feasible set. A primal feasible point $x^* \in X$ is called *primal optimal* if x^* solves (8.25) and a dual feasible point $(\lambda^*, \mu^*) \in Y$ is called *dual optimal* if (λ^*, μ^*) solves (8.26). We also called such an (x^*, λ^*, μ^*) *primal-dual* optimal. It is important that the minimization over x in the dual problem (8.26) is unconstrained. It converts the constrained minimization (8.25) into an unconstrained minimization over x under certain conditions; see Remark 8.5.

The dual problem (8.26) always provides a lower bound on the primal problem (8.25) for arbitrary cost and constraint functions f, g, h (even extended real-valued functions studied in Chapter 12).

Lemma 8.13 (Weak duality)**.** If $(x, \lambda, \mu) \in X \times Y$ is a primal-dual feasible point then $d(\lambda, \mu) \leq f(x)$.

Proof Since (x,λ,μ) is primal-dual feasible we have $\lambda^\mathsf{T} g(x) = 0$ and $\mu^\mathsf{T} h(x) \le 0$ and hence $L(x,\lambda,\mu) \le f(x)$ from (8.26a). Therefore

$$d(\lambda,\mu) \;:=\; \min_{x'\in\mathbb{R}^n} L(x',\lambda,\mu) \;\le\; L(x,\lambda,\mu) \;\le\; f(x)$$

as desired. □

The weak duality Lemma 8.13 implies in particular that the dual objective value d^* lower bounds the primal objective value f^*:

$$d^* \;:=\; \max_{\lambda,\mu\ge 0} d(\lambda,\mu) \;\le\; \min_{x\in X} f(x) \;=:\; f^* \tag{8.27}$$

This holds whether or not the primal problem is convex and whether or not these values are bounded: if the primal optimal value is $f^* = -\infty$ then the dual problem is infeasible; if the dual optimal value is $d^* = \infty$ then the primal problem is infeasible. The gap $f^* - d^*$ is called the *duality gap*. For general nonlinear optimization the duality gap can be strictly positive, and even unbounded. If the primal problem (8.25) is convex and a certain constraint qualification is satisfied, then the duality gap is zero (Theorem 8.17). In this case we say *strong duality* holds. Before we study in Chapters 8.3.3 and 8.3.4 the existence of primal and dual optimal solutions (x^*,λ^*,μ^*) that closes the duality gap, we first characterize them.

Saddle point.

For the duality gap to be zero and for the primal and dual problems to both attain their optimal values, it is necessary and sufficient that a saddle point exists for arbitrary f, g, h. To define a saddle point we first claim that the primal problem can be written in terms of L:

$$f^* \;=\; \min_x \max_{\lambda,\mu\ge 0} L(x,\lambda,\mu) \tag{8.28}$$

To prove (8.28), note that given any infeasible $x \notin X := \{x : g(x) = 0,\ h(x) \le 0\}$, it is clear that $\max_{\lambda,\mu\ge 0} L(x,\lambda,\mu)$ is unbounded. Therefore

$$\min_x \max_{\lambda,\mu\ge 0} L(x,\lambda,\mu) \;=\; \min_{x\in X} \max_{\lambda,\mu\ge 0} L(x,\lambda,\mu) \tag{8.29a}$$

Fix any $x \in X$. On the one hand, $L(x,\lambda,\mu) \le f(x)$ for any $\mu \ge 0$, and hence

$$\min_{x\in X} \max_{\lambda,\mu\ge 0} L(x,\lambda,\mu) \;\le\; \min_{x\in X} f(x) \;=:\; f^* \tag{8.29b}$$

On the other hand, $\max_{\lambda,\mu\ge 0} L(x,\lambda,\mu) \ge L(x,\lambda,0) = f(x)$ since $x \in X$, and hence

$$\min_{x\in X} \max_{\lambda,\mu\ge 0} L(x,\lambda,\mu) \;\ge\; \min_{x\in X} f(x) \;=:\; f^* \tag{8.29c}$$

Combining (8.29) gives

$$f^* \;=\; \min_x \max_{\lambda,\mu\ge 0} L(x,\lambda,\mu) \;=\; \min_{x\in X} \max_{\lambda,\mu\ge 0} L(x,\lambda,\mu) \tag{8.30}$$

proving (8.28). Therefore weak duality (8.27) can also be expressed symmetrically in terms of the Lagrangian L:

$$d^* := \max_{(\lambda,\mu)\in Y} \min_{x\in\mathbb{R}^n} L(x,\lambda,\mu) \le \min_{x\in\mathbb{R}^n} \max_{(\lambda,\mu)\in Y} L(x,\lambda,\mu) =: f^* \tag{8.31}$$

An important feature of (8.31) is that the minimization over x is unconstrained.[4]

Definition 8.8 (Saddle point)**.** A point $(x^*,\lambda^*,\mu^*) \in \mathbb{R}^n \times Y$ is called a *saddle point* of the Lagrangian L if it satisfies

$$\max_{(\lambda,\mu)\in Y} L(x^*,\lambda,\mu) = L(x^*,\lambda^*,\mu^*) = \min_{x\in\mathbb{R}^n} L(x,\lambda^*,\mu^*) \tag{8.32}$$

where $Y := \{(\lambda,\mu) \in \mathbb{R}^{m+l} : \mu \ge 0\}$. □

Remark 8.3 (Equivalent definitions of saddle point)**.** 1 If $(x^*,\lambda^*,\mu^*) \in \mathbb{R}^n \times Y$ is a saddle point then necessarily $x^* \in X$ is primal feasible because otherwise, $\max_{(\lambda,\mu)\in Y} L(x^*,\lambda,\mu)$ is unbounded but $L(x^*,\lambda^*,\mu^*)$ is finite since f,g,h are real-valued. Therefore, when f,g,h are real-valued, we can define a saddle point without loss of generality as a primal-dual feasible point $(x^*,\lambda^*,\mu^*) \in X \times Y$ that satisfies (8.32).

2 An equivalent specification of a saddle point (x^*,λ^*,μ^*) is (Exercise 8.15):

$$(x^*,\lambda^*,\mu^*) \in X\times Y, \quad L(x^*,\lambda^*,\mu^*) = \min_{x\in\mathbb{R}^n} L(x,\lambda^*,\mu^*), \quad \mu^{*\mathsf{T}} h(x^*) = 0 \tag{8.33}$$

i.e., $\max_{(\lambda,\mu)\in Y} L(x^*,\lambda,\mu) = L(x^*,\lambda^*,\mu^*)$ in (8.32) can be replaced by primal feasibility and complementary slackness. □

Remark 8.4 (Partial dualization)**.** The minimization over x in Definition 8.8 is unconstrained because all constraints of (8.25) have been dualized. The constraints can also be partially dualized. Specifically suppose (8.25) takes the form

$$f^* := \min_{x\in\mathbb{R}^n} f(x) \quad \text{s.t.} \quad x \in X',\ g(x) = 0,\ h(x) \le 0 \tag{8.34a}$$

where $X' \subseteq \mathbb{R}^n$. The Lagrangian L is still defined by (8.26a), but the dual function is now defined to be $d(\lambda,\mu) := \min_{x\in X'} L(x,\lambda,\mu)$ and the dual problem is

$$d^* := \max_{(\lambda,\mu)\in Y} \min_{x\in X'} L(x,\lambda,\mu) \tag{8.34b}$$

where $Y := \{(\lambda,\mu) \in \mathbb{R}^{m+l} : \mu \ge 0\}$. Instead of (8.31) and (8.32), strong duality holds if

$$\max_{(\lambda,\mu)\in Y} \min_{x\in X'} L(x,\lambda,\mu) = \min_{x\in X'} \max_{(\lambda,\mu)\in Y} L(x,\lambda,\mu)$$

and $(x^*,\lambda^*,\mu^*) \in X' \times Y$ is a saddle point if

$$\max_{(\lambda,\mu)\in Y} L(x^*,\lambda,\mu) = L(x^*,\lambda^*,\mu^*) = \min_{x\in X'} L(x,\lambda^*,\mu^*)$$

[4] The weak duality (8.31) can be interpreted as a two-person zero-sum game where a player tries to maximize $L(x,\lambda,\mu)$ over $(\lambda,\mu) \in Y$ and the other player tries to minimize $L(x,\lambda,\mu)$ over $x \in \mathbb{R}^n$. The inequality (8.31) expresses the second-mover advantage: the player that makes the first move is generally disadvantaged. A saddle point (x^*,λ^*,μ^*) is a Nash equilibrium of this game.

All saddle point results extend to the case of partial dualization with obvious modifications (see also Chapter 12.7). □

The next result Theorem 8.14 states that a saddle point (x^*, λ^*, μ^*) of L solves both the primal and the dual problems and closes the duality gap. It does not require any of the functions f, g, h to be convex or smooth (e.g., differentiable or continuous) or the feasible sets X, Y to be compact (Y is obviously not compact). It is simply a reinterpretation of a saddle point in terms of the primal problem (8.28) and dual problem (8.26). It only characterizes a saddle point but does not ensure its existence. We will study the existence of primal and dual optimal solutions in Chapters 8.3.3 and 8.3.4.

Theorem 8.14 (Saddle-point Theorem)**.** Consider the primal problem (8.25) and its dual (8.26). A point (x^*, λ^*, μ^*) is a saddle point if and only if

1 It is primal-dual optimal, i.e., x^* is optimal for (8.25) and (λ^*, μ^*) is optimal for (8.26); and
2 The duality gap is zero at (x^*, λ^*, μ^*), i.e.,

$$d(\lambda^*, \mu^*) = d^* = f^* = f(x^*) \tag{8.35}$$

In particular a saddle point (x^*, λ^*, μ^*), if it exists, attains both the primal and dual objective values (f^*, d^*).

Proof Suppose (x^*, λ^*, μ^*) is a saddle point, i.e., it satisfies (8.32). As explained in Remark 8.3, when the functions f, g, h are real-valued, a saddle point is necessarily primal-dual feasible, in particular, $x^* \in X$. Then we have

$$f(x^*) = L(x^*, \lambda, 0) \leq \max_{(\lambda,\mu)\in Y} L(x^*, \lambda, \mu) = \min_{x\in\mathbb{R}^n} L(x, \lambda^*, \mu^*) =: d(\lambda^*, \mu^*)$$

where the second equality follows from (8.32) and the last equality follows from the definition of the dual objective function d. Since $(x^*, \lambda^*, \mu^*) \in X \times Y$ is feasible, the weak duality Lemma 8.13 implies that

$$f(x^*) = d(\lambda^*, \mu^*)$$

The definition of f^* and d^* and weak duality (8.27) then imply

$$d(\lambda^*, \mu^*) \leq d^* \leq f^* \leq f(x^*) = d(\lambda^*, \mu^*)$$

which is (8.35). This also shows that (x^*, λ^*, μ^*) is primal-dual optimal.

Conversely suppose $(x^*, \lambda^*, \mu^*) \in X \times Y$ is primal-dual optimal and satisfies (8.35). Since $g(x) = 0$ and $\mu^\mathsf{T} h(x) \leq 0$ for any $(x, \lambda, \mu) \in X \times Y$, we have

$$L(x^*, \lambda^*, \mu^*) \leq \max_{(\lambda,\mu)\in Y} L(x^*, \lambda, \mu) \leq f(x^*) = d(\lambda^*, \mu^*) := \min_{x\in\mathbb{R}^n} L(x, \lambda^*, \mu^*) \leq L(x^*, \lambda^*, \mu^*)$$

where the second inequality follows because $g(x^*) = 0$ and $h(x^*) \leq 0$, the first equality follows from (8.35), and the second equality follows from the definition of d. Hence all inequalities above hold with equality, proving that (x^*, λ^*, μ^*) is a saddle point. □

Theorem 8.14 and (8.33) lead to a common characterization of attainment of optimality and strong duality: (x^*, λ^*, μ^*) attains primal-dual optimality and strong duality $f^* = d^*$ if and only if $(x^*, \lambda^*, \mu^*) \in X \times Y$ is primal-dual feasible and

$$x^* \in \arg\min_{x \in \mathbb{R}^n} L(x, \lambda^*, \mu^*), \qquad \mu^{*\mathsf{T}} g(x^*) = 0$$

Remark 8.5 (Solving dual problems)**.** It is important that the minimization over $x \in \mathbb{R}^n$ in the primal problem (8.25) and its dual (8.26c), reproduced here:

$$f^* := \min_{x \in \mathbb{R}^n} \max_{(\lambda,\mu) \in Y} L(x, \lambda, \mu) \tag{8.36}$$

$$d^* := \max_{(\lambda,\mu) \in Y} \min_{x \in \mathbb{R}^n} L(x, \lambda, \mu) \tag{8.37}$$

is unconstrained. We can interpret the dual problem as converting the constrained primal problem (8.25) into an unconstrained minimization where the primal constraints are replaced by the penalty terms $\lambda^{\mathsf{T}} g(x) + \mu^{\mathsf{T}} h(x)$ in the Lagrangian $L(x, \lambda, \mu)$. Given an $(\lambda, \mu) \in Y$, solving the inner unconstrained problem $\min_x L(x, \lambda, \mu)$ can be much easier than solving (8.25), e.g., when $\nabla_x L(x, \lambda, \mu) = 0$ can be solved explicitly. In this case, if strong duality holds, we can solve (8.25) by solving the dual problem (8.37).

When the primal constraints are partially dualized, as explained in Remark 8.4, the primal and dual problems become

$$f^* := \min_{x \in X'} \max_{(\lambda,\mu) \in Y} L(x, \lambda, \mu)$$

$$d^* := \max_{(\lambda,\mu) \in Y} \min_{x \in X'} L(x, \lambda, \mu)$$

Solving the dual problem is advantageous if strong duality holds and, given an $(\lambda, \mu) \in Y$, solving the inner problem $\min_{x \in X'} L(x, \lambda, \mu)$ is much easier than solving (8.25).

Even if strong duality does not hold, solving the dual problem yields a lower bound on the primal objective value f^* which can be useful in practice. □

8.3.2 Characterization: KKT point = saddle point

We now consider the primal problem (8.25) and its dual problem (8.26) under the assumption that the cost function f and the inequality function h are convex and continuously differentiable (see Chapter 12.3.1 on continuously differentiability), and the equality function $g(x) = Ax - b$ is affine. While the duality theory can be developed when some or all of the constraints are dualized (see Remark 8.4), the KKT theory needs all constraints to be dualized.

KKT condition.

The *KKT condition* on (x,λ,μ) associated with the primal and dual problems (8.25)(8.26) is defined by the following system of equations:

$$\begin{aligned}
&\text{Stationarity :} && \nabla_x L(x,\lambda,\mu) = 0 && \text{(8.38a)}\\
&\text{Primal feasibility :} && g(x) = 0,\quad h(x) \le 0 && \text{(8.38b)}\\
&\text{Dual feasibility :} && \mu \ge 0 && \text{(8.38c)}\\
&\text{Complementary slackness :} && \mu^{\mathsf{T}} h(x) = 0 && \text{(8.38d)}
\end{aligned}$$

where $\nabla_x L$ is the column vector whose ith entry is $\frac{\partial L}{\partial x_i}$. The stationarity (8.38a) is explicitly:

$$\text{Stationarity :} \qquad \nabla f(x) + \nabla g(x)\lambda + \nabla h(x)\mu = 0 \qquad \text{(8.38e)}$$

where $\nabla g(x) = \left[\frac{\partial g}{\partial x}\right]^{\mathsf{T}} \in \mathbb{R}^{n\times m}$ and $\nabla h(x) = \left[\frac{\partial h}{\partial x}\right]^{\mathsf{T}} \in \mathbb{R}^{n\times l}$ are the Jacobian functions of g and h respectively.

Definition 8.9 (KKT point)**.** A primal variable x^* is called a *stationary point* and a dual variable (λ^*,μ^*) a *Lagrange multiplier* (vector) of (8.25) if (x^*,λ^*,μ^*) satisfies (8.38), i.e., if

$$\nabla_x L(x^*,\lambda^*,\mu^*) = 0,\quad g(x^*) = 0,\quad h(x^*) \le 0,\quad \mu^* \ge 0,\quad \mu^{*\mathsf{T}} h(x^*) = 0 \qquad (8.39)$$

We also call such a point $(x^*,\lambda^*,\mu^*) \in X \times Y$ a *KKT point*. □

Like a saddle point, a KKT point is necessarily primal-dual feasible. For general functions f,g,h, the KKT condition is necessary for (x^*,λ^*,μ^*) to be primal-dual optimal. If f,h are convex and continuously differentiable functions and g is affine, then it is also sufficient; moreover a KKT point is a saddle point and attains strong duality.

Theorem 8.15 (KKT Theorem)**.** Consider the primal problem (8.25) and its dual (8.26). Suppose f,h are convex and continuously differentiable and $g(x) = Ax - b$ is affine. Consider an arbitrary point (x^*,λ^*,μ^*). The following are equivalent:

1 (x^*,λ^*,μ^*) is a saddle point.
2 (x^*,λ^*,μ^*) satisfies the KKT condition (8.39).
3 (x^*,λ^*,μ^*) is primal-dual optimal and closes the duality gap, i.e., $d(\lambda^*,\mu^*) = d^* = f^* = f(x^*)$.

Proof As discussed above, a saddle point (Remark 8.3), a KKT point and a primal-dual optimum are necessarily primal-dual feasible and hence we can restrict ourselves without loss of generality to $(x^*,\lambda^*,\mu^*) \in X \times Y$. The equivalence of the first and the third assertions is proved in Theorem 8.14 and holds for arbitrary functions f,g,h, not necessarily convex or continuously differentiable. To show the equivalence of the first two assertions, since (x^*,λ^*,μ^*) is primal-dual feasible, we only need to show

the complementary slackness condition (8.38d) and the stationarity condition (8.38a). As we will see complementary slackness does not require f, g, h to be convex or continuously differentiable; stationarity being a first-order condition requires both.

Suppose (x^*, λ^*, μ^*) is a saddle point, i.e.,

$$\max_{(\lambda,\mu)\in Y} L(x^*, \lambda, \mu) \;=\; L(x^*, \lambda^*, \mu^*) \;=\; \min_x L(x, \lambda^*, \mu^*) \tag{8.40}$$

The second equality in (8.40) means that x^* is an unconstrained minimizer of $L(x, \lambda^*, \mu^*)$. It is therefore necessary that $\nabla_x L(x, \lambda^*, \mu^*) = 0$ as long as f, g, h are continuously differentiable, proving stationarity (8.38a). The first equality in (8.40) reads, substituting $g(x^*) = 0$ (since $x^* \in X$),

$$f(x^*) + \max_{(\lambda,\mu)\in Y} \mu^{\mathsf{T}} h(x^*) \;=\; f(x^*) + \mu^{*\mathsf{T}} h(x^*)$$

But $\max_{(\lambda,\mu)\in Y} \mu^{\mathsf{T}} h(x^*) = 0$ since $h(x^*) \le 0$ and $\mu \ge 0$, and hence $\mu^{*\mathsf{T}} h(x^*) = 0$. Hence if (x^*, λ^*, μ^*) is a saddle point, then the KKT condition (8.39) is satisfied, for arbitrary (continuously differentiable) functions f, g, h.

Conversely suppose (x^*, λ^*, μ^*) satisfies the KKT condition (8.39). We now show that the saddle point condition (8.40) is satisfied. Since f, h are convex and $g(x) = Ax - b$ is affine, $L(x, \lambda^*, \mu^*)$ is convex in x and hence the stationarity condition $\nabla_x L(x, \lambda^*, \mu^*) = 0$ implies that $L(x^*, \lambda^*, \mu^*) = \min_x L(x, \lambda^*, \mu^*)$, proving the second equality of (8.40). For the first equality, since $g(x^*) = 0$ and $\mu^{*\mathsf{T}} h(x^*) = 0$, we have $f(x^*) = L(x^*, \lambda^*, \mu^*)$. Hence

$$L(x^*, \lambda^*, \mu^*) \;=\; f(x^*) \;\ge\; \max_{(\lambda,\mu)\in Y} f(x^*) + \lambda^{\mathsf{T}} g(x^*) + \mu^{\mathsf{T}} h(x^*) \;\ge\; L(x^*, \lambda^*, \mu^*)$$

proving $L(x^*, \lambda^*, \mu^*) = \max_{(\lambda,\mu)\in Y} L(x^*, \lambda, \mu)$. This completes the proof of the theorem. □

Remark 8.6 (Comparison: Saddle point and KKT theorems)**.** 1 The saddle point Theorem 8.14 holds without requiring f, g, h in the primal problem (8.25) to be convex or differentiable. It says that a saddle point (x^*, λ^*, μ^*) is primal-dual optimal and closes the duality gap.

2 The KKT Theorem 8.15 requires that f, h be convex and continuously differentiable and g be affine. It implies that, for a primal-dual feasible point (x^*, λ^*, μ^*), the saddle point condition (8.40) is equivalent to stationarity and complementary slackness conditions:

$$\nabla_x L(x^*, \lambda^*, \mu^*) \;=\; 0, \qquad \mu^{*\mathsf{T}} h(x^*) = 0$$

The consequence of $\nabla_x L(x^*, \lambda^*, \mu^*) = 0$ is that x^* is an unconstrained minimizer of L, i.e., $L(x^*, \lambda^*, \mu^*) = \min_x L(x, \lambda^*, \mu^*)$. As demonstrated in the proof of Theorem 8.15, if f, g, h are not convex, then primal-dual optimality of (x^*, λ^*, μ^*) and strong duality imply the KKT condition (8.39), but the converse may not hold.

3 Like Theorem 8.14, Theorem 8.15 only shows that a KKT point (x^*, λ^*, μ^*) is

primal-dual optimal and closes the duality gap, but does not guarantee its existence. We now study the existence and uniqueness of a KKT point.

□

8.3.3 Existence: primal optimal solutions

In general an optimal primal solution of a constrained optimization may not exist, even when the optimal primal value is finite, dual optimal solutions exist and strong duality holds, as the next two examples show.

Example 8.9 (Nonexistence of primal optimal)**.** Consider

$$f^* \quad := \quad \inf_{x\in\mathbb{R}} \; f(x) := x^2 \qquad \text{s.t.} \qquad x > 1$$

Clearly the primal optimal value is finite, $f^* = 1$, but no primal optimal x^* exists such that $f(x^*) = f^*$.

The Lagrangian is $L(x,\mu) := x^2 + \mu(1-x) = x^2 - \mu x + \mu$, the dual function is

$$d(\mu) \; := \; \min_x L(x,\mu) \; = \; -\frac{\mu^2}{4} + \mu$$

and hence $d^* := \max_{\mu\geq 0} d(\mu) = d(2) = 1 = f^*$, i.e., strong duality holds and $\mu^* = 2$ attains the dual optimal.

Theorem 8.15 says that for a feasible x^* to be optimal, (x^*,μ^*) must satisfy the KKT condition. In particular $2x^* = \mu^*$ and $\mu^*(1-x^*) = 0$, which cannot be satisfied when $\mu^* = 2$ and $x^* > 1$. □

The reason the primal optimal is not attained in Example 8.9 is that the primal feasible set is not closed. The next example possesses a closed (but unbounded) feasible set and has no primal optimal solution either.

Example 8.10 (Nonexistence of primal optimal)**.** Consider

$$f^* \quad := \quad \inf_{x\in\mathbb{R}} \; f(x) := e^{-x} \qquad \text{s.t.} \qquad x \geq 0$$

Clearly the primal optimal value is finite, $f^* = 0$, but no finite $x^* \in \mathbb{R}$ exists such that $f(x^*) = f^*$.

The Lagrangian is $L(x,\mu) := e^{-x} - \mu x$, the dual function is

$$d(\mu) \; := \; \min_x \; e^{-x} - \mu x \; = \; \begin{cases} 0, & \mu = 0 \\ -\infty, & \mu > 0 \end{cases}$$

and hence $d^* := \max_{\mu\geq 0} d(\mu) = d(0) = 0 = f^*$, i.e., strong duality holds and $\mu^* = 0$ attains the dual optimal.

Theorem 8.15 says that for a feasible x^* to be optimal, (x^*,μ^*) must satisfy the KKT

condition. In particular $e^{-x^*} = -\mu^*$, which cannot be satisfied by any finite x^* when $\mu^* = 0$. □

We now formalize the intuition from these two examples. Consider the general optimization problem (8.13), reproduced here

$$\min_{x \in \mathbb{R}^n} \; f(x) \;\text{ subject to }\; x \in X \tag{8.41}$$

where $X \subseteq \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$ is an arbitrary real-valued function. The next result provides a sufficient condition for the existence of a primal optimal solution x^*.

Theorem 8.16 (Existence and uniqueness of x^*)**.** Consider the optimization problem (8.41). Suppose X is nonempty and compact (closed and bounded) and f is continuous on X. Then

1 An optimal solution x^* exists.
2 Moreover the optimal solution x^* is unique if f is strictly convex.

The sufficient condition in Theorem 8.16 is a consequence of the Weierstrass theorem (Theorem 12.22 in Chapter 12.6). For an exact condition see Theorem 12.26 in Chapter 12.6. The existence of an optimal solution x^* only requires f to be continuous, not necessarily convex. Convexity is important for the efficient computation of an optimal solution because a local first-order condition is not only necessary but also sufficient for optimality when the cost function is a convex function and the feasible set is a convex set. Note that a real-valued convex function is continuous on the interior of its domain, according to Lemma 8.4.

8.3.4 Existence: dual optimal solutions and constraint qualifications

Consider the primal and dual problems (8.25)(8.26) where the feasible set is specified by a set of equalities and inequalities. Conditions that guarantee the existence and uniqueness of Lagrange multipliers (λ^*, μ^*) are called constraint qualification conditions. We describe three of them.

Constraint qualifications.

Suppose x^* is a local optimal of (8.25). Let $Y(x^*)$ be the set of Lagrange multipliers associated with x^*:

$$Y(x^*) := \left\{ \; (\lambda, \mu) \in \mathbb{R}^{m+l} \;:\; (x^*, \lambda^*, \mu^*) \text{ satisfies KKT condition (8.39)} \; \right\}$$

If $Y(x^*)$ is nonempty then it is a convex polyhedral set whether or not (8.25) is a convex program. (Recall that a set $B \subseteq \mathbb{R}^n$ is a polyhedral set if $B = \{x \in \mathbb{R}^n : Ax \leq b\}$ for some matrix A and vector b of appropriate sizes; see Chapter 8.1.2.)

The set $Y(x^*)$ of Lagrange multipliers associated with a local optimal x^* is nonempty if and only if the following condition holds at x^*:

$$\text{rank } \frac{\partial g}{\partial x}(x^*) \;=\; m, \qquad \exists \xi \in N\left(\frac{\partial g}{\partial x}(x^*)\right) \;\text{ s.t. }\; \frac{\partial h_{I(x^*)}}{\partial x}(x^*)\,\xi < 0 \tag{8.42}$$

where $N(A)$ is the null space of matrix A and $I(x^*)$ is the set of indices of inequality constraints that are active at x^* and $\frac{\partial h_{I(x^*)}}{\partial x}(x^*)$ is the $|I(x^*| \times n$ matrix of partial derivatives of h_i that are active at x^*:

$$I(x^*) \;:=\; \{i \;:\; h_i(x^*) = 0\}, \qquad \frac{\partial h_{I(x^*)}}{\partial x}(x^*) \;:=\; \left(\frac{\partial h_i}{\partial x}(x^*),\; i \in I(x^*)\right)$$

The condition (8.42) is called the *Mangasarian-Fromovitz constraint qualification* (MFCQ). The second condition of MFCQ says that the local optimal x^* can move infinitesimally in the direction of ξ and become strictly feasible.

The second constraint qualification guarantees not only the existence, but also the uniqueness, of the Lagrangian multiplier associated with a local optimal x^*:

$$\text{the rows of } \frac{\partial g}{\partial x}(x^*),\; \frac{\partial h_{I(x^*)}}{\partial x}(x^*) \text{ are linearly independent} \tag{8.43}$$

This is called the *linear independence constraint qualification* (LICQ) and it guarantees that $Y(x^*)$ is a singleton. This originates from the unique representation of the normal cone vector in terms of $\nabla g(x^*)$ and $\nabla h_{I(x^*)}(x^*)$; see Theorem 12.4 and Example 12.4. Using the Farkas lemma (Theorem 8.12) it can be shown that LICQ implies MFCQ (Exercise 8.17).

Both LICQ and MFCQ presume the existence of an optimal solution x^* for the primal problem (8.25). When an optimal x^* exists and if one of the condition is satisfied then an optimal Lagrange multiplier $(\lambda^*, \mu^*) \in Y(x^*)$ exists and (x^*, λ^*, μ^*) is a KKT point. Theorem 8.15 then implies that (x^*, λ^*, μ^*) is a saddle point that closes the duality gap and solves both the primal and the dual problems, provided f, h are convex and continuously differentiable and g is affine.

We next discuss the third constraint qualification, called the Slater condition, that does not require the existence of a primal optimal solution x^*. We will restrict ourselves to the version of the primal problem (8.25) where the equality constraint function $g(x)$ is affine. Consider the following problem:

$$f^* \;:=\; \min_{x \in \mathbb{R}^n} \; f(x) \quad \text{s.t.} \quad Ax = b, \;\; h(x) \le 0 \tag{8.44}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}^l$ are real-valued functions, and $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$. Suppose $h_1(x), \ldots, h_{\bar{l}}(x)$ are affine functions and $h_{\bar{l}+1}(x), \ldots, h_l(x)$ are nonlinear convex functions. Then the constraint qualification is:

Slater condition: There exists $\bar{x}$ such that

$$A\bar{x} \;=\; b, \quad h_i(\bar{x}) \;\le\; 0, \quad i = 1, \ldots, \bar{l}, \quad h_i(\bar{x}) \;<\; 0, \quad i = \bar{l}+1, \ldots, l \tag{8.45}$$

The Slater condition is often stated as having a strictly feasible point $\bar{x}$ because $\bar{x}$ satisfies the nonlinear inequality constraints strictly. If all $h_i(x)$ are affine then the Slater condition reduces to primal feasibility.

Strong duality and dual optimality.

Let the Lagrangian function $L : \mathbb{R}^{n+m+l} \to \mathbb{R}$ associated with the primal problem (8.44) be

$$L(x,\lambda,\mu) \;:=\; f(x)+\lambda^{\mathsf{T}}(Ax-b)+\mu^{\mathsf{T}}h(x), \quad x \in \mathbb{R}^n,\ \lambda \in \mathbb{R}^m,\ \mu \in \mathbb{R}^l \tag{8.46a}$$

The dual function is

$$d(\lambda,\mu) \;:=\; \min_{x\in\mathbb{R}^n} L(x,\lambda,\mu), \qquad \lambda \in \mathbb{R}^m,\ \mu \in \mathbb{R}^l \tag{8.46b}$$

and the *dual problem* is

$$d^* \;:=\; \max_{\lambda,\mu\geq 0} d(\lambda,\mu) \tag{8.46c}$$

Let $X := \{x \in \mathbb{R}^n : Ax = b,\, h(x) \leq 0\}$ denote the primal feasible set and $Y := \{(\lambda,\mu) \in \mathbb{R}^m \times \mathbb{R}^l : \mu \geq 0\}$ the dual feasible set.

When f, h are convex the Slater condition ensures that strong duality and the existence of a dual optimal solution (λ^*,μ^*) that attains the dual optimal value, $d(\lambda^*,\mu^*) = d^*$.

Theorem 8.17 (Slater Theorem)**.** Consider the primal problem (8.44) and its dual (8.46). Suppose the following conditions hold:

- *Finite primal value*: f^* is finite, i.e., $-\infty < f^* < \infty$.
- *Convexity*: f, h are convex.
- *Slater condition*: (8.45) holds.

Then

1. $f^* = d^*$.
2. There exists a dual optimal solution (λ^*,μ^*) with $d(\lambda^*,\mu^*) = d^*$. Moreover the set of dual optimal solutions is nonempty, convex and closed.
3. If there exists $\bar{x}$ such that $h(\bar{x}) < 0$ (i.e., the Slater condition is strict and there is no equality constraint), then the set of dual optimal solutions is nonempty, convex and compact.

Note that Theorem 8.17 does not require f, h to be smooth but only convex, e.g., it may not be continuously differentiable or even continuous. This result will be extended and proved in a nonsmooth setting in Chapter 12.7.1 as Theorem 12.28. Part 3 of Theorem 8.17 on the compactness and convexity of the dual optimal set is proved in Exercise 8.19. In particular it shows that the set D^* of dual optimal solutions is

bounded by the weak duality gap at the strict Slater point $\bar{x}$ divided by the worst-case "constraint gap" [55, Lemma 1]

$$\max_{\mu \in D^*} \|\mu\|_2 \;\leq\; \max_{\mu \in D^*} \|\mu\|_1 \;\leq\; \frac{f(\bar{x}) - d^*}{\min_i(-h_i(\bar{x}))} \;=\; \frac{f(\bar{x}) - f^*}{\min_i(-h_i(\bar{x}))}$$

Since f^* is finite, weak duality implies that the dual problem can only be finite feasible or infeasible. The Slater condition in Theorem 8.17 guarantees that it is feasible and attained. It does not however guarantee that the finite primal optimal is attained, i.e., there may not be a feasible x^* such that $f(x^*) = f^*$ when the feasible set is not compact, as Examples 8.9 and 8.10 show. In these examples, both conditions in Theorem 8.17 are satisfied and hence f^* is finite, dual optimal solutions exist and strong duality holds. If a primal optimal solution x^* does exist and (λ^*, μ^*) is the associated Lagrange multiplier, i.e., (x^*, λ^*, μ^*) is a KKT point, then Theorem 8.15 implies that (x^*, λ^*, μ^*) is also a saddle point that is primal-dual optimal and closes the duality gap. Note that for both the Slater Theorem 8.17 and the KKT Theorem 8.15, it is not enough for the feasible set to be convex. It has to be specified by a convex constraint function $h(x)$ for these theorems to apply. We will discuss in Chapter 8.3.7 potential issues that may arise when the convex feasible set is represented by nonconvex constraint functions.

The next example shows that the importance of the Salter condition.

Example 8.11 (Nonexistence of dual optimal solution)**.** Consider

$$f^* \;:=\; \inf_{x \in \mathbb{R}} \; f(x) := 2x \quad \text{s.t.} \quad x^2 \leq 0$$

The feasible set is $\{x = 0\}$ and the Slater condition does not hold. We now show that the dual problem is feasible, but dual optimality is not attained even though f^* is finite and attained, $f^* = f(0) = 0$, all functions are convex, and strong duality holds.

The Lagrangian is $L(x, \mu) := 2x + \mu x^2$ and the dual function $d(\mu) := \inf_{x \in \mathbb{R}} L(x, \mu)$ is

$$d(\mu) \;=\; \begin{cases} -1/\mu & \text{if } \mu > 0 \\ -\infty & \text{if } \mu \leq 0 \end{cases}$$

Hence

$$d^* \;:=\; \sup_{\mu > 0} \; d(\mu) = -\inf_{\mu > 0} \frac{1}{\mu} \;=\; 0$$

i.e., dual optimal μ^* does not exists in $\mathbb{R}$ even though $d^* = 0 = f^* = f(0)$. □

The counterexamples to primal optimality (Theorem 8.16) and dual optimality (Slater Theorem 8.17) are summarized in Table 8.2. These examples are all primal and dual feasible. They show that one of the (primal and dual) problems having an optimal solution generally does not guarantee that the other also has an optimal solution, except for linear programs (see Chapter 8.4.2).

Compact feasible set	Primal optimality	Slater condition	Dual optimality	Strong duality	Example
no	no x^*	yes	$d^* = d(\mu^*)$	finite $f^* = d^*$	8.9, 8.10
yes	$f^* = f(x^*)$	no	no μ^*	finite $f^* = d^*$	8.11

Table 8.2 Primal-dual feasible counterexamples to Theorems 8.16 and 8.17.

In summary Theorems 8.14 and 8.15 characterize a primal-dual optimal solution (x^*, λ^*, μ^*) as a saddle point and a KKT point that closes the duality gap. Theorems 8.16 and 8.17 provide sufficient conditions for the existence of primal and dual solutions. These conditions combine to give the following result.

Corollary 8.18 (Existence, uniqueness, characterizations)**.** Consider the primal problem (8.44) and its dual (8.46). Suppose

- *Convexity and smoothness*: f, h are convex and continuously differentiable.
- *Compact* X: The primal feasible set $X := \{x \in \mathbb{R}^n : Ax = b, h(x) \leq 0\}$ is compact;
- *Finite primal value*: f^* is finite, i.e., $-\infty < f^* < \infty$;
- *Slater condition*: (8.45) holds;

Then there exists a primal-dual optimal solution $(x^*, \lambda^*, \mu^*) \in X \times Y$ to (8.44)(8.46), i.e., both the primal and dual optimal values are attained, $f^* = f(x^*)$ and $d^* = d(\lambda^*, \mu^*)$. Moreover

1 Strong duality holds $f^* = d^*$.
2 $(x^*, \lambda^*, \mu^*) \in X \times Y$ is a saddle point of the Lagrangian L.
3 $(x^*, \lambda^*, \mu^*) \in X \times Y$ is a KKT point.
4 If f is strictly convex then the primal optimal solution x^* is unique.
5 If LICQ (8.43) holds, i.e., if the rows of A and $\left\{ \frac{\partial h_i}{\partial x}(x^*) : h_i(x^*) = 0 \right\}$ are linearly independent, then the dual optimal solution (λ^*, μ^*) is unique.

8.3.5 Perturbed problem and local sensitivity

A dual optimal solution (λ^*, μ^*) can be interpreted as the sensitivity of the optimal *value* f^* to constraint perturbations. Specifically, for any $(u, v) \in \mathbb{R}^{m+l}$, consider the perturbed problem with the perturbation vector (u, v):

$$f^*(u, v) \quad := \quad \min_{x \in \mathbb{R}^n} \quad f(x) \quad \text{s.t.} \quad g(x) = u, \; h(x) \leq v \tag{8.47}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^m$ and $h : \mathbb{R}^n \to \mathbb{R}^l$ are real-valued functions. We do *not* assume that these functions are convex. The function $f^* : \mathbb{R}^{m+l} \to \mathbb{R}$ maps a perturbation vector (u, v) to a (primal) optimal value. If the perturbed problem is

infeasible at (u,v), then $f^*(u,v) := \infty$. The primal problem (8.44) is a special case of (8.47) with $(u,v) = (0,0)$ and $g(x) = Ax - b$. We discuss two properties of the function $f^*(u,v)$: an affine lower bound on $f^*(u,v)$ and the local sensitivity $\frac{\partial f^*}{\partial (u,v)}(0,0)$.

Suppose strong duality holds and dual optimality is attained for the unperturbed problem (8.47) with the perturbation vector $(0,0)$, e.g., when the conditions of the Slater Theorem 8.17 hold. Let (λ^*, μ^*) be any dual optimal solution of the unperturbed problem. The first property is an affine lower bound on the function $f^*(u,v)$ in terms of the optimal value $f^*(0,0)$ and the dual optimal solution (λ^*, μ^*) of the unperturbed problem:

$$f^*(u,v) \;\geq\; f^*(0,0) - \lambda^{*\mathsf{T}} u - \mu^{*\mathsf{T}} v, \qquad \forall (u,v) \in \mathbb{R}^{m+l} \tag{8.48}$$

The inequality (8.48) bounds the function $f^*(u,v)$ by an affine function in (u,v). To prove (8.48) let $\bar{x}$ be any feasible solution of the perturbed problem with the perturbation vector $(u,v) \in \mathbb{R}^{m+l}$, i.e., $g(\bar{x}) = u$ and $h(\bar{x}) \leq v$. Then

$$\begin{aligned} f^*(0,0) \;&=\; d(\lambda^*, \mu^*) \;:=\; \inf_{x \in \mathbb{R}^n} \; f(x) + \lambda^{*\mathsf{T}} g(x) + \mu^{*\mathsf{T}} h(x) \\ &\leq\; f(\bar{x}) + \lambda^{*\mathsf{T}} g(\bar{x}) + \mu^{*\mathsf{T}} h(\bar{x}) \;\leq\; f(\bar{x}) + \lambda^{*\mathsf{T}} u + \mu^{*\mathsf{T}} v \end{aligned}$$

where the first equality follows from strong duality for the unperturbed problem, and the last inequality follows since $\bar{x}$ is feasible for the perturbed problem and $\mu^* \geq 0$. Hence

$$f(\bar{x}) \;\geq\; f^*(0,0) - \lambda^{*\mathsf{T}} u - \mu^{*\mathsf{T}} v \qquad \text{for all feasible } \bar{x} \text{ of perturbed problem}$$

from which (8.48) follows.

The second property is local sensitivity of the optimal value $f^*(u,v)$ to constraint perturbations around $(u,v) = (0,0)$. Suppose again strong duality holds and dual optimality is attained for the unperturbed problem (8.47) with the perturbation vector $(0,0)$. Suppose further that the function $f^*(u,v)$ is differentiable at $(u,v) = (0,0)$. Then the lower bound (8.48) implies

$$f^*(te_i, 0) - f^*(0,0) \;\geq\; -t\lambda_i^*, \qquad t \in \mathbb{R}$$

where $e_i \in \mathbb{R}^m$ is the ith unit vector with a single 1 in the ith entry. Therefore, taking the limit $t \to 0$ from above and below, we have

$$\lim_{t \to 0, t > 0} \frac{f^*(te_i, 0) - f^*(0,0)}{t} \;\geq\; -\lambda_i^*, \qquad \lim_{t \to 0, t < 0} \frac{f^*(te_i, 0) - f^*(0,0)}{t} \;\leq\; -\lambda_i^*$$

Similarly for μ^*, and we conclude (since $f^*(u,v)$ is differentiable at $(u,v) = (0,0)$ by assumption):

$$\frac{\partial f^*}{\partial u_i}(0,0) \;=\; -\lambda_i^*, \qquad \frac{\partial f^*}{\partial v_i}(0,0) \;=\; -\mu_i^* \tag{8.49}$$

8.3.6 Envelope theorems

Consider a constrained optimization. Let (x, y) denote the primal and dual variables and p a parameter in the objective and/or constraint functions. For example, in optimal power flow problems, p may be line limits or nodal powers of uncontrollable loads or renewable generations. Let $(x(p), y(p))$ denote the primal-dual optimal point given the parameter p. Define the *value function* $V(p) := L(x(p), y(p); p)$ to be the Lagrangian evaluated at the optimal point $(x(p), y(p))$, as a function of p. Envelope theorems provide sufficient conditions for the differentiability of $V(p)$. The main conclusion is that the derivative of $V(p)$ is the *partial* derivative $\nabla_p L(x(p), y(p); p)$ of the Lagrangian with respect to p, evaluated at the optimal point $(x(p), y(p))$. When the feasible set is independent of p or is defined by only equality constraints then $V(p) := f(x(p); p)$ can be defined to be the optimal cost as a function of p. Its derivative is the sensitivity of the optimal cost to p and therefore of great interest in applications. This subsection collects several variants of envelope theorems.

The following saddle-point envelope theorem is from [56, Theorem 298]. It makes mild assumptions, e.g., does not need convexity or differentiability (except differentiability in parameter p), and unifies several variants.

Theorem 8.19 (Saddle-point Envelope Theorem [56])**.** Let X and Y be metric spaces and $P \subseteq \mathbb{R}^n$ be an open set. Let $L : X \times Y \times P \to \mathbb{R}$. For each $p \in P$, let $(x^*(p), y^*(p)) \in X \times Y$ be a saddle point of L, i.e.,

$$L(x^*, y; p) \;\le\; L(x^*(p), y^*(p); p) \;\le\; L(x, y^*(p); p), \qquad x \in X,\ y \in Y \tag{8.50}$$

and define the value function as

$$V(p) := L(x^*(p), y^*(p); p)$$

Suppose:

1. $x^*(p)$ and $y^*(p)$ are continuous functions (in particular, this assumes that there is a unique saddle point $(x^*(p), y^*(p))$ for each $p \in P$).
2. $\nabla_p L(x, y; p)$ exists and is jointly continuous on $X \times Y \times P$.

Then V is continuously differentiable and

$$\nabla V(p) = \nabla_p\, L(x^*(p), y^*(p); p)$$

i.e., $\frac{\partial V}{\partial p_i}(p) = \frac{\partial L}{\partial p_i}(x, y; p)$ evaluated at $(x, y) = (x^*(p), y^*(p))$.

Proof We will prove that the directional derivative of V at each $p \in P$ in each direction $h \in \mathbb{R}^n$:

$$dV(p; h) := \lim_{t \downarrow 0} \frac{V(p + th) - V(p)}{t}$$

exists[5] and equals $\frac{\partial V}{\partial p}(p)\cdot h$. This is equivalent to the differentiability of f. Moreover we will show that $\nabla V(p)$ is continuous on P.

Let $h\in\mathbb{R}^n$ be such that $[p,p+h]\subset P$ where $[p,p+h]:=\{p+th : 0\le t\le 1\}$ (such h always exists since P is open). By definition we have

$$V(p+h)-V(p)=L(x^*(p+h),y^*(p+h);p+h)\ -\ L(x^*(p),y^*(p);p)$$

The saddle point property (8.50) then implies the inequalities in the following:

$$\begin{aligned} V(p+h)-V(p)\ =\ & \underbrace{L(x^*(p+h),y^*(p+h);p+h)-L(x^*(p+h),y^*(p);p+h)}_{\ge 0}\\ &+\ L(x^*(p+h),y^*(p);p+h)-L(x^*(p+h),y^*(p);p) \qquad (8.51)\\ &+\ \underbrace{L(x^*(p+h),y^*(p);p)-L(x^*(p),y^*(p);p)}_{\ge 0} \end{aligned}$$

Since $L(x,y;p)$ is differentiable with respect to p for each (x,y), we can apply the mean value theorem to (8.51) to get

$$V(p+h)-V(p)\ge \frac{\partial L}{\partial p}(x^*(p+h),y^*(p);p_1(h))\cdot h$$

for some $p_1(h)\in[p,p+h]$. Similarly we have

$$\begin{aligned} V(p+h)-V(p) &= \underbrace{L(x^*(p+h),y^*(p+h);p+h)\ -\ L(x^*(p),y^*(p+h);p+h)}_{\le 0}\\ &\quad +\ L(x^*(p),y^*(p+h);p+h)\ -\ L(x^*(p),y^*(p+h);p)\\ &\quad +\ \underbrace{L(x^*(p),y^*(p+h);p)\ -\ L(x^*(p),y^*(p);p)}_{\le 0}\\ &\le \frac{\partial L}{\partial p}(x^*(p),y^*(p+h);p_2(h))\cdot h \end{aligned}$$

for some $p_2(h)\in[p,p+h]$. Combining, and replacing h by th, we have

$$\frac{\partial L}{\partial p}(x^*(p+th),y^*(p);p_1(th))\cdot th\le\ V(p+th)-V(p)\ \le \frac{\partial L}{\partial p}(x^*(p),y^*(p+th);p_2(th))\cdot th$$

Dividing throughout by t, taking $t\downarrow 0$ and using the continuity of $\frac{\partial L}{\partial p}$, $x(p)$ and $y(p)$ we get

$$dV(p;h)=\frac{\partial L}{\partial p}(x^*(p),y^*(p);p)\cdot h$$

for all $p\in P$ and all $h\in\mathbb{R}^n$. Hence

$$\frac{\partial V}{\partial p}(p)=\frac{\partial L}{\partial p}(x^*(p),y^*(p);p)$$

exists. Moreover it is continuous since $\frac{\partial L}{\partial p}$ is continuous on $X\times Y\times P$. □

[5] Since $V(p)$ is not assumed to be convex, the limit in the definition of $dV(p;h)$ may not exist.

Remark 8.7. It is important that the feasible sets (X,Y) are independent of p. The saddle point property (8.50) can still hold if the feasible sets (X_p,Y_p) depend on p, i.e., for all $p \in P$,

$$L(x^*(p),y;p) \;\le\; L(x^*(p),y^*(p);p) \;\le\; L(x,y^*(p);p), \qquad x \in X_p,\ y \in Y_p$$

Yet the conclusion of Theorem 8.19 in general does not hold. This is because the inequalities in $V(p+h)-V(p)$ above rely on inequalities of the form:

$$\begin{aligned} L(x^*(p),y^*(p);p) &\ge L(x^*(p),y^*(q);p) \\ L(x^*(q),y^*(p);p) &\ge L(x^*(p),y^*(p);p) \end{aligned}$$

which may not hold if $y^*(q)$ is in $Y_q \setminus Y_p$ or $x^*(q)$ is in $X_q \setminus X_p$. See Exercise 8.20 for more. □

An important implication of Remark 8.7 is that for a two-stage stochastic program with recourse, since the feasible set for the second-stage problem usually depends on the first-stage decision x, envelope theorems generally do not guarantee the differentiability of the second-stage value function or recourse function $Q(x)$ in (13.118). Hence a rigorous study of the two-stage stochastic program requires nonsmooth convex optimization theory covered in Chapter 12.

The following version is the classical envelope theorem. The key condition is that the first-order stationarity condition holds with equality, which is the reason for X to be open so that the optimal point $x^*(p)$ is in the interior of X. Note that convexity is not assumed since the proof only needs the necessity of the stationarity condition.

Theorem 8.20 (Saddle-point Envelope Theorem [56])**.** Let $X \subseteq \mathbb{R}^n$ and $P \subseteq \mathbb{R}^l$ be open sets. Consider the constrained optimization for each $p \in P$:

$$\min_{x \in X} \; f(x,p) \;\text{ s.t. }\; g(x,p) = 0$$

where $f : X \times P \to \mathbb{R}$ and $g := (g_1,\ldots,g_m) : X \times P \to \mathbb{R}^m$. Let $x^*(p)$ denote an optimal solution and $V(p) := f(x^*(p),p)$ the optimal value. Let $y \in \mathbb{R}^m$ denote the dual variable and define the Lagrangian

$$L(x,y;p) := f(x,p) + y^{\mathsf{T}} g(x,p), \qquad x \in X,\ y \in \mathbb{R}^m$$

Suppose

1 $f, g_1,\ldots,g_m$ are continuously differentiable on $X \times P$.
2 Stationarity in x: For each $p \in P$, there exist $y^*(p) \in \mathbb{R}^m$ such that the first-order stationarity condition holds with equality:

$$\nabla_x L(x^*(p),y^*(p);p) = \nabla_x f(x^*(p),p) \;+\; \nabla_x g(x^*(p),p) y^*(p) \;=\; 0$$

3 $x^*(p)$ and $y^*(p)$ are continuously differentiable functions (in particular, this assumes that the optimal primal and dual solutions exist and are unique).

Then $V(p)$ is continuously differentiable and

$$\nabla V(p) = \nabla_p\, L(x^*(p), y^*(p); p) \;=\; \nabla_p f(x^*(p), p) \,+\, \nabla_p g(x^*(p), p) y^*(p)$$

The theorem can be proved by appealing to Theorem 8.19 but a direct proof is simpler.

Proof of Theorem 8.20 $V(p)$ is continuously differentiable since $f(x,p)$ and $x^*(p)$ are. Since $x^*(p)$ satisfies $g(x^*(p), p) = 0$ we have

$$V(p) = L(x^*(p), y^*(p); p) \;=\; f(x^*(p), p) \,+\, \sum_j y_j^*(p)\, g_j(x^*(p), p)$$

Differentiability assumptions yield

$$\frac{\partial V}{\partial p_l}(p) = \sum_i \frac{\partial f}{\partial x_i}(x^*(p), p) \cdot \frac{\partial x_i^*}{\partial p_l}(x^*(p), p) \,+\, \frac{\partial f}{\partial p_l}(x^*(p), p) \,+\, \sum_j \frac{\partial y_j^*}{\partial p_l}(p) \cdot g_j(x^*(p), p)$$
$$+ \sum_j y_j^*(p) \left(\sum_i \frac{\partial g_j}{\partial x_i}(x^*(p), p) \cdot \frac{\partial x_i^*}{\partial p_l}(x^*(p), p) \,+\, \frac{\partial g_j}{\partial p_l}(x^*(p), p) \right)$$

Feasibility and stationarity in x imply:

$$g_j(x^*(p), p) \;=\; 0, \qquad \frac{\partial f}{\partial x_i}(x^*(p), p) + \sum_j y_j^*(p) \frac{\partial g_j}{\partial x_i}(x^*(p), p) \;=\; 0$$

Substituting into $\partial V / \partial p_l$ yields $\frac{\partial V}{\partial p_l} = \frac{\partial f}{\partial p_l} + \sum_j y_j \frac{\partial g_j}{\partial p_l}$, i.e.,

$$\nabla_p V(p) = \nabla_p f(x^*(p), p) \,+\, \nabla_p g(x^*(p), p) y^*(p)$$

as desired. □

Remark 8.8. It is important that the set X is open so that the first-order stationarity condition holds with equality. If the feasible set X_p depends on p, then either X_p is assumed open or $x^*(p)$ is in the interior of X_p. This means that if the constraint $x \in X_p$ is represented by $h(x,p) \leq 0$, the corresponding Lagrange multipliers will be zero at optimality so that the stationarity condition and the conclusion of the theorem will remain unchanged.

When the feasible set does not depend on p, only the cost function does, the saddle-point envelope theorems reduce to Danskin's Theorem. If the function $f(x,p)$ in Theorem 8.21 represents the Lagrangian function of a constrained optimization and (x,p) represents primal and dual variables, then the theorem implies the differentiability of the dual function when the optimal $x(p)$ is unique.

Theorem 8.21 (Danskin's Theorem)**.** Let $X \subseteq \mathbb{R}^n$ be nonempty and $f : X \times \mathbb{R}^l \to \mathbb{R}$ be a continuous function. Suppose $f(x,p)$ is convex in p for every $x \in X$. Let

$$V(p) := \sup_{x \in X} f(x,p)$$

1 Suppose X is compact so that a maximizer $x^*(p)$ always exists with $V(p) = f(x^*(p), p)$. Let the set of maximizers be

$$X^*(p) := \{x \in X : V(p) = f(x, p)\}$$

Then

1 The function $V : \mathbb{R}^l \to \mathbb{R}$ is convex and has directional derivative $dV(p; h)$ at p in the direction of $h \in \mathbb{R}^m$ given by:

$$dV(p; h) := \lim_{t \downarrow 0} \frac{V(p + th) - V(p)}{t} = \max_{x \in X^*(p)} df(x, h; p)$$

where $df(x, h; p) := \lim_{t \downarrow 0} \frac{f(x+th,p)-f(x,p)}{t}$ is the directional derivative of the function $f(\cdot, p)$.

2 If $X^*(p) = \{x^*(p)\}$ is a singleton and $f(x^*(p), \cdot)$ is differentiable in its second argument at p, then $V(p)$ is differentiable at p and

$$\nabla_p V(p) = \nabla_p f(x^*(p), p) = \left(\frac{\partial f}{\partial p_j}(x^*(p), p),\ j = 1, \ldots, m \right)$$

3 If X is compact and convex and $f(x, p)$ is convex in x for every $p \in \mathbb{R}^m$, then $X^*(p)$ is nonempty, convex and compact (according to Theorem 12.26).

2 The conclusions of 1 hold if, instead of assuming X is compact, we assume that

- $X^*(p)$ is nonempty for every $p \in \mathbb{R}^m$; and
- For every sequence $\{p_k\}$ converging to some p, there exists a bounded sequence $\{x_k^*\}$ of maximizes $x_k^* \in X^*(p)$ for all k (so that $\{x_k^*\}$ has a convergent subsequence).

Remark 8.9. 1 As for Theorem 8.19, it is important that the feasible set X does not depend on p, for the same reason discussed in Remark 8.7.

2 Theorem 8.21 is generalized in Theorem 12.19 to the case where f may not be continuous in x, X may not be compact, and $X^*(p)$ may not be a singleton. □

Theorem 8.21 guarantees the existence of directional derivative of $V(p)$ if f is jointly continuous in (x, p) and convex in p for every $x \in X$. Differentiability of V however needs uniqueness of the maximizer $x^*(p)$ and differentiability of $f(x^*(p), \cdot)$. It $f(x, p)$ is convex in p for every $x \in X$ then $V(p) := \sup_{x \in X} f(x, p)$ is convex in p. For $U(p) := \inf_{x \in X} f(x, p)$ when $f(x, p)$ is *jointly* convex in (x, p) (this is not the case with Lagrangian functions), see Remark 12.7.

8.3.7 Equivalent representations

Consider the following two convex optimization programs:

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t.} \quad Ax = b,\ h_1(x) \leq 0 \tag{8.52a}$$

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t.} \quad Ax = b,\ h_2(x) \leq 0 \tag{8.52b}$$

where f is a convex function. Suppose the feasible sets $\{x \in \mathbb{R}^n : Ax = b, h_1(x) \leq 0\}$ and $\{x \in \mathbb{R}^n : Ax = b, h_2(x) \leq 0\}$ are the same, so (8.52a) and (8.52b) are equivalent representations of the same problem in the sense that they have the same cost function f and the same feasible set.

Equivalent representations of the same problem can have different structural and computational properties. Judicious choice of problem formulation is therefore important in application. For example, the dual problem, the optimal dual value and strong duality generally depend on the primal and dual representations and may be different for different (even if equivalent) representations.

1 If both $h_1(x)$ and $h_2(x)$ are convex functions, the Slater condition is satisfied for both representations in (8.52), and their optimal primal value is finite, then the Slater Theorem 8.17 applies to both representations and hence strong duality holds and dual optimality is attained for both representations. The KKT Theorem 8.15 also applies to both representations. Even though they may have different dual problems and different KKT conditions, their optimal dual values will be the same.
2 If on the other hand $h_1(x)$ is convex but $h_2(x)$ is not, then even if the Slater condition is satisfied for both problems and their common optimal primal value is finite and the same, neither the Slater Theorem 8.17 nor the KKT Theorem 8.15 applies to (8.52b). Indeed, for (8.52b), strong duality may not hold and its dual problem may be infeasible, as the following example shows.

Example 8.12 (Equivalent representations)**.** Consider what is called a second-order cone program (studied in Chapter 8.4.4):

$$f_1^* := \min_{x \in \mathbb{R}^n} c^T x \quad \text{s.t.} \quad \|x^{n-1}\|_2 \leq x_n \tag{8.53a}$$

where $c \in \mathbb{R}^n$. Its constraint function $h_1(x) := \|x^{n-1}\|_2 - x_n$ is not differentiable at x where $x^{n-1} = 0$. To bypass this difficulty the following formulation is often solved instead:

$$f_2^* := \min_{x \in \mathbb{R}^n} c^T x \quad \text{s.t.} \quad \|x^{n-1}\|_2^2 \leq x_n^2, \ x_n \geq 0 \tag{8.53b}$$

Both problems have the same convex feasible set, the standard second-order cone $K_{\text{soc}} \subseteq \mathbb{R}^n$ defined in (8.16), and therefore have the same optimal primal value. They arise from two equivalent representations of K_{soc} using different constraint functions.

The constraint function $h_1(x) := \|x^{n-1}\|_2 - x_n$ in (8.53a) is a convex function while the constraint function $h_2(x) := \|x^{n-1}\|_2^2 - x_n^2$ in (8.53b) is nonconvex (Exercise 8.11). If the optimal primal value $f_1^* = f_2^*$ is finite, the Slater Theorem 8.17 applies to problem (8.53a) (the Slater condition is always satisfied) and hence strong duality holds and a dual optimal solution exists. The KKT Theorem 8.15 also applies at at x where $x^{n-1} \neq 0$ and h_1 is continuously differentiable. Since $h_2(x)$ is nonconvex, neither theorem applies to problem (8.53b) even though its feasible set is convex.

Indeed Exercise 8.27 shows that, if $\|c^{n-1}\|_2 \leq c_n$, then strong duality holds and dual optimality is attained for (8.53a) with $f_1^* = f_2^* = 0$, but as long as $0 \neq \|c^{n-1}\|_2 \leq c_n$, $f_2^* = 0 > -\infty = d_2^*$, i.e., the duality gap is unbounded and the dual problem is infeasible for (8.53b).

Hence when we formulate different representations of a convex program:

1 It is important to check that the Slater Theorem 8.17 and the KKT Theorem 8.15 are applicable so that strong duality and optimality conditions hold.
2 If points of nonsmoothness are relevant for the application, nonsmooth analysis studied in Chapter 12 should be used to derive optimality conditions at these points. For (8.53a), $x^* = 0$, where $h_1(x)$ is nondifferentiable, is optimal if and only if $c \in K_{\text{soc}}$ (see Figure 8.15 or (12.61c)). □

8.4 Special convex programs

In this section we apply the general theory developed in Chapter 8.3 to special classes convex optimization problems widely used in applications.

8.4.1 Summary: general method

Consider the convex problem (8.25) reproduced here:

$$f^* := \min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t.} \quad Ax = b,\ h(x) \leq 0 \tag{8.54}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}^l$ are convex functions, and $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$. The classes of problems studied in this section and in Chapter 12.8 using nonsmooth methods are summarized in Figure 8.14 and the conclusions are summarize in Table

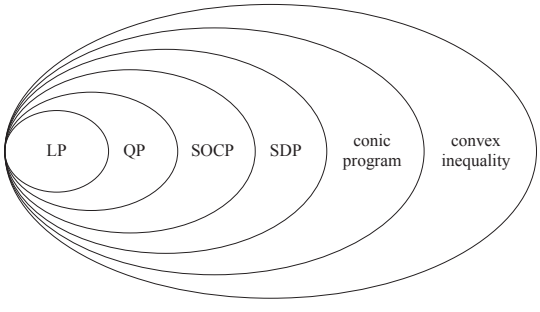


Figure 8.14 Special classes of convex problems studied in this section and Chapter 12.8 using nonsmooth methods.

8.3.

	$f(x)$	$h(x) \leq 0$	sufficient condition	$f^* = d^* = d(\lambda^*, \mu^*)$ KKT, saddle pt
LP	linear	affine: $Bx+d \in \mathbb{R}^l_+$	finite f^*	Th 8.23
QP	quadratic	affine: $Bx+d \in \mathbb{R}^l_+$	feasibility (if $Q \succ 0$)	Th 8.24, 8.25
SOCP	convex	$h(x) \in K_{\text{soc}}$ $h(x) := \tilde{B}x+\tilde{d}$	finite f^*, $A\bar{x}=b$ $h(\bar{x}) \in \text{ri}(K_{\text{soc}})$	Th 8.26, 8.27
SDP	convex	$h(x) \in K_{\text{psd}}$ $h(x) := B_0 + \sum_{i=1}^n x_i B_i$	finite f^*, $A\bar{x}=b$ $h(\bar{x}) \in \text{ri}(K_{\text{psd}})$	Th 8.28
Conic prog.	convex	$h(x) \in K$ $h(x) := Bx+d$	finite f^*, $A\bar{x}=b$ $h(\bar{x}) \in \text{ri}(K)$	Th 12.31, 12.32
Convex prog.	convex	convex	finite f^*, $A\bar{x}=b$ $h(\bar{x}) < 0$	Exercise 12.21

Table 8.3 Summary: strong duality, dual optimality and KKT condition.

The classes in Figure 8.14 differ mainly in the convex constraint $h(x) \leq 0$:

1 Linear program (LP): $f(x) = c^\mathsf{T}x$ and $h(x) \leq 0$ specifies $Bx+d \in \mathbb{R}^l_+ := \{x \in \mathbb{R}^l : x \geq 0\}$, i.e., an affine transformation of x is in the nonnegativity cone.
2 Quadratic program (QP): $f(x) = x^\mathsf{T}Qx + 2cx$ with a positive semidefinite cost matrix Q and an affine constraint $Bx+d \in R^l_+$.[6]
3 Second-order cone program (SOCP): $h(x) \leq 0$ specifies $Bx+d \in K_{\text{soc}} := \{x \in \mathbb{R}^l : \|x^{l-1}\|_2 \leq x_l\}$, i.e., an affine transformation of x is in the second-order cone.
4 Semidefinite program (SDP): $h(x) \leq 0$ specifies $Bx+d \in K_{\text{psd}} \subset \mathbb{S}^l$, i.e., an affine transformation of x is in the semidefinite cone.
5 Conic program: $h(x) \leq 0$ specifies $Bx+d \in K \subseteq \mathbb{R}^l$, i.e., an affine transformation of x is in a closed convex cone K.
6 Convex inequality: $h : \mathbb{R}^n \to \mathbb{R}^l$ is a convex function.

The theory developed in Chapter 8.3 are used to derive three types of results for these convex programs. The general derivation method is as follows. Some of the results in Chapter 8.3 (Saddle-point Theorem 8.14 and primal optimality Theorem 8.16) apply to nonconvex problems as well.

1 *Dual problem*. Given the primal problem (8.54), define the Lagrangian function $L(x,\lambda,\mu) : \mathbb{R}^{n+m+l} \to \mathbb{R}$:

$$L(x,\lambda,\mu) := f(x) - \lambda^\mathsf{T}(Ax-b) + \mu h(x), \qquad x \in \mathbb{R}^n,\ (\lambda,\mu) \in \mathbb{R}^{m+l} \tag{8.55a}$$

[6] Sometimes QP is used to denote problems with a convex quadratic cost f and a general conic constraint $Bx+d \in K$.

Then the dual function is $d(\lambda,\mu) := \min_{x\in\mathbb{R}^n} L(x,\lambda,\mu)$ and the dual problem is

$$d^* := \max_{(\lambda,\mu)\in\mathbb{R}^{m+l}} d(\lambda,\mu) \quad \text{s.t.} \quad \mu \geq 0 \tag{8.55b}$$

2 *Strong duality and dual optimality*. Recall that (i) f and h are convex functions. Suppose (ii) the Slater condition is satisfied, i.e., there exists $\bar{x}$ with $A\bar{x} = b$ and $h(\bar{x}) < 0$, and (iii) the optimal primal value f^* is finite, i.e., $-\infty < f^* < \infty$. Then the Slater Theorem 8.17 implies strong duality and the existence of a dual optimal solution (λ^*,μ^*) with $f^* = d^* = d(\lambda^*,\mu^*)$. This does not guarantee the existence of a primal optimal x^*.

3 *KKT condition and primal optimality*. Recall that (i) f and h are convex functions. Suppose (ii) the Slater condition is satisfied, i.e., there exists $\bar{x}$ with $A\bar{x} = b$ and $h(\bar{x}) < 0$. Then the KKT Theorem 8.15 implies that a feasible $x^* \in \mathbb{R}^n$ is optimal if and only if there exists dual feasible $(\lambda^*,\mu^*) \in \mathbb{R}^{m+l}$ such that

$$\nabla f(x^*) = A^{\mathsf{T}}\lambda^* - \nabla h(x^*)\mu, \qquad \mu^{*\mathsf{T}} h(x^*) = 0, \qquad \mu^* \geq 0 \tag{8.55c}$$

where (only) the first condition is $\nabla_x L(x^*,\lambda^*,\mu^*) = 0$ and requires continuous differentiability of f and h. Such a point (x^*,λ^*,μ^*) is a saddle point that closes the duality gap and attains primal and dual optimality, i.e., $f^* = f(x^*) = d(\lambda^*,\mu^*) = d^*$. Hence the KKT condition can be derived simply by taking the derivative of L with respect to x and it is sufficient for primal-dual optimality when f and h are convex. This method is not applicable if f or h are not continuously differentiable.

Remark 8.10 (Nonsmooth extension)**.** Smoothness (differentiability) of the cost and constraint functions f,h is not important. As long as f,h are convex functions these results hold verbatim at points of differentiability and extends naturally at nondifferentiable points using set-theoretic tools. These tools, developed in Chapter 12, exploit convexity properties, are conceptually simple and can treat a larger class of convex problems (e.g., see Theorem 8.26 and Remark 8.11).

For example for a general conic program whose feasible set is specified, not explicitly by a constraint function $h(x) \leq 0$, but abstractly by a closed convex cone K as $x \in K$, the Lagrangian dual problem is defined by (8.55a)(8.55b), where the penalty term $\mu h(x)$ in $L(x,\lambda,\mu)$ is replaced by μx and dual feasibility $\mu \geq 0$ is replaced by $\mu \in K^*$. Here $K^* := \{\mu \in \mathbb{R}^l : \mu^{\mathsf{T}} z \geq 0 \ \forall z \in K\}$ is called the dual cone of K (ses Chapter 12.1.1). Strong duality and dual optimality hold verbatim. The KKT condition (8.55c) defined only at points where f and h are continuously differentiable can be generalized to a nondifferentiable point using the concept of subgradients $\xi^* \in \partial f(x^*)$ and normal cones (see Chapter 12.8.4). □

In the rest of this section we apply this general method to LP, SOCP and SDP. Referring to Table 8.3, the results on strong duality, dual optimality and the KKT condition for QP are derived in Exercise 8.23 and those for convex problems specified by the convex inequality $h(x) \leq 0$ are derived in Exercise 12.21. General conic programs are studied in Chapter 12.8 using nonsmooth methods.

8.4.2 Linear program (LP)

Consider the linear program:

$$f^* := \min_{x\in\mathbb{R}^n} c^\mathsf{T} x \qquad \text{s.t.} \qquad Ax \geq b \tag{8.56a}$$

where $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. From (8.26) the Lagrangian $L : \mathbb{R}^{n+m} \to \mathbb{R}$ of (8.56) is

$$L(x,\mu) := \left(c - A^\mathsf{T}\mu\right)^\mathsf{T} x + b^\mathsf{T}\mu \qquad x \in \mathbb{R}^n,\ \mu \in \mathbb{R}^m$$

the dual function is

$$d(\mu) := \min_{x\in\mathbb{R}^n} L(x,\mu) = \begin{cases} b^\mathsf{T}\mu & \text{if} \quad A^\mathsf{T}\mu = c \\ -\infty & \text{if} \quad A^\mathsf{T}\mu \neq c \end{cases}$$

and the dual problem is

$$d^* := \max_{\mu\geq 0} d(\mu) = \max_{\mu\geq 0} b^\mathsf{T}\mu \qquad \text{s.t.} \qquad A^\mathsf{T}\mu = c \tag{8.56b}$$

Let $X := \{x \in \mathbb{R}^n : Ax \geq b\}$ and $Y := \{\mu \in \mathbb{R}^m : A^\mathsf{T}\mu = c, \mu \geq 0\}$ be the feasible sets.

The primal and dual problems in (8.56) can each be finite feasible, feasible but unbounded, or infeasible. By definition the primal problem is feasible if $f^* < \infty$ and the dual problem is feasible if $d^* > -\infty$. Strong duality of LP implies that only four, instead of nine, scenarios are possible (see Table 8.4 and its caption). Moreover a feasible solution (x^*, μ^*) is optimal if and only if it satisfies complementary slackness, as we will show. We start by stating in the next lemma that a finite f^* (feasibility is insufficient) implies the existence of a primal optimal solution $x^* \in X$ with $f(x^*) = f^*$; indeed a finite f^* also implies the existence of dual optimal μ^* and strong duality (Theorem 8.23). Lemma 8.22 applies to the dual problem (8.56b) if d^* is finite.

Lemma 8.22 (LP primal optimality)**.** Consider the linear program (8.56a). If $-\infty < f^* < \infty$ then an optimal solution $x^* \in X$ exists with $c^\mathsf{T}x^* = f^*$.

Proof Let $X := \{x \in \mathbb{R}^n : Ax \geq b\}$ be the feasible set of (8.56a). Since f^* is finite, X is nonempty and closed. If the feasible set X is bounded or if there is a $\gamma \in \mathbb{R}$ such that the level set V_γ is nonempty and bounded, then $X \cap V_\gamma$ is a compact (closed and bounded) set. The minimization (8.56a) can be taken over $X \cap V_\gamma$ and a minimizer x^* therefore exists by Theorem 8.16.

Consider then the case where X is unbounded and every nonempty level set $V_\gamma := \{x \in \mathbb{R}^n : c^\mathsf{T}x \leq \gamma\}$ is unbounded. Let $\{V_{\gamma_k}\}$ be a nested sequence of level sets with $\gamma_k \downarrow f^*$. The set of solutions of (8.56a) is $X^* := \cap_{k=1}^\infty \left(X \cap V_{\gamma_k}\right)$. The finiteness of f^* means that $X \cap V_{\gamma_k} \neq \emptyset$ for each k. Moreover $X \cap V_{\gamma_k}$ is polyhedral for each k. Then $X^* \neq \emptyset$ follows from the following fact (see e.g. [54, Props. 1.4.9, 1.4.10, pp.58–61] for a proof) which underlies the simplicity of linear programs.
Fact. Consider a sequence $\{C_k\}$ of nonempty sets C_k.

1 The intersection $\cap_{k=1}^{\infty} C_k \neq \emptyset$ if and only if there is a sequence $\{x_k\}$ that is bounded where $x_k \in C_k$, i.e., there is r with $\|x_k\| \leq r$ for all k.
2 If $\{C_k\}$ are polyhedral (which is the case for linear programs), then $\cap_{k=1}^{\infty} C_k \neq \emptyset$.

□

The next theorem is the main result on LP duality and optimality. Though the proof below appeals to the Slater Theorem 8.17, it can also be proved directly using the Farkas Lemma (Theorem 8.12); see Exercise 8.22.

Theorem 8.23 (LP duality and KKT)**.** Consider the linear program and its dual (8.56).

1 *Strong duality and primal-dual optimality*. Exactly one of the following holds:
   1 If $-\infty < f^* < \infty$ or $-\infty < d^* < \infty$ then both primal and dual problems attain optimality and strong duality holds, i.e., there exists $(x^*, \mu^*) \in X \times Y$ such that
   $$c^{\mathsf{T}} x^* \;=\; f^* \;=\; d^* \;=\; b^{\mathsf{T}} \mu^*$$
   2 If the primal problem is feasible but unbounded then $f^* = -\infty = d^*$, i.e., the dual problem is infeasible.
   3 If the dual problem is feasible but unbounded then $d^* = \infty = f^*$, i.e., the primal problem is infeasible.
   4 Otherwise, both are infeasible, i.e., $f^* = \infty$ and $d^* = -\infty$.
2 *KKT characterization*. A feasible $x^* \in X$ is optimal if and only if there is a dual feasible $\mu^* \in Y$ that satisfies complementary slackness, i.e.,
   $$A^{\mathsf{T}} \mu^* \;=\; c, \qquad \mu^{*\mathsf{T}} (Ax^* - b) \;=\; 0, \qquad \mu^* \;\geq\; 0 \tag{8.57}$$
   Such a point (x^*, μ^*) is a saddle point and a KKT point and is hence primal-dual optimal with $c^{\mathsf{T}} x^* = b^{\mathsf{T}} \mu^*$.

Proof Suppose f^* is finite (If d^* is finite, Lemma 8.22 applies to the dual problem (8.56b) and the argument below is symmetric and omitted). Then Lemma 8.22 implies the existence of a primal optimal solution $x^* \in X$ with $c^{\mathsf{T}} x^* = f^*$. This also implies that the Slater condition (8.45) is satisfied. The Slater Theorem 8.17 then implies that there exists a dual optimal solution $\mu^* \in Y$ such that $f^* = d^* = d(\mu^*)$. (For linear programs, this step can be proved using the Farkas Lemma (Theorem 8.12); see Exercise 8.22.)

If $f^* = -\infty$ then weak duality Lemma 8.13 implies that $d^* \leq f^* = -\infty$. Similarly if $d^* = \infty$ then $f^* = \infty$ by weak duality. The only case that is not covered by the three cases above is when both $f^* = \infty$ and $d^* = -\infty$. This is possible as Example 8.14 shows.

Finally given any primal feasible point $x^* \in X$ and any $\mu^* \geq 0$, we need to show that (x^*, μ^*) is primal-dual optimal if and only if (x^*, μ^*) satisfies (8.57). Suppose (x^*, μ^*) satisfies (8.57). Then $\mu^* \in Y$ and
$$b^{\mathsf{T}} \mu^* \;=\; c^{\mathsf{T}} x^* - \mu^{*\mathsf{T}} (Ax^* - b) \;\leq\; c^{\mathsf{T}} x^* \tag{8.58}$$

where the first equality follows from $\mu^* \in Y$ and the inequality follows from $(x^*, \mu^*) \in X \times Y$. Moreover the complementary slackness in (8.57) implies that equality is attained in (8.58), i.e., $b^\mathsf{T}\mu^* = L(x^*, \mu^*) = c^\mathsf{T}x^*$. The weak duality Lemma 8.13 then implies that (x^*, μ^*) is primal-dual optimal and closes the duality gap. Conversely suppose $(x^*, \mu^*) \in X \times Y$ is primal-dual optimal. Then both $f^* = f(x^*)$ and $d^* = d(\mu^*)$ are finite and therefore by part 1, strong duality holds, i.e., $b^\mathsf{T}\mu^* = c^\mathsf{T}x^*$. This and (8.58) then imply $\mu^{*\mathsf{T}}(Ax^* - b)$ and therefore (x^*, μ^*) satisfies (8.57). Such a point is a saddle-point and a KKT point according to Theorem 8.15. □

Example 8.13 (Equality and nonnegativity constraints)**.** Adapt Theorem 8.23 to linear program of the form:

1 $f^* := \min_{x \in \mathbb{R}^n} c^\mathsf{T}x$ s.t. $Ax = b$, $x \geq 0$ where $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$.
2 $f^* := \min_{x \in \mathbb{R}^n} c^\mathsf{T}x$ s.t. $Ax = b$, $Bx + d \geq 0$ where $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $B \in \mathbb{R}^{l \times n}$ and $d \in \mathbb{R}^l$.

Solution. For part 1 we will show that the condition for strong duality and primal-dual optimality remains that same as in Theorem 8.23 but the KKT condition (8.57) is modified to

$$A^\mathsf{T}\lambda^* + \mu^* = c, \qquad \mu^{*\mathsf{T}}x^* = 0, \qquad \mu^* \geq 0 \tag{8.59}$$

The Lagrangian $L : \mathbb{R}^{2n+m} \to \mathbb{R}$ is

$$L(x, \lambda, \mu) := \Big(c - A^\mathsf{T}\lambda - \mu\Big)^\mathsf{T} x + b^\mathsf{T}\lambda \qquad x \in \mathbb{R}^n,\ \lambda \in \mathbb{R}^m,\ \mu \in \mathbb{R}^n$$

the dual function is

$$d(\lambda, \mu) := \min_{x \in \mathbb{R}^n} L(x, \lambda, \mu) = \begin{cases} b^\mathsf{T}\lambda & \text{if} \quad A^\mathsf{T}\lambda + \mu = c \\ -\infty & \text{if} \quad A^\mathsf{T}\lambda + \mu \neq c \end{cases}$$

and the dual problem is

$$d^* = \max_{\lambda \in \mathbb{R}^m, \mu \geq 0} b^\mathsf{T}\lambda \qquad \text{s.t.} \qquad A^\mathsf{T}\lambda + \mu = c$$

Let $X := \{x \in \mathbb{R}^n : Ax = b, x \geq 0\}$ and $Y := \{(\lambda, \mu) \in \mathbb{R}^{m+n} : A^\mathsf{T}\lambda + \mu = c, \mu \geq 0\}$ be the feasible sets. All the structural results of Theorem 8.23 holds. The only change is that (8.58) becomes, since $Ax^* = b$,

$$b^\mathsf{T}\mu^* = c^\mathsf{T}x^* - \mu^{*\mathsf{T}}x^* \leq c^\mathsf{T}x^*$$

and hence a feasible $x^* \in X$ is optimal if and only if there exists a dual optimal $(\lambda^*, \mu^*) \in \mathbb{R}^{m+n}$ that satisfies (8.59).

Part 2 can be converted to the problem in part 1 by introducing the slack variable $s \in \mathbb{R}^l$: $f^* := \min_{(x,s) \in \mathbb{R}^{n+l}} c^\mathsf{T}x$ s.t. $Ax = b$, $Bx + d - s = 0$, $s \geq 0$. Then (8.59) becomes

$$A^\mathsf{T}\lambda^* + B^\mathsf{T}\mu^* = c, \qquad \mu^{*\mathsf{T}}s^* = 0, \qquad \mu^* \geq 0$$

□

As Theorem 8.23 shows, weak and strong duality imply that only 4 feasibility cases are possible for the primal and dual problems, instead of 9 cases, as explained in Table 8.4 and its caption. The only case where the optimal values are attained at finite (x^*, λ^*, μ^*) is when both problems are bounded feasible.

		primal		
		bounded feasible	unbounded feasible	infeasible
dual	bounded feasible	(x^*, λ^*, μ^*)	$\times$ (sd)	$\times$ (sd)
	unbounded feasible	$\times$ (sd)	$\times$ (wd)	$f^* = d^* = \infty$
	infeasible	$\times$ (sd)	$f^* = d^* = -\infty$	$d^* = -\infty < \infty = f^*$

Table 8.4 Four possibilities: Strong duality in Theorem 8.23 excludes 4 possibilities labeled "×(sd)". The 5th impossibility, labeled "×(wd)", violates weak duality. Optimal values are attained only in one case.

Example 8.14 (LPs with infinite values)**.** 1 *Infeasible LP pair*. Consider the LP $\min_x x$ such that $\begin{bmatrix} 1 \\ -1 \end{bmatrix} x \geq \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Its dual is $\max_{\mu \geq 0} \mu_2$ such that $-\mu_2 = 1$. Clearly neither the primal nor the dual is feasible and hence $d^* = -\infty < \infty = f^*$.

2 *Unbounded primal, infeasible dual*. Consider

$$f^* \quad := \quad \min_{x \geq 0} \; -x_1 + \alpha x_2 \qquad \text{s.t.} \qquad x_1 - x_2 = 0$$

where $\alpha < 1$. Then the optimal primal value is $f^* = -\infty$ and there is no finite x that attains it. From Example (8.13) the dual function is

$$d(\lambda, \mu) := \begin{cases} 0 & \text{if } \begin{bmatrix} -1 \\ \alpha \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \lambda + \mu \\ -\infty & \text{otherwise} \end{cases}$$

Multiplying both sides of the equality constraint by $(1, 1)$ yields $\mu_1 + \mu_2 = -(1 - \alpha) < 0$. Hence there is no (λ, μ) that satisfies $\mu \geq 0$. Therefore the dual problem is infeasible, or $d^* := -\infty = f^*$. □

Optimal basic feasible solution.

The first widely used algorithm for solving linear programs is the simplex algorithm which makes use of the fact that if a LP has a *finite* optimal solution x^* then it has an optimal solution that is an extreme point (vertex) of the feasible set $X := \{x \in \mathbb{R}^n : Ax \geq b\}$ where $A \in \mathbb{R}^{m \times n}$. A finite f^* implies $m \geq n$. For each feasible point $x \in X$ let $\hat{I}(x) := \{i \in \{1, \ldots, m\} : a_i^\mathsf{T} x = b_i\}$ be the set of all active constraints at x, where a_i^T is the ith row of A. A feasible x is an extreme point of X if and only if $\hat{I}(x)$ contains $n \leq m$ linearly independent active constraints at x, i.e., $\{a_i : i \in \hat{I}(x)\}$ contains n linearly independent a_i. In the simplex algorithm literature a feasible extreme point is called a *basic feasible solution* and an optimal extreme point is called an *optimal basic feasible solution*. The simplex algorithm starts from a basic feasible solution

and moves to another basic feasible solution with a lower cost until an optimal basic feasible solution is found. Even though the simplex algorithm is usually replaced by interior point methods in modern LP solvers, it reveals the following useful structure of an optimal basic feasible solution x^* of linear programs.

For each extreme point x of X let $I(x) \subseteq \hat{I}(x)$ denote any collection of n linearly independent constraints, i.e., $\{a_i : i \in I(x)\}$ is a set of n linearly independent vectors. Decompose (A, b) according to $I = I(x)$:

$$A =: \begin{bmatrix} A_{I(x)} \\ A_{-I(x)} \end{bmatrix}, \qquad b =: \begin{bmatrix} b_{I(x)} \\ b_{-I(x)} \end{bmatrix}$$

so that $A_{I(x)}x = b_{I(x)}$ and $A_{-I(x)}x > b_{-I(x)}$. Then $A_{I(x)}$ is a $n \times n$ nonsingular matrix whose columns form a basis of $\mathbb{R}^n$. Hence an optimal basic feasible solution (extreme point) x^* of the linear program (8.56) satisfies

$$x^* = A_{I(x^*)}^{-1} b_{I(x^*)} \tag{8.60}$$

In Exercise 8.22 the set $I(x)$ (or $\hat{I}(x)$) is used to construct an optimal dual variable μ^*. The basic idea is to use the Farkas lemma to show that $c \in \text{cone}\left(A_{I(x^*)}^{\mathsf{T}}\right)$ and hence $c = A_{I(x^*)}^{\mathsf{T}} \mu_{I(x^*)}^*$ for some $\mu_{I(x^*)}^* \geq 0$.

8.4.3 Convex quadratic program (QP)

A quadratic program (QP) has a quadratic cost function and affine constraints and a quadratically constrained quadratic program (QCQP) has a quadratic cost function and quadratic constraints. In this subsection we study QPs that are convex.

Convex quadratic program (QP).

Consider first an unconstrained convex quadratic program:

$$f_1^* := \min_{x \in \mathbb{R}^n} f(x) := x^{\mathsf{T}} Q x + 2c^{\mathsf{T}} x \tag{8.61}$$

where $Q \in \mathbb{R}^{n \times n}$ is positive semidefinite, i.e., $Q \succeq 0$, and $c \in \mathbb{R}^n$. The cost function f is convex if and only if $Q \succeq 0$. Since Q is positive semidefinite it has a spectral decomposition

$$Q = U \Lambda U^{\mathsf{T}} = \begin{bmatrix} U_r & U_{n-r} \end{bmatrix} \begin{bmatrix} \Lambda_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_r^{\mathsf{T}} \\ U_{n-r}^{\mathsf{T}} \end{bmatrix} = U_r \Lambda_r U_r^{\mathsf{T}} \tag{8.62a}$$

where r is the rank of Q, Λ_r is a diagonal (sub)matrix of the r positive eigenvalues of Q and the columns of $U_r \in \mathbb{R}^{n \times r}$ are the corresponding $r \leq n$ (real) orthonormal eigenvectors. The columns of $U_{n-r} \in \mathbb{R}^{n \times (n-r)}$ are $n-r$ orthonormal (real) eigenvectors corresponding to the 0 eigenvalue, if any. The matrix Q is positive definite if $r = n$ and

positive semidefinite but not positive definite if $r < n$. The range space, null space and the pseudo-inverse $Q^\dagger$ of Q are respectively:

$$\text{range}(Q) = \text{span}(U_r), \;\; \text{null}(Q) = \text{span}(U_{n-r}), \;\; Q^\dagger := U_r \Lambda_r^{-1} U_r^\mathsf{T}, \quad r \le n \tag{8.62b}$$

because $U_r^\mathsf{T} U_{n-r} = 0$ (see Chapter A.7 on pseudo-inverse and Theorem A.16 on orthogonal diagonalization for psd matrices). If $r = n$ then $Q^\dagger = Q^{-1}$. Unconstrained convex QP can be solved explicitly, as stated below and proved in Exercise 8.23.

Theorem 8.24 (Unconstrained convex QP)**.** Consider the unconstrained QP (8.61).

1 If $c \in \text{range}(Q)$ then a minimizer x^* and the minimal value f_1^* are respectively:

$$x^* \;=\; -Q^\dagger c, \qquad\qquad f_1^* = -c^\mathsf{T} Q^\dagger c$$

where $Q^\dagger$ is the pseudo-inverse of Q defined in (8.62b). Moreover the set of minimizer is $x^* = -Q^\dagger c + \text{null}(Q)$.

2 If $c \notin \text{range}(Q)$ then $f_1^* = -\infty$.

3 If $Q \succ 0$ is positive definite then the unique minimizer x^* and the minimum value f_1^* are respectively:

$$x^* \;=\; -Q^{-1} c, \qquad\qquad f_1^* = -c^\mathsf{T} Q^{-1} c$$

In particular $\text{range}(Q) = \mathbb{R}^n$ and $Q^\dagger = Q^{-1}$.

Consider next an affinely constrained version of (8.61):

$$f_2^* \;:=\; \min_{x \in \mathbb{R}^n} f(x) := x^\mathsf{T} Q x \,+\, 2c^\mathsf{T} x \quad \text{s.t.} \quad Ax = b,\; Bx + d \ge 0 \tag{8.63}$$

where $Q \succeq 0$, $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $B \in \mathbb{R}^{l \times n}$ and $d \in \mathbb{R}^l$. The quadratic program (8.63) reduces to a linear program if $Q = 0$. We next state strong duality and the KKT condition for (8.63) when $Q \succ 0$ is positive definite. The result is proved in Exercise 8.24 for the more general case when $Q \succeq 0$. When $Q \succ 0$ let

$$\hat{Q} := \begin{bmatrix} A \\ B \end{bmatrix} Q^{-1} \begin{bmatrix} A^\mathsf{T} & B^\mathsf{T} \end{bmatrix}, \qquad\qquad \hat{c} := \begin{bmatrix} -b \\ d \end{bmatrix} - \begin{bmatrix} A \\ B \end{bmatrix} Q^{-1} c \tag{8.64}$$

Theorem 8.25 (Constrained convex QP)**.** Suppose the QP (8.63) is feasible and $Q \succ 0$.

1 *Dual problem*. The dual problem is

$$d^* \;:=\; -c^\mathsf{T} Q^{-1} c \,-\, \min_{\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^l_+} \left(\begin{bmatrix} \lambda^\mathsf{T} & \mu^\mathsf{T} \end{bmatrix} \hat{Q} \begin{bmatrix} \lambda \\ \mu \end{bmatrix} + 2\,\hat{c}^\mathsf{T} \begin{bmatrix} \lambda \\ \mu \end{bmatrix} \right)$$

where $\mathbb{R}^l_+ := \{\mu \in \mathbb{R}^l : \mu \ge 0\}$.

2 *Strong duality, dual optimality, KKT condition*. Strong duality holds and dual optimality is attained. Moreover a feasible x^* is optimal if and only if there exists $(\lambda^*, \mu^*) \in \mathbb{R}^{m+l}$ such that $\mu^* \ge 0$ and

$$x^* \;=\; Q^{-1}(A^\mathsf{T}\lambda^* + B^\mathsf{T}\mu^* - c), \qquad\qquad \mu^{*\mathsf{T}}(Bx^* + d) \;=\; 0 \tag{8.65}$$

Such a point is a saddle point and a KKT point that is primal-dual optimal and closes the duality gap, i.e., $f_2^* = f(x^*) = d(\lambda^*, \mu^*) = d^*$.

Exercise 8.25 studies the following convex quadratically constrained quadratic program (QCQP):

$$f^* \;:=\; \min_{x\in\mathbb{R}^n} f(x) \;:=\; x^\mathsf{T} Q_0 x \,+\, 2c_0^\mathsf{T} x \quad \text{s.t.} \quad x^\mathsf{T} Q_1 x \,+\, 2c_1^\mathsf{T} x \le d$$

where $Q_0 \succ 0$ is positive definite, $Q_1 \succeq 0$ is positive semidefinite, $c_0, c_1 \in \mathbb{R}^n$ and $d \in \mathbb{R}$. It shows that the dual problem is:

$$d^* \;:=\; -\min_{\mu\in\mathbb{R}_+} d\mu + (c_0+\mu c_1)^\mathsf{T}(Q_0+\mu Q_1)^{-1}(c_0+\mu c_1)$$

strong duality holds and dual optimality is attained if f^* is finite and there exists $\bar{x}$ such that $\bar{x}^\mathsf{T} Q_1 \bar{x} \,+\, 2c_1^\mathsf{T}\bar{x} < d$. In that case a feasible x^* is optimal if and only if there exists $\mu^* \in \mathbb{R}$ such that $\mu^* \ge 0$ and

$$(Q_0+\mu^* Q_1)x^* + (c_0+\mu^* c_1) \;=\; 0, \qquad \mu^*(x^{*\mathsf{T}} Q_1 x^* \,+\, 2c_1^\mathsf{T} x^* - d) \;=\; 0$$

8.4.4 Second-order cone program (SOCP)

A second-order cone program (SOCP) is a convex optimization problem where either the variable x or its affine transformation $\tilde{B}x + \tilde{d}$ is in the standard second-order cone $K_{\text{soc}} := \{x \in \mathbb{R}^n : \|x^{n-1}\|_2 \;\le\; x_n\}$ defined in (8.16),

Second-order cone.

Consider the convex optimization problem:

$$f^* \;:=\; \min_{x\in\mathbb{R}^n} f(x) \;\text{ s.t. }\; Ax = b,\; x \in K_{\text{soc}} \tag{8.66a}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a real-valued convex function, $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, and $K_{\text{soc}} \subseteq \mathbb{R}^n$ is the standard second-order cone defined in (8.16), reproduced here ($x^k := (x_1, \cdots, x_k)$ denotes the vector consisting of the first k entries of x),

$$K_{\text{soc}} \;:=\; \{x \in \mathbb{R}^n : \|x^{n-1}\|_2 \;\le\; x_n\} \tag{8.66b}$$

This problem is called a *second-order cone program (SOCP)*. It reduces to a linear program (8.56a) if K_{soc} is polyhedral (e.g., $K_{\text{soc}} = \{x \in \mathbb{R}^n : x \ge 0\}$) and f is linear. In this chapter we assume f is continuously differentiable though this is not important (see the extension to nonsmooth convex setting in Chapter 12.8.3). Let $h(x) := \|x^{n-1}\|_2 - x_n$. Then $h(x)$ is convex, differentiable if and only if $x^{n-1} \neq 0$, and $x \in K_{\text{soc}}$ is equivalent to $h(x) \le 0$.

To derive the dual problem of (8.66) and the KKT condition, let the Lagrangian function $L : \mathbb{R}^{n+m+1} \to \mathbb{R}$ be

$$L(x,\lambda,\mu) \;:=\; f(x) - \lambda^\mathsf{T}(Ax-b) + \mu\left(\|x^{n-1}\|_2 - x_n\right), \qquad x \in \mathbb{R}^n,\; \lambda \in \mathbb{R}^m, \mu \in \mathbb{R}$$

Then the dual function is $d(\lambda,\mu) := \min_{x\in\mathbb{R}^n} L(x,\lambda,\mu)$ and the dual problem is

$$d^* := \max_{\lambda,\mu\geq 0} d(\lambda,\mu) \tag{8.66c}$$

Let $X := \left\{x \in \mathbb{R}^n : Ax = b, \|x^{n-1}\|_2 \leq x_n\right\}$ and $Y := \left\{(\lambda,\mu) \in \mathbb{R}^{m+1} : \mu \geq 0\right\}$ be the feasible sets.

Theorem 8.26 (SOCP duality and KKT)**.** Consider the SOCP and its dual (8.66).

1 *Strong duality and dual optimality*. Suppose f^* is finite, and there exists $\bar{x}$ such that $A\bar{x} = b$ and $\|\bar{x}^{n-1}\|_2 < \bar{x}_n$. Then there exists a dual optimal solution $(\lambda^*,\mu^*) \in Y$ that closes the duality gap, i.e., $f^* = d^* = d(\lambda^*,\mu^*)$.
2 *KKT characterization*. A primal and dual feasible point $(x^*,\lambda^*,\mu^*) \in X\times Y$ with $[x^*]^{n-1} \neq 0$ is primal-dual optimal and closes the duality gap if and only if and

$$\nabla f(x^*) = A^\mathsf{T}\lambda^* + \mu^* \begin{bmatrix} -[x^*]^{n-1} \\ \|[x^*]^{n-1}\|_2 \end{bmatrix}, \qquad \mu^*\left(\|[x^*]^{n-1}\|_2 - x_n^*\right) = 0 \tag{8.67}$$

Such a point (x^*,λ^*,μ^*) is a saddle point and a KKT point.

Proof Part 1 follows from the Slater Theorem 8.17 since the constraint function $h(x) := \|x^{n-1}\|_2 - x_n$ in (8.66b) is convex (and differentiable if and only if $x^{n-1} \neq 0$). Part 2 follows from the KKT Theorem 8.15 because (8.67) in the theorem are the stationarity condition $\nabla_x L(x^*,\lambda^*,\mu^*) = 0$ and the complementary slackness condition. □

Remark 8.11. 1 *Primal optimality*. Unlike for a linear program, a finite f^* and the Slater condition do not guarantee that the optimal value f^* is attained at a finite x^*. In particular, even when a dual optimal solution exists that closes the duality gap under the Slater condition, there may not be any feasible x^* that satisfies the KKT condition; see Examples 8.9 and 8.10.
2 *KKT under Slater condition.* If we assume the Slater condition, i.e., there exists $\bar{x}$ with $A\bar{x} = b$ and $\|\bar{x}^{n-1}\|_2 < \bar{x}_n$, then the KKT characterization in Theorem 8.26 can be strengthened to: a feasible $x^* \in X$ is optimal if and only if there exist $(\lambda^*,\mu^*) \in Y$ such that (8.67) holds. Without the Slater condition, the existence of a primal optimal x^* (and hence finite f^*) does not guarantee the existence of a dual optimal (λ^*,μ^*). □

The condition $[x^*]^{n-1} \neq 0$ in Theorem 8.26 is needed because the constraint function $h(x) := \|x^{n-1}\|_2 - x_n$ is nondifferentiable if $x^{n-1} = 0$. The differentiability of the cost and constraint functions is however unimportant as long as these functions are convex. When $[x^*]^{n-1} = 0$, the KKT condition requires the Slater condition that there exists $\bar{x}$ such that $A\bar{x} = b$ and $\|\bar{x}^{n-1}\|_2 < \bar{x}_n$. Then

1 Case $x_n^* > \|[x^*]^{n-1}\|_2 = 0$: x^* is optimal if and only if there exists $\lambda^* \in \mathbb{R}^m$ such that $\nabla f(x^*) = A^\mathsf{T}\lambda^*$.

2 Case $x^* = 0$: $x^* = 0$ is optimal if and only if there exist $\lambda^* \in \mathbb{R}^m$ and $\eta^* \in K_{\text{soc}}$ such that $\nabla f(0) = A^\mathsf{T}\lambda^* + \eta^*$.

This is derived in Chapter 12.8.3 using techniques for nonsmooth analysis (see (12.61)). Figure 8.15 illustrates why the optimality condition does not depend on differentiability.

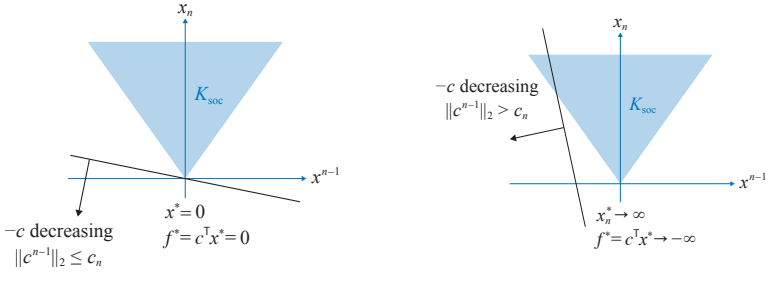


Figure 8.15 Theorem 8.26: optimality condition at $x^* = 0$ where $h(x)$ is nondifferentiable with $f(x) := c^\mathsf{T}x$ and without $Ax = b$. (a) $c \in K_{\text{soc}}$: $x^* = 0$ and $f^* = 0$. (b) $c \notin K_{\text{soc}}$: $f^* = -\infty$.

SOC constraint.

Consider the convex optimization

$$f^* \;:=\; \min_{x \in \mathbb{R}^n} f(x) \;\text{ s.t. }\; Ax = b,\; \|Bx + d\|_2 \le \beta^\mathsf{T}x + \delta \tag{8.68}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a real-valued convex continuously differentiable function, $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$, $B \in \mathbb{R}^{(l-1)\times n}$, $d \in \mathbb{R}^{l-1}$, $\beta \in \mathbb{R}^n$ and $\delta \in \mathbb{R}$. The constraint $\|Bx + d\|_2 \le \beta^\mathsf{T}x + \delta$ is the second-order cone constraint studied in Chapter 8.2.1. The problem (8.68) is also called a second-order cone program (SOCP) because the quadratic constraint says that an affine transformation of x lies in the second-order cone K_{soc}. It reduces to a linear program when $B = 0$ or $l = 1$ and $f(x)$ is linear. It subsumes (8.66) as a special case. The assumption that f is continuously differentiable is relaxed in Chapter 12.8.3.

To derive the dual problem of (8.68) and the KKT condition, we reduce it to the case of (8.66) with an auxiliary variable z and an additional equality constraint. Consider the equivalent problem:

$$f^* \;:=\; \min_{(x,z) \in \mathbb{R}^{n+l}} f(x) \qquad \text{s.t.} \qquad Ax = b,\; z = \tilde{B}x + \tilde{d},\; z \in K_{\text{soc}} \tag{8.69a}$$

where K_{soc} is the second-order cone defined in (8.66b) and

$$\tilde{B} \;:=\; \begin{bmatrix} B \\ \beta^\mathsf{T} \end{bmatrix}, \qquad \tilde{d} \;:=\; \begin{bmatrix} d \\ \delta \end{bmatrix} \tag{8.69b}$$

The Lagrangian $L : \mathbb{R}^{n+l} \times \mathbb{R}^{m+l+1} \to \mathbb{R}$ is: for $x \in \mathbb{R}^n, z \in \mathbb{R}^l, \lambda \in \mathbb{R}^m, \gamma \in \mathbb{R}^l, \mu \in \mathbb{R}$,

$$L(x,z,\lambda,\gamma,\mu) \;:=\; f(x) - \lambda^\mathsf{T}(Ax-b) - \gamma^\mathsf{T}(\tilde{B}x+\tilde{d}-z) + \mu\left(\|z^{l-1}\|_2 - z_l\right)$$

The dual problem is (Exercise 8.26):

$$d^* \quad := \quad \max_{\lambda,\gamma} \left(b^\mathsf{T}\lambda - \tilde{d}^\mathsf{T}\gamma\right) + d_0(\lambda,\gamma) \qquad \text{s.t.} \qquad \gamma \in K_{\text{soc}} \tag{8.69c}$$

where

$$d_0(\lambda,\gamma) \;:=\; \min_{x\in\mathbb{R}^n} \left(f(x) - (A^\mathsf{T}\lambda + \tilde{B}^\mathsf{T}\gamma)^\mathsf{T}x\right) \tag{8.69d}$$

For example when the cost function in (8.69a) is linear $c^\mathsf{T}x$ the dual problem is:

$$d^* \quad := \quad \max_{(\lambda,\gamma)\in\mathbb{R}^{m+l}} \; b^\mathsf{T}\lambda - \tilde{d}^\mathsf{T}\gamma \qquad \text{s.t.} \qquad A^\mathsf{T}\lambda + \tilde{B}^\mathsf{T}\gamma = c, \;\; \|\gamma^{l-1}\|_2 \le \gamma_l$$

Let $X := \left\{x \in \mathbb{R}^n : Ax = b, \|Bx+d\|_2 \le \beta^\mathsf{T}x + \delta\right\}$ and $Y := \left\{(\lambda,\mu) \in \mathbb{R}^{m+1} : \mu \ge 0\right\}$. Note that $X \times Y$ does not contain the auxiliary variable z and the corresponding dual variable γ. Even though the dual problem does not depend on μ, the complementary slackness in the KKT condition does.

Theorem 8.27 (SOCP duality and KKT)**.** Consider the SOCP and its dual (8.69). Suppose there exists $\bar{x}$ such that $A\bar{x} = b$ and $\|B\bar{x}+d\|_2 < \beta^\mathsf{T}\bar{x} + \delta$ so that the Slater condition (8.45) is satisfied.

1 *Strong duality and dual optimality*. Suppose f^* is finite. Then there exists a dual optimal solution $(\lambda^*,\gamma^*,\mu^*)$ that closes the duality gap, i.e., $f^* = d^* = d(\lambda^*,\gamma^*,\mu^*)$.
2 *KKT characterization*. A point $x^* \in X$ with $Bx^* + d \neq 0$ is optimal if and only if there exist $(\lambda^*,\mu^*) \in Y$ such that

$$\begin{aligned} \nabla f(x^*) \;&=\; A^\mathsf{T}\lambda^* + \mu^*\left(-B^\mathsf{T}(Bx^*+d) + \beta(\beta^\mathsf{T}x^* + \delta)\right) \\ 0 \;&=\; \mu^*\left(\|Bx^*+d\|_2 - (\beta^\mathsf{T}x^* + \delta)\right) \end{aligned}$$

Such a point (x^*,λ^*,μ^*), together with $z^* := \tilde{B}x^* + \tilde{d}$ and $\gamma^* = \mu^*\begin{bmatrix} -[z^*]^{l-1} \\ \|[z^*]^{l-1}\|_2 \end{bmatrix} \in K_{\text{soc}}$, is a saddle point and a KKT point for (8.69).

Proof If there exists an $\bar{x}$ such that $A\bar{x} = b$ and $\|B\bar{x}+d\|_2 < \beta^\mathsf{T}\bar{x} + \delta$ then there exists a $\bar{z}$ such that $\bar{z} = \tilde{B}\bar{x} + \tilde{d}$ and $\|\bar{z}^{l-1}\|_2 < \bar{z}_l$. This is the Slater condition for (8.69a) and hence part 1 follows from Theorem 8.26.

For part 2 we derive the stationarity condition $\nabla_x L(x^*,z^*,\lambda^*,\gamma^*,\mu^*) = 0$ and $\nabla_z L(x^*,z^*,\lambda^*,\gamma^*,\mu^*) = 0$ as well as the complementary slackness condition in the KKT Theorem 8.15. When $z^{l-1} \neq 0$ we have

$$\nabla_x L(x,z,\lambda,\gamma,\mu) \;=\; \nabla f(x) - A^\mathsf{T}\lambda - \tilde{B}^\mathsf{T}\gamma, \quad \nabla_z L(x,z,\lambda,\gamma,\mu) \;=\; \gamma + \mu\begin{bmatrix} \frac{z^{l-1}}{\|z^{l-1}\|_2} \\ -1 \end{bmatrix}$$

Hence the KKT condition in terms of (x^*, z^*) and $(\lambda^*, \gamma^*, \mu^*)$ is:

$$\nabla f(x^*) \;=\; A^{\mathsf{T}}\lambda^* + \tilde{B}^{\mathsf{T}}\gamma^*, \quad \gamma^* \;=\; \mu^* \begin{bmatrix} \frac{-[z^*]^{l-1}}{\|[z^*]^{l-1}\|_2} \\ 1 \end{bmatrix}, \quad \mu^* \left(\|[z^*]^{l-1}\|_2 - z_l^* \right) \;=\; 0$$

Eliminating z^* and γ^* yields the KKT condition in the theorem. The remaining claim follows from the KKT Theorem 8.15. □

As in Theorem 8.26, the condition $Bx^* + d \neq 0$ is needed because the constraint function $h(z) := \|z^{l-1}\|_2 - z_l$ is nondifferentiable if $z^{l-1} = Bx + d = 0$. When $Bx^* + d = 0$, the KKT condition requires the Slater condition in the theorem that there exists $\bar{x}$ such that $A\bar{x} = b$ and $\|B\bar{x} + d\|_2 < \beta^{\mathsf{T}}\bar{x} + \delta$. Then a point $x^* \in X$ with $Bx^* + d = 0$ is optimal if and only if

1 Case $\beta^{\mathsf{T}}x^* + \delta > 0$: there exists $\lambda^* \in \mathbb{R}^m$ such that $\nabla f(x^*) = A^{\mathsf{T}}\lambda^*$.
2 Case $\beta^{\mathsf{T}}x^* + \delta = 0$: there exists $\lambda^* \in \mathbb{R}^m$ and $\eta^* \in K_{\text{soc}}$ such that $\nabla f(0) = A^{\mathsf{T}}\lambda^* + \tilde{B}^{\mathsf{T}}\eta^*$.

This is derived in Chapter 12.8.3 using techniques for nonsmooth analysis (see (12.64)).

Conic program.

A generalization of SOCP (8.66) and (8.68) is the following convex optimization

$$f^* \quad := \quad \min_{x \in \mathbb{R}^n} \; f(x) \;\; \text{s.t.} \;\; Ax = b, \; x \in K \tag{8.70}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a real-valued convex function, $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, and $K \subseteq \mathbb{R}^n$ is a general closed convex cone. The Slater Theorem 8.17 and the KKT Theorem 8.15 are formulated in this chapter for problems where the constraint functions are explicitly given and continuously differentiable. Even though part of the constraints in (8.70) is not explicit, since K is a convex cone, a dual problem can be formulated in terms of what is called its dual cone. We derive in Chapter 12.8.4 a sufficient condition for strong duality and dual optimality and the KKT condition for the general conic program (8.70) where the constraint functions are not fully specified and the cost function f is convex but not necessarily continuously differentiable (Theorem 12.31).

8.4.5 Semidefinite program (SDP)

Recall the vector space $\mathbb{S}^n$ of Hermitian matrices over the field $\mathbb{R}$ of real numbers, not over $\mathbb{C}$, and the cone K_{psd} of positive semidefinite matrices in the vector space $\mathbb{S}^n$, studied in Chapter 8.2.2. For two Hermitian matrices $x, y \in \mathbb{S}^n$, their inner product is $x \cdot y := \text{tr}\left(y^{\mathsf{H}}x\right) = \sum_{j,k} x_{jk}\overline{y}_{jk}$ is a real number and satisfies $x \cdot y = y \cdot x$. Furthermore K_{psd} is a proper self-dual cone.

Consider the following convex optimization

$$f^* \;:=\; \min_{x\in\mathbb{R}^n} f(x) \;\text{ s.t. }\; Ax=b,\; h(x)\in K_{\text{psd}} \tag{8.71a}$$

where $f:\mathbb{R}^n\to\mathbb{R}$ is a real-valued convex function, $A\in\mathbb{R}^{m\times n}$, $b\in\mathbb{R}^m$, $K_{\text{psd}}\subseteq\mathbb{S}^l$ is the cone of positive semidefinite matrices, and $h:\mathbb{R}^n\to\mathbb{S}^l$ is the function

$$h(x) \;:=\; B_0+\sum_{i=1}^n x_iB_i, \qquad B_i\in\mathbb{S}^l,\; i\geq 0 \tag{8.71b}$$

The constraint $h(x)\in K_{\text{psd}}$ is called a *linear matrix inequality* and is sometimes denoted as $h(x)\succeq_{K_{\text{psd}}} 0$ or simply $h(x)\succeq 0$ if the underlying cone K_{psd} is understood. SDP (8.71) reduces to LP if $l=1$ (see Example 8.13 of Chapter 8.4.2). It also includes SOCP (8.66a) as a special case because $x\in K_{\text{soc}}$ if and only if the "arrow matrix" $\begin{bmatrix} x_n & [x^{n-1}]^{\mathsf{T}} \\ x^{n-1} & x_n\mathbb{I}_{n-1}\end{bmatrix}\in K_{\text{psd}}$. This is because, if $x_n>0$, then the arrow matrix is psd if and only if its Schur complement $x_n-\|x^{n-1}]\|_2^2/x_n\geq 0$. (Theorem A.4 in Chapter A.3.1).

To define the dual problem let $\lambda\in\mathbb{R}^m$ and $Z\in K_{\text{psd}}^*\subseteq\mathbb{S}^l$ denote dual variables, where K_{psd}^* is the dual cone of K_{psd}. The Lagrangian is[7], for $x\in\mathbb{R}^n,\lambda\in\mathbb{R}^m,Z\in K_{\text{psd}}^*$,

$$L(x,\lambda,Z) := f(x)-\lambda^{\mathsf{T}}(Ax-b)-Z\cdot\left(B_0+\sum_{i=1}^n x_iB_i\right) \tag{8.72a}$$

The dual function $d(\lambda,Z):=\min_{x\in\mathbb{R}^n} L(x,\lambda,Z)$ is:

$$\begin{aligned} d(\lambda,Z) \;&=\; \left(b^{\mathsf{T}}\lambda - Z\cdot B_0\right)+d_0(\lambda,Z)\\ d_0(\lambda,Z) \;&:=\; \min_{x\in\mathbb{R}^n}\; f(x)-\lambda^{\mathsf{T}}Ax-\sum_i x_i(Z\cdot B_i)\end{aligned} \tag{8.72b}$$

Hence the dual problem is

$$d^* \;:=\; \max_{\lambda\in\mathbb{R}^m,Z\in\mathbb{S}^l}\left(b^{\mathsf{T}}\lambda-\text{tr}(B_0^{\mathsf{H}}Z)\right)+d_0(\lambda,Z) \qquad \text{s.t.} \qquad Z\in K_{\text{psd}}^* \tag{8.72c}$$

If $f(x)=c^{\mathsf{T}}x$ then

$$d^* \;:=\; \max_{\lambda\in\mathbb{R}^m,Z\in K_{\text{psd}}^*}\left(b^{\mathsf{T}}\lambda-\text{tr}(B_0^{\mathsf{H}}Z)\right) \quad \text{s.t.} \quad c_i=\sum_j A_{ji}\lambda_j+\text{tr}(B_i^{\mathsf{H}}Z),\; i=1,\ldots,n$$

A point $(x^*,\lambda^*,Z^*)\in\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{S}^l$ is a saddle point if

$$\min_{x\in\mathbb{R}^n} L(x,\lambda^*,Z^*) \;=\; L(x^*,\lambda^*,Z^*) \;=\; \max_{\lambda\in\mathbb{R}^m,Z\in K_{\text{psd}}^*} L(x^*,\lambda,Z)$$

Strong duality, dual optimality and KKT characterization of SDP (8.71) is stated in the following theorem. Since K_{psd} is self-dual, i.e., $K_{\text{psd}}^*=K_{\text{psd}}$, K_{psd}^* above can all be replaced by K_{psd}. This property is not important and therefore we continue to use K_{psd}^* in the theorem.

[7] A justification for the definition of Lagrangian is weak duality: for any feasible x and any $Z\in K_{\text{psd}}^*$, $L(x,\lambda,Z)\leq f(x)$ since $h(x)\in K_{\text{psd}}$ and hence $Z\cdot h(x)\geq 0$.

Theorem 8.28 (SDP strong duality and KKT)**.** Consider the SDP (8.71) and its dual (8.72). Suppose there exists $\bar{x} \in \mathbb{R}^n$ such that $A\bar{x} = b$ and $h(\bar{x}) \in \mathrm{ri}(K_{\mathrm{psd}})$. Then

1 *Strong duality and dual optimality*. If f^* is finite then there exists a dual optimal solution $(\lambda^*, Z^*) \in \mathbb{R}^m \times K_{\mathrm{psd}}^*$ that closes the duality gap, i.e., $f^* = d^* = d(\lambda^*, Z^*)$.
2 *KKT characterization*. A feasible x^* is optimal if and only if there exists a dual feasible $(\lambda^*, Z^*) \in \mathbb{R}^m \times K_{\mathrm{psd}}^*$ such that

$$\mathrm{tr}(h(x^*)^{\mathsf{H}} Z^*) \;=\; 0, \qquad \frac{\partial f}{\partial x_i}(x^*) \;=\; \sum_j A_{ji}\lambda_j + \mathrm{tr}(B_i^{\mathsf{H}} Z^*), \;\; i = 1, \ldots, n$$

In this case (x^*, Z^*) is a saddle point that closes the duality gap and is primal-dual optimal. □

Theorem 8.28, as well as Theorem 8.29 below, are obtained by applying the Slater Theorem 12.27 and the generalized KKT Theorem 12.21 to the vector space of $\mathbb{S}^n$ and SDP.

We often use the following form of the semidefinite program with inequality constraints:

$$d^* \;:=\; \min_{Z \in K_{\mathrm{psd}}} \mathrm{tr}\left(B_0^{\mathsf{H}} Z\right) \quad \text{s.t.} \quad \mathrm{tr}\left(B_i^{\mathsf{H}} Z\right) \le c_i, \;\; i = 1, \ldots, n \tag{8.73}$$

where $K_{\mathrm{psd}} \subset \mathbb{S}^l$. For instance the semidefinite relaxation of optimal power flow problems in Chapter 10.1.1 takes this form. This is equivalent to problem (8.72) without the affine constraint $Ax = b$, noting that $K_{\mathrm{psd}}^* = K_{\mathrm{psd}}$. We now derive its dual problem.

Let the Lagrangian be

$$L(Z, x) \;:=\; \mathrm{tr}\left(B_0^{\mathsf{H}} Z\right) + \sum_{i=1}^n x_i \left(\mathrm{tr}(B_i^{\mathsf{H}} Z) - c_i\right), \qquad Z \in K_{\mathrm{psd}}, \; x \in \mathbb{R}_+^n$$

and the dual function be $f(x) := \min_{Z \in K_{\mathrm{psd}}} L(Z, x) = -c^{\mathsf{T}} x + \min_{Z \in K_{\mathrm{psd}}} Z \cdot h(x)$, where $h(x)$ is defined in (8.71b). Since the constraint $Z \in K_{\mathrm{psd}}$ is not dualized, the minimization over Z in $f(x)$ is over K_{psd}, not $\mathbb{S}^l$. If $h(x) \in K_{\mathrm{psd}}^*$ then $Z \cdot h(x) \ge 0$ for all $Z \in K_{\mathrm{psd}}$ whereas if $h(x) \notin K_{\mathrm{psd}}^*$ then, by the definition of dual cone, there exists $\bar{Z} \in K_{\mathrm{psd}}$ such that $\bar{Z} \cdot h(x) < 0$. Hence

$$\min_{Z \in K_{\mathrm{psd}}} Z \cdot h(x) \;=\; \begin{cases} 0 & \text{if } \; h(x) \in K_{\mathrm{psd}}^* \\ -\infty & \text{otherwise} \end{cases}$$

Since $K_{\mathrm{psd}}^* = K_{\mathrm{psd}}$, the dual function is then (recalling that $x \ge 0$)

$$f(x) \;=\; \begin{cases} -c^{\mathsf{T}} x & \text{if } \; x \ge 0, \; h(x) \in K_{\mathrm{psd}} \\ -\infty & \text{otherwise} \end{cases} \tag{8.74a}$$

The dual problem $f^* := \max_{x \in \mathbb{R}^n} f(x)$ is

$$f^* \;:=\; -\min_{x \in \mathbb{R}^n} c^{\mathsf{T}} x \quad \text{s.t.} \quad x \ge 0, \; h(x) \in K_{\mathrm{psd}} \tag{8.74b}$$

Recall that all eigenvalues of a matrix $Z \in K_{\text{psd}}$ are nonnegative. The interior $\text{int}(K_{\text{psd}})$ of K_{psd} is the set of all positive definite matrices whose eigenvalues are strictly positive.

Theorem 8.29 (SDP strong duality and KKT)**.** Consider the SDP (8.73) and its dual (8.74). Suppose there exists a positive definite matrix $\bar{Z} \in \text{int}(K_{\text{psd}})$ such that $\text{tr}\left(B_i^{\mathsf{H}} \bar{Z}\right) \leq c_i$, for $i = 1, \ldots, n$. Then

1 *Strong duality and dual optimality*. If d^* is finite then there exists a dual optimal solution $x \in \mathbb{R}^n$ that closes the duality gap, i.e., $d^* = f^* = f(x^*)$.
2 *KKT characterization*. A feasible $Z^* \in K_{\text{psd}}$ is optimal if and only if there exists an $x^* \in \mathbb{R}^n$ such that

$$h(x^*) \in K_{\text{psd}}, \qquad x^* \geq 0, \qquad x_i^* \left(c_i - Z^* \cdot B_i\right) = 0, \quad i = 1, \ldots, n$$

In this case (x^*, Z^*) is a saddle point that closes the duality gap and is primal-dual optimal. □

8.5 Optimization algorithms

Even though OPF can be formulated as an optimization problem in the complex domain using the complex form of power flow equations (e.g., in (9.9) or (9.16) for single-phase OPF in BIM), in computing a solution, it is first converted into a problem in the real domain; see Remark 9.2. OPF can also be formulated directly in the real domain using the polar form (4.27) or the Cartesian form (4.28) of the power flow equations. We therefore present and analyze algorithms for solving OPF in the real domain.

Consider the problem

$$\min_x \ f(x) \ \text{ subject to } \ x \in X \tag{8.75}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and $X \subseteq \mathbb{R}^n$ is nonempty, closed and convex. Let the column vector $\nabla f(x)$ denote the gradient of f evaluated at x, i.e., $[\nabla f(x)]_i := \partial f / \partial x_i$, $i = 1, \ldots, n$. Recall that a point x^* is a *local minimizer* if $f(x^*)$ is minimum on a neighborhood of x^*, i.e., there exists $r > 0$ such that $f(x^*) \leq f(x)$ for all $x \in B_r(x^*) \cap X$. It is a *global minimizer* if $f(x^*) \leq f(x)$ for all $x \in X$.

If $X = \mathbb{R}^n$ then the minimization is unconstrained. The condition $\nabla f(x^*) = 0$ is necessary for x^* to be a local minimizer; if f is convex then it is also sufficient for x^* to be a global minimizer. For constrained minimization where X is a strict subset of $\mathbb{R}^n$, the condition $\nabla f(x^*) = 0$ is generalized to: if $x^* \in X$ is a local minimizer for (8.75) then there is a neighborhood $B_r(x^*)$ for some $r > 0$ such that

$$\left(\nabla f(x^*)\right)^{\mathsf{T}} (x - x^*) \geq 0 \qquad \forall x \in B_r(x^*) \cap X \tag{8.76}$$

i.e., moving away from x^* to any other feasible point x in $B_r(x^*)$ can only increase the function value f. If f is a convex function (X is assumed convex) then this is both

necessary and sufficient for x^* to be a global optimum of (8.75). This is illustrated in Figure 8.16.

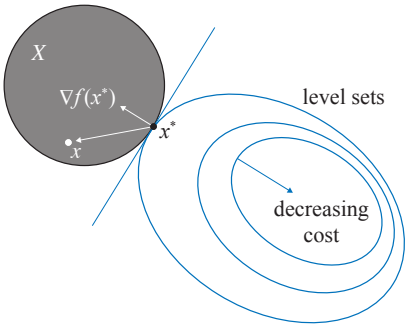


Figure 8.16 Moving away from x^* to another feasible x locally increases the cost.

In most applications an optimum of (8.75) cannot be solved in closed form and must be computed iteratively. Iterative algorithms generally take the form

$$x(t+1) \;=\; g(x(t)) \tag{8.77a}$$

i.e., the next iterate $x(t+1)$ is determined from the current iterate $x(t)$ according to an algorithm represented by the function $g : \mathbb{R}^n \to \mathbb{R}^n$, until a certain termination criterion is met, e.g., when $x(t+1)$ satisfies the optimality condition (8.76) approximately. This is also called a fixed-point iteration that computes a fixed point x^* satisfying $x^* = g(x^*)$ (fixed-point iteration is studied in Chapter 8.6.1). For example a gradient descent algorithm can be interpreted as the following fixed-point iteration:

$$x(t+1) \;=\; [x(t) - \gamma G(x(t))\nabla f(x(t)]_X \;=:\; g(x(t)) \tag{8.77b}$$

where $\gamma > 0$ is a stepsize, $G(x) \succ 0$ is a scaling matrix, and $[\cdot]_X$ is the projection to the feasible set X. A fixed point x^* of the gradient algorithm (8.77b) satisfies the optimality condition (8.76).

In this section we present several algorithms for solving (8.75). The convergence of some of these algorithms are analyzed in Chapter 8.6.

8.5.1 Steepest descent algorithm

Steepest descent is the most widely used class of iterative algorithms for solving optimization problems. For (8.75), it is given by the following iteration: starting from an initial point $x(0) = x_0$,

$$x(t+1) := [\; x(t) - \gamma \nabla f(x(t)) \;]_X \tag{8.78}$$

where $\gamma > 0$ is a constant stepsize and X is a nonempty, closed and convex subset of $\mathbb{R}^n$. Here $[x]_X$ denotes the projection of x onto the nonempty, closed and convex set X, i.e., for any $x \in \mathbb{R}^n$,

$$[x]_X := \arg\min_{y \in X} \ \|x - y\|_2$$

where $\|\cdot\|_2$ is the *Euclidean* norm. Hence $[x]_X$ is the unique point in X that is closest to $x \in \mathbb{R}^n$ in the Euclidean norm. As mentioned above, a fixed point x^* defined by $x^* = [x^* - \gamma \nabla f(x^*)]_X$ satisfies the optimality condition (8.76):

$$\frac{\partial f}{\partial x}(x^*)(x - x^*) \ \geq \ 0 \qquad \forall x \in X$$

when f is a convex function and X a convex set. A termination criterion for (8.78) can be $\|x(t+1) - x(t)\| < \epsilon$ for a pre-determined $\epsilon > 0$.

Variants of the steepest descent algorithm can be obtained by using an iteration-dependent stepsize $\gamma(t) > 0$ or a scaling matrix $\gamma \in \mathbb{R}^{n \times n}$. The steepest descent algorithm is called a first-order algorithm because it uses only the first derivative of the objective function f. A second-order algorithm, such as the Newton-Raphson algorithm widely used for solving optimal power flow problems, uses the second derivative to construct a time-dependent scaling matrix $\gamma(t)$ in each iteration, as we now explain.

A popular algorithm for regression and machine learning applications is the *stochastic gradient descent*. In the simplest form it is an algorithm to solve

$$\min_{x \in \mathbb{R}^n} \ \sum_{i=1}^{m} f_i(x)$$

The standard gradient descent algorithm computes $x(t+1) := x(t) - \gamma \sum_i \nabla f_i(x(t))$ at time t. A stochastic gradient descent algorithm approximates the true gradient $\sum_i \nabla f_i(x(t))$ by the gradient at a sample i, chosen randomly or in an online fashion:

$$x(t+1) \ := \ x(t) - \gamma \nabla f_i(x(t))$$

Typically, each i represents a sample. For example we are given m samples $(u_i, y_i) \in \mathbb{R}^n \times \mathbb{R}$, $i = 1, \ldots, m$, and we are to choose weights x to minimize the mean square error $\sum_i (u_i^\mathsf{T} x - y_i)^2$ or a loss function $\sum_i f(x; u_i, y_i)$. In an online setting, the tth sample (u_t, y_t) may be revealed only at time t and the existing weights $x(t)$ are then updated with the approximate gradient $\nabla f_t(x(t)) = 2(u_t^\mathsf{T} x(t) - y_t)u_t$ or $\nabla f_t(x(t)) = \nabla_x f(x(t); u_t, y_t)$ respectively. Even though solution may take more iterations with approximate gradient, each iteration can be much easier to compute than with true gradient. A generalization is to use a small number $k \ll m$ of gradients in each iteration, i.e., $x(t+1) \ := \ x(t) - \gamma \sum_{i=1}^{k} \nabla f_i(x(t))$.

8.5.2 Newton-Raphson algorithm

As explained in Chapter 4.4.2, Newton-Raphson is an iterative algorithm for solving nonlinear equations $F(y) = 0$ where $F : \mathbb{R}^n \to \mathbb{R}^n$. It computes iteratively

$$y(t+1) \;=\; y(t) \;+\; \Delta y(t) \qquad \text{where} \qquad J(y(t))\Delta y(t) \;=\; -F(y(t)) \tag{8.79}$$

where $J(y) := \frac{\partial F}{\partial y}(y)$ is the Jacobian of F at y. In this section we apply it to optimization problems where the equation $F(y) = 0$ represents the KKT condition. A solution y^{opt} of $F(y) = 0$ then produces an optimal solution if the underlying optimization problem is convex. For simplicity we assume solutions exist for all the optimization problems considered unless otherwise specified.

Specifically we will present algorithms for:

1 *Linear equality constrained problems*. The idea is to approximate the cost function by a quadratic function around the next iterate (to be determined). This results in a quadratic program in each iteration whose KKT condition is a system of linear equations that can be solved analytically for the next iterate. We will also describe another algorithm that generalizes to nonlinear constraints.
2 *Nonlinear equality constrained problems*. In contrast to the KKT condition of an approximating quadratic program, the KKT condition of these problems is generally nonlinear and cannot be solved analytically. The idea is to solve the KKT condition iteratively using the Newton-Raphson method.
3 *Inequality constrained problems*. The KKT condition of these problems involves inequalities and Newton-Raphson is not directly applicable. The idea is to replace the inequality constraint by a penalty term in the cost function to obtain an approximate problem that has no inequality constraints.

Nonlinear program with linear equality constraint.

Consider the following problem with an equality constraint:

$$\min_{x \in \mathbb{R}^n} \quad f(x) \qquad \text{s.t.} \qquad Ax \;=\; b \tag{8.80}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable and $A \in \mathbb{R}^{m \times n}$. We will derive two equivalent algorithms. The first algorithm relies on the linearity of the constraint and is generally not applicable to problems with nonlinear constraints. It approximates the cost function $f(x)$ by a quadratic function in each iteration and solves the resulting quadratic program directly. The second algorithm solves the KKT condition for (8.80) and extends directly to problems with nonlinear equality constraints.

For the first algorithm, given the current iterate $x(t)$, approximate the cost $f(x(t) + \Delta x(t))$ at the *next* iterate by

$$\hat{f}(x(t)+\Delta x(t)) \;:=\; f(x(t)) + \frac{\partial f}{\partial x}(x(t))\,\Delta x(t) + \frac{1}{2}\Delta x(t)^{\mathsf{T}} \frac{\partial^2 f}{\partial x^2}(x(t))\,\Delta x(t) \tag{8.81a}$$

and consider the optimization over $\Delta x(t)$

$$\min_{\Delta x \in \mathbb{R}^n} \quad \hat{f}(x(t)+\Delta x(t)) \qquad \text{s.t.} \qquad A(x(t)+\Delta x(t)) \;=\; b \tag{8.81b}$$

This is a quadratic program in $\Delta x(t)$ with a fixed $x(t)$ and can be solved analytically. Let $\lambda(t) \in \mathbb{R}^m$ be the Lagrange multiplier of (8.81). If f is convex then (8.81) is a convex program and the KKT condition is both necessary and sufficient for optimality. The KKT condition is (Exercise 8.28)

$$\underbrace{\begin{bmatrix} \frac{\partial^2 f}{\partial x^2}(x(t)) & A^\mathsf{T} \\ A & 0 \end{bmatrix}}_{K(t)} \begin{bmatrix} \Delta x(t) \\ \lambda(t) \end{bmatrix} \;=\; -\underbrace{\begin{bmatrix} \nabla f(x(t)) \\ Ax(t)-b \end{bmatrix}}_{d(x(t))} \tag{8.82a}$$

This is system of $n+m$ linear equations in $n+m$ unknowns $(\Delta x(t), \lambda(t))$. The matrix $K(t)$ on the left-hand side of (8.82a) is called a KKT matrix. If $K(t)$ is nonsingular[8] then $\Delta x(t)$ can be computed directly. If $K(t)$ is singular but the given vector $d(x(t))$ on the right-hand side is orthogonal to the null space of $K(t)$, then there is a subspace of solutions $(\Delta x(t), \lambda(t))$ to (8.82a) and $-K^\dagger(t)d(x(t))$ is the minimum-norm solution where $K^\dagger(t)$ is the pseudo inverse of $K(t)$. Neither $K(t)$ nor $d(x(t))$ depends on $\lambda(t)$. Hence in both cases $\Delta x(t)$ can be computed from just the current iterate $x(t)$ and (8.82a) always allows pure primal iterations,

$$x(t+1) \;=\; x(t)+\Delta x(t) \tag{8.82b}$$

for solving (8.80).

The second algorithm does not use the second-order approximation of $f(x)$ and considers (8.80) directly. Specifically let $\lambda \in \mathbb{R}^m$ denote the Lagrange multiplier associated with the m constraints in (8.80). The Lagrangian is

$$L(x;\lambda) \;:=\; f(x) \;+\; \lambda^\mathsf{T}(Ax-b)$$

Let $y := (x,\lambda) \in \mathbb{R}^{n+m}$ and define $F : \mathbb{R}^{n+m} \to \mathbb{R}^{n+m}$ by

$$F(y) \;:=\; \begin{bmatrix} \nabla_x L(x,\lambda) \\ \nabla_\lambda L(x,\lambda) \end{bmatrix} \;=\; \begin{bmatrix} \nabla f(x) + A^\mathsf{T}\lambda \\ Ax-b \end{bmatrix}$$

The KKT condition is $F(y)=0$. This specifies a system of $n+m$ nonlinear equations in $n+m$ unknowns (x,λ), in contrast to the linear KKT condition (8.82a) for the second-order approximation (8.81). It generally needs to be solved iteratively. The Jacobian $J(y) := \frac{\partial F}{\partial y}$ of F is:

$$J(y) \;=\; \begin{bmatrix} \frac{\partial^2 f}{\partial x^2}(x) & A^\mathsf{T} \\ A & 0 \end{bmatrix}$$

(which is the KKT matrix $K(t)$ in (8.82a).) Hence the Newton-Raphson iteration is

$$\begin{bmatrix} x(t+1) \\ \lambda(t+1) \end{bmatrix} \;=\; \begin{bmatrix} x(t) \\ \lambda(t) \end{bmatrix} + \begin{bmatrix} \Delta x(t) \\ \Delta\lambda(t) \end{bmatrix} \tag{8.83a}$$

[8] See [57, Chapter 10.1, p.523] for equivalent conditions of the nonsingularity of the KKT matrix $K(t)$.

where the increment $\Delta y(t)$ is given by $J(y(t))\,\Delta y(t) = -F(y(t))$, i.e.,

$$\begin{bmatrix} \frac{\partial^2 f}{\partial x^2}(x(t)) & A^\mathsf{T} \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x(t) \\ \Delta \lambda(t) \end{bmatrix} = -\begin{bmatrix} \nabla f(x(t)) + A^\mathsf{T}\lambda(t) \\ Ax(t) - b \end{bmatrix} \tag{8.83b}$$

We compare the two algorithms (8.82) and (8.83). Both algorithms solve a linear equation with the KKT matrix $K(t)$ in each iteration. As mentioned above, the approach of (8.82) solves the KKT condition for the second-order approximation (8.81) of f. This is possible because the linearity of the constraint allows a second-order approximation of only the cost function but not of the constraint, resulting in a quadratic program that can be solved analytically. It leads to a primal algorithm that iterates only on $x(t)$. This is generally inapplicable if the constraint is nonlinear. The approach of (8.83), on the other hand, solves the KKT condition $F(x,\lambda) = 0$ for the original problem (8.80) iteratively using the Newton-Raphson algorithm (8.79). It leads to a primal-dual algorithm that updates both the primal and the dual variables. It will be extended to problems with a nonlinear constraint in (8.85).

These two algorithms are equivalent in that both produce the same sequence of $(x(t),\lambda(t))$ starting from the same initial point. Indeed, given the current iterate $(x(t),\lambda(t))$ of the primal and dual variables, $(\Delta x(t),\Delta\lambda(t))$ satisfies (8.83) if and only if $(\Delta x(t),\lambda := \lambda(t)+\Delta\lambda(t))$ satisfies (8.82). To see this, suppose $(\Delta x(t),\lambda(t)+\Delta\lambda(t))$ satisfies (8.82), i.e.,

$$\begin{bmatrix} \frac{\partial^2 f}{\partial x^2}(x(t)) & A^\mathsf{T} \\ A & 0 \end{bmatrix} \left(\begin{bmatrix} \Delta x(t) \\ \Delta \lambda(t) \end{bmatrix} + \begin{bmatrix} 0 \\ \lambda(t) \end{bmatrix} \right) = -\begin{bmatrix} \nabla f(x(t)) \\ Ax(t) - b \end{bmatrix}$$

which yields (8.83). Suppose the converse holds. Write the right-hand side of (8.83) as

$$\begin{bmatrix} \nabla f(x(t)) + A^\mathsf{T}\lambda(t) \\ Ax(t) - b \end{bmatrix} = \begin{bmatrix} \nabla f(x(t)) \\ Ax(t) - b \end{bmatrix} + \begin{bmatrix} \frac{\partial^2 f}{\partial x^2}(x(t)) & A^\mathsf{T} \\ A & 0 \end{bmatrix} \begin{bmatrix} 0 \\ \lambda(t) \end{bmatrix}$$

which, together with (8.83), yields (8.82). The only difference between these algorithms is that (8.83) computes $\Delta\lambda(t)$ from $(x(t),\lambda(t))$ and forms $\lambda(t+1)$ whereas (8.82) computes $\lambda(t+1)$ directly from $x(t)$.

Nonlinear program with equality constraint.

Consider the following problem with a possibly nonlinear equality constraint

$$\min_{x\in\mathbb{R}^n} \; f(x) \qquad \text{s.t.} \qquad g(x) = 0 \tag{8.84}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ are twice continuously differentiable. The approach of (8.83) generalizes directly to this problem. Let $\lambda \in \mathbb{R}^m$ denote the Lagrange multiplier associated with the m constraints. The Lagrangian is

$$L(x;\lambda) := f(x) + \lambda^\mathsf{T} g(x)$$

Let $y := (x, \lambda) \in \mathbb{R}^{n+m}$ and define $F : \mathbb{R}^{n+m} \to \mathbb{R}^{n+m}$ by

$$F(y) \;:=\; \begin{bmatrix} \nabla_x L(x,\lambda) \\ \nabla_\lambda L(x,\lambda) \end{bmatrix} \;=\; \begin{bmatrix} \nabla f(x) \;+\; \frac{\partial g}{\partial x}(x)^{\mathsf{T}} \lambda \\ g(x) \end{bmatrix} \tag{8.85a}$$

The KKT condition is $F(y) = 0$ which specifies a system of $n+m$ nonlinear equations in $n+m$ unknowns (x, λ). Hence the Jacobian $J(y) := \frac{\partial F}{\partial y}$ of F is:

$$J(y) \;:=\; \begin{bmatrix} \frac{\partial^2 L}{\partial x^2} & \frac{\partial^2 L}{\partial \lambda \partial x} \\ \frac{\partial^2 L}{\partial x \partial \lambda} & \frac{\partial^2 L}{\partial \lambda^2} \end{bmatrix} \;=\; \left[\begin{array}{c|c} \frac{\partial^2 f}{\partial x^2}(x) \;+\; \sum_k \frac{\partial^2 g_k}{\partial x^2} \lambda_k & \frac{\partial g}{\partial x}(x)^{\mathsf{T}} \\ \hline \frac{\partial g}{\partial x}(x) & 0 \end{array}\right] \tag{8.85b}$$

which reduces to the Jacobian in (8.83b) when $g(x) = Ax - b$. Here $\frac{\partial^2 L}{\partial \lambda \partial x} = \left(\frac{\partial^2 L}{\partial \lambda \partial x}\right)^{\mathsf{T}}$ is $n \times m$. The Newton-Raphson algorithm for solving (8.84) is the iteration (8.79) where $F(y)$ and its Jacobian $J(y)$ are given by (8.85). It is a primal-dual algorithm that iterates on both $x(t)$ and $\lambda(t)$.

When the cost function $f(x)$ or the feasible set $\{x \in \mathbb{R}^n : g(x) = 0\}$ is nonconvex, there is generally no guarantee that the Newton-Raphson algorithm will converge and if it does, it will produce a local or global optimum. In practice, for OPF problems, the algorithm often converges to a local, and even global, optimum despite nonconvexity.

When f and g are homogeneous quadratic functions the nonlinear program reduces to the following QCQP with equality constraints:

$$\min_{x \in \mathbb{R}^n} \quad \frac{1}{2} x^{\mathsf{T}} C_0 x \qquad \text{s.t.} \qquad \frac{1}{2} x^{\mathsf{T}} C_k x \;=\; b_k, \qquad k = 1, \ldots, m$$

where $C_k \in \mathbb{R}^{n \times n}$, $k \geq 0$, are real symmetric matrices and $b_k \in \mathbb{R}$, $k \geq 1$. Then (8.85) reduces to:

$$F(y) \;:=\; \begin{bmatrix} \nabla_x L(y) \\ \nabla_\lambda L(y) \end{bmatrix} \;=\; \left[\begin{array}{c} A(\lambda)^{\mathsf{T}} x \\ \hline \frac{1}{2} x^{\mathsf{T}} C_1 x - b_1 \\ \vdots \\ \frac{1}{2} x^{\mathsf{T}} C_m x - b_m \end{array}\right]$$

where $A(\lambda) := C_0 + \sum_k \lambda_k C_k$ and

$$J(y) \;:=\; \begin{bmatrix} \frac{\partial^2 L}{\partial x^2} & \frac{\partial^2 L}{\partial \lambda \partial x} \\ \frac{\partial^2 L}{\partial x \partial \lambda} & \frac{\partial^2 L}{\partial \lambda^2} \end{bmatrix} \;=\; \left[\begin{array}{c|ccc} A(\lambda)^{\mathsf{T}} & C_1^{\mathsf{T}} x & \cdots & C_m^{\mathsf{T}} x \\ \hline x^{\mathsf{T}} C_1 & & & \\ \vdots & & 0 & \\ x^{\mathsf{T}} C_m & & & \end{array}\right]$$

Nonlinear program with inequality constraint.

Consider the following problem with an inequality constraint

$$\min_{x \in \mathbb{R}^n} \quad f(x) \qquad \text{s.t.} \qquad h(x) \;\leq\; 0 \tag{8.86}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}^l$ are twice continuously differentiable. Let $\mu \in \mathbb{R}^l$ denote the Lagrange multiplier associated with the m constraints. The KKT condition involves inequalities, of the form

$$\begin{aligned} \nabla_x L(x,\mu) &= \nabla_x f(x) + \frac{\partial h}{\partial x}(x)^\mathsf{T}\mu = 0, \qquad \nabla_\mu L(x,\mu) = h(x) \leq 0 \\ \mu &\geq 0, \qquad\qquad\qquad\qquad \mu^\mathsf{T} h(x) = 0 \end{aligned}$$

The standard Newton-Raphson method cannot be applied directly to solve this system of equalities and inequalities. There are however many Newton-like methods that have been developed for inequality constrained problems.

One approach is to introduce a slack variable $z \in \mathbb{R}^l$ and convert (8.86) into a problem with a 'simple' inequality constraint:

$$\min_{(x,z)\in\mathbb{R}^{n+l}} \quad f(x) \qquad \text{s.t.} \qquad h(x) + z = 0, \quad z \geq 0$$

Algorithms for solving equality constrained problems can be modified by projecting $z(t)$ to the nonnegative quadrant in each iteration; see e.g. [58]. Another approach is to replace the constraint $h(x) \leq 0$ in (8.86) by a penalty term $(1/t)\phi(x)$ in the cost function and solve the resulting unconstrained approximate problem

$$\min_{x\in\mathbb{R}^n} \quad f(x) + \frac{1}{t}\phi(x)$$

where $t > 0$ is a parameter that controls the accuracy of the approximation. Newton-Raphson can be applied to solve the optimality condition $\nabla f(x) + (1/t)\nabla\phi(x) = 0$. This is the approach of the interior point methods which we describe next.

8.5.3 Interior-point algorithm

Consider the following problem with an equality and an inequality constraints:

$$f^* := \min_{x\in\mathbb{R}^n} f(x) \qquad \text{s.t.} \qquad g(x) = 0,\ h(x) \leq 0 \tag{8.87}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^m$, and $h : \mathbb{R}^n \to \mathbb{R}^l$ are twice continuously differentiable. The idea is to approximate (8.87) by an equality constrained problem by replacing the inequality constraint $h(x) \leq 0$ by a penalty term in the cost function, and then solving the equality constrained problem using Newton methods.

Log barrier function.

A popular barrier function is $\varphi : \mathbb{R}_- \to \mathbb{R}$ defined by:

$$\varphi_t(u) := -\frac{1}{t}\log(-u), \qquad u < 0$$

where $t > 0$ is a parameter. For each $t > 0$, the function $\varphi_t(u)$ is convex increasing over its domain $u < 0$ and approaches ∞ as $u \to 0$. It is an approximation of the indicator

function which takes the value 0 if $u \leq 0$ and ∞ if $u > 0$. The larger the parameter t is, the more accurate the approximation will be. While the indicator function is discontinuous, the log barrier function $\varphi_t(u)$ is continuously differentiable over its domain $u < 0$ for each $t > 0$.

The *logarithmic barrier* $\phi : \mathbb{R}^n \to \mathbb{R}$ is

$$\phi(x) := -\sum_{i=1}^{l} \log(-h_i(x)) \tag{8.88a}$$

over the domain

$$\text{dom}(\phi) := \{x \in \mathbb{R}^n : h_i(x) < 0, i = 1, \ldots, l\}$$

The log barrier $\phi(x)$ grows without bound as $h_i(x) \to 0$ for any i. Its gradient and Hessian are:

$$\nabla\phi(x) = \sum_{i=1}^{l} \frac{1}{-h_i(x)} \nabla h_i(x) \tag{8.88b}$$

$$\frac{\partial^2 \phi}{\partial x^2}(x) = \sum_i \frac{1}{h_i^2(x)} \nabla h_i(x) \nabla h_i^{\mathsf{T}}(x) + \sum_i \frac{1}{-h_i(x)} \frac{\partial^2 h_i}{\partial x^2}(x) \tag{8.88c}$$

The approximate problem.

Fix any $t > 0$. An approximate problem to (8.87) with an equality constraint is

$$\min_{x \in \mathbb{R}^n} f(x) + \frac{1}{t}\phi(x) \qquad \text{s.t.} \qquad g(x) = 0$$

It is more convenient to consider the following equivalent approximate problem (they have the same minimizers):

$$\text{Problem}(t): \qquad \min_{x \in \mathbb{R}^n} t f(x) + \phi(x) \qquad \text{s.t.} \qquad g(x) = 0 \tag{8.89}$$

Unlike (8.87) Problem(t) (8.89) has only equality constraints and therefore can be solved using the Newton-Raphson algorithm defined by (8.79)(8.85). If f is convex and g is affine then Problem(t) is a convex problem. In that case, if the Newton-Raphson algorithm converges to a solution (x_t, λ_t), then the solution satisfies the KKT condition and is therefore primal and dual optimal, i.e., x_t solves (8.89) and λ_t solves its dual. Otherwise, Problem(t) is nonconvex and there is generally no guarantee that the Newton-Raphson algorithm will converge. If it does converge, it will produce a feasible solution but there is no guarantee that it is a local or global optimum. In practice, for OPF problems, the algorithm often converges to a local, and even global, optimum despite nonconvexity.

Suppose the Newton-Raphson algorithm converges and produces a solution $(x_t, \hat{\lambda}_t)$ that satisfies the KKT condition of Problem(t) (8.89). We now show that x_t, together

with an associated (λ_t, μ_t) to be defined, satisfy approximately the KKT condition of the original problem (8.87). Define the Lagrangian of Problem(t):

$$L_t(x, \hat{\lambda}) \;:=\; tf(x) + \phi(x) + \hat{\lambda}^{\mathsf{T}} g(x)$$

The KKT condition for Problem(t) consists of primal feasibility and stationarity, $F_t(x_t, \hat{\lambda}_t) = 0$ where

$$F_t(x_t, \hat{\lambda}_t) \;:=\; \begin{bmatrix} \nabla_x L_t(x_t, \hat{\lambda}_t) \\ \nabla_{\hat{\lambda}} L_t(x_t, \hat{\lambda}_t) \end{bmatrix} \;=\; \begin{bmatrix} t\nabla f(x_t) + \nabla\phi(x_t) + \nabla g(x_t)\hat{\lambda}_t \\ g(x_t) \end{bmatrix}$$

Substitute $\nabla\phi(x) = \sum_{i=1}^{l} \frac{1}{-h_i(x)} \nabla h_i(x)$ from (8.88b) and define:

$$\lambda_t \;:=\; \frac{\hat{\lambda}_t}{t}, \qquad \mu_{t,i} \;:=\; \frac{1}{-th_i(x_t)} \;>\; 0$$

with $\mu_t := (\mu_{t,i}, i = 1, \ldots l)$. Then $F_t(x_t, \hat{\lambda}_t) = 0$ becomes:

$$\nabla f(x_t) + \nabla g(x_t)\lambda_t + \nabla h(x_t)\mu_t = 0, \qquad g(x_t) = 0 \tag{8.90a}$$

We have also, from the strict feasibility of x_t and the definition of μ_t,

$$h(x_t) < 0, \qquad \mu_t > 0, \qquad \mu_t^{\mathsf{T}} h(x_t) = -\frac{l}{t} \tag{8.90b}$$

This would be the KKT condition for the original problem (8.87) were the condition $\mu_t^{\mathsf{T}} h(x_t) = -l/t$ in (8.90b) replaced by complementary slackness $\mu_t^{\mathsf{T}} h(x_t) = 0$. Hence the KKT condition for Problem(t) is approximately the KKT condition for (8.87) for large t.

A popular interior point method, called the barrier method, is based on solving a sequence of the approximate problems (8.89) with increasing t until the approximation is sufficiently accurate. To describe it we first explain how to estimate the gap between the optimal value of the original problem (8.87) and the objective value of a solution of its approximation (8.89).

Suboptimality gap.

The theory of the barrier method is most complete for convex problems. For simplicity, we make the following assumptions:

C8.1: The original problem (8.87) is convex, i.e., f, h are convex functions and $g(x) = Ax - b$ for some $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$.
C8.2: For every $t > 0$ the approximate problem (8.89) has a unique primal solution $x(t)$ and the Newton-Raphson algorithm converges to $x(t)$.

We call the optimal solution $x(t)$ of (8.89) a *central point* and the set $\{x(t) : t > 0\}$ of central points the *central path*. The assumption of unique $x(t)$ for each $t > 0$ means that there is a unique central path. In this case the barrier method will use the Newton-Raphson algorithm to follow this unique path, as we will see.

Let f^* denote the optimal value of the original problem (8.87). The next result shows that a central point $x(t)$ is a feasible solution of (8.87) with a suboptimality gap that is strictly decreasing in $t > 0$. A certificate for the suboptimality gap is provided by a dual feasible solution for (8.87) associated with a central point $x(t)$.

Theorem 8.30 (Central point $x(t)$)**.** Under assumptions C8.1 and C8.2, for each $t > 0$:

1 The central point $x(t)$ is feasible for the original problem (8.87).
2 Its objective value is at most l/t away from the optimal value f^*, i.e.,

$$f(x(t)) - f^* \;\leq\; \frac{l}{t}$$

In particular $f(x(t)) \to f^*$ as $t \to \infty$.

Proof Since (8.89) is convex by assumption, the optimality of $x(t)$ means that the Slater condition is satisfied and hence strong duality holds and an optimal dual variable $\hat{\lambda}(t) \in \mathbb{R}^m$ exists by the Slater Theorem 8.17. Moreover the KKT Theorem 8.15 implies that $\big(x(t), \hat{\lambda}(t)\big)$ satisfies the KKT condition for (8.89):

$$t\nabla f\,(x(t)) + \nabla\phi\,(x(t)) + \frac{\partial g^{\mathsf{T}}}{\partial x}\,(x(t))\,\hat{\lambda}(t) \;=\; 0, \quad g(x(t)) \;=\; Ax(t) - b \;=\; 0 \tag{8.91a}$$

Because of the log barrier ϕ we must have $h_i(x(t)) < 0$ for all $i = 1, \ldots, l$. This means that $x(t)$ is also (strictly) feasible for the original problem (8.87), i.e., $x(t)$ satisfies

$$h(x(t)) < 0, \qquad g(x(t)) = Ax(t) - b = 0 \tag{8.91b}$$

We now use (8.91) to estimate the suboptimality gap of $x(t)$. Define the Lagrangian of the original problem (8.87)

$$L(x, \mu, \lambda) \;:=\; f(x) + \lambda^{\mathsf{T}} g(x) + \mu^{\mathsf{T}} h(x)$$

where the dual variables are $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^l_+$. Let the dual function be

$$d(\mu, \lambda) \;:=\; \min_{x \in \mathbb{R}^n} L(x, \mu, \lambda)$$

Define

$$\mu_i(t) \;:=\; \frac{1}{-t h_i(x(t))}, \qquad \lambda_i(t) \;:=\; \frac{\hat{\lambda}_i(t)}{t}$$

and let $\lambda(t) := (\lambda_i(t), i = 1, \ldots, m)$ and $\mu(t) := (\mu_i(t), i = 1, \ldots, l)$. Since $h_i(x(t)) < 0$, we have $\mu_i(t) > 0$ and hence $(\mu(t), \lambda(t))$ is dual feasible for (8.87). Dividing by t the first condition in (8.91a) and substituting (8.88b) we have

$$\nabla_x\, L(x, \mu(t), \lambda(t)) \;=\; \nabla f\,(x(t)) + \sum_{i=1}^{l} \mu_i(t) \nabla h_i(x(t)) + \frac{\partial g^{\mathsf{T}}}{\partial x}\,(x(t))\,\lambda(t) \;=\; 0$$

which implies that $x(t)$ minimizes $L(x, \mu(t), \lambda(t))$ over x. Hence the dual function of the original problem (8.87) evaluated at $(\mu(t), \lambda(t))$ is

$$d(\mu(t), \lambda(t)) = L(x(t), \mu(t), \lambda(t)) = f(x(t)) + \lambda^{\mathsf{T}}(t) g(x(t)) + \mu^{\mathsf{T}}(t) h(x(t)) \quad (8.92)$$

But $g(x(t)) = 0$ from (8.91) and $d(\mu(t), \lambda(t)) \leq f^*$ from weak duality for (8.87). Hence

$$f(x(t)) - f^* \leq -\sum_{i=1}^{l} \mu_i(t) h_i(x(t)) = \frac{l}{t}$$

from the definition of $\mu_i(t)$. □

The central point $x(t)$ and the dual variable $(\lambda(t), \mu(t))$ are primal-dual feasible for the original problem (8.87). By (8.92) their duality gap is exactly l/t:

$$d(\mu(t), \lambda(t)) - f(x(t)) = \mu^{\mathsf{T}}(t) h(x(t)) = \frac{l}{t}$$

Hence Problem(t) is an approximation of (8.87) both in the sense that, for large t, $(x(t), \lambda(t), \mu(t))$ is feasible and an approximate KKT point for (8.87) (see (8.90)) and that the suboptimality gap $f(x(t)) - f^*$ is small when the problem is convex.

The barrier method.

Theorem 8.30 says that, when (8.87) is convex, the central point $x(t)$ computed by the Newton-Raphson algorithm is feasible for the original problem (8.87) and its objective value $f(x(t))$ is at most l/t away from the optimal value f^*. This motivates the *barrier method*, also known as the *path-following method*, that solves Problem(t) in (8.89) to compute a central point $x(t)$, sequentially for increasing $t > 0$.

Specifically the barrier method solves a sequence of the approximate problems (8.89) with increasing $t > 0$, using the solution of the previous problem as the initial point for the current problem, as follows. Fix a parameter $\gamma > 1$ and solve Problem(t) in (8.89) with parameter t using the Newton-Raphson algorithm. Geometrically increase the parameter t by multiplying it by $\gamma > 1$ and solve (8.89) again starting from the solution of the previous problem. Repeat until t is sufficiently large so that the solution produced by Newton-Raphson is an accurate enough solution to the original problem (8.87). This method is described more precisely as Algorithm 3 when (8.87) is convex. Under C8.1 and C8.2, Algorithm 3 returns a feasible solution x that is ϵ-optimal, i.e., $f(x) - f^* \leq \epsilon$ by Theorem 8.30. The barrier method is also widely used for nonconvex problems even though convergence or optimality is not guaranteed. In the nonconvex case, a different stopping criterion based on the primal or dual iterates may be used.

In principle one can solve Problem(t) in (8.89) with parameter $t := l/\epsilon$ instead of solving a sequence of (8.89) with increasing t as in Algorithm 3. In practice this method does not work well unless the problem is small, the required accuracy ϵ is moderate and a good starting point is available. Therefore the barrier method is usually preferred.

Algorithm 3: Barrier method

Input: *strictly* feasible x, initial $t := t_0$, scaling factor $\gamma > 1$, tolerance ϵ.

Output: an ϵ-optimal solution x when (8.87) is convex.

1. **while** $t \leq \frac{l}{\epsilon}$ **do**
   1. Solve Problem(t) in (8.89) to compute $x(t)$ using the Newton-Raphson algorithm starting from x.
   2. $x \leftarrow x(t)$.
   3. $t \leftarrow \gamma t$.
2. **Return**: x.

Strictly feasible initial point.

Algorithm 3 requires an initial point x that is *strictly* feasible for the original problem (8.87), i.e. x satisfies

$$g(x) = 0, \qquad f(x) < 0$$

There are various methods to produce a strictly feasible point and we explain a simplest one (see [57, Chapter 11.4] for others). When necessary, such a method can be used to compute a strictly feasible x before the barrier method is executed. Starting from such an initial point, all subsequent iterates, across Problem(t) for different t, will remain strictly feasible because of the log barrier ϕ.

Consider the feasibility problem

$$\inf_{(x,s)\in\mathbb{R}^{n+1}} \quad s \qquad \text{s.t.} \qquad g(x) = 0,\ h_i(x) \leq s,\ i = 1,\ldots,l \tag{8.93}$$

where $s \in \mathbb{R}$, and as before, $g : \mathbb{R}^n \to \mathbb{R}^m$, and $h : \mathbb{R}^n \to \mathbb{R}^l$ are twice continuously differentiable. Suppose we are given an initial x_0 such that $g(x_0) = 0$ and $x_0 \in \mathrm{dom}(h_1) \cap \cdots \cap \mathrm{dom}(h_l)$, i.e., $h_i(x_0) < \infty$, $i = 1,\ldots,l$. Then (8.93) is feasible because (x_0, s_0) is a feasible point with $s_0 := \max_{i=1}^m f_i(x_0)$. Note that the feasible set is closed but not necessarily bounded and hence an optimal point of (8.93) may not exist or the infimum may not be attained by any x.

A strictly feasible point x for (8.87) exists if and only if the optimal value s^{opt} of (8.93) is strictly negative (can be $-\infty$); see Exercise 8.29. Solving (8.93) either produces such an x or proves that none exists, according to the sign of s^{opt}.

8.5.4 Dual and primal-dual gradient algorithms

Consider again the problem (8.87) reproduced here:

$$f^* \quad := \quad \min_{x\in\mathbb{R}^n} f(x) \qquad \text{s.t.} \qquad g(x) = 0,\ h(x) \leq 0 \tag{8.94}$$

where $f:\mathbb{R}^n\to\mathbb{R}$, $g:\mathbb{R}^n\to\mathbb{R}^m$, and $h:\mathbb{R}^n\to\mathbb{R}^l$ are twice continuously differentiable. The Lagrangian is

$$L(x,\lambda,\mu) \;:=\; f(x)+\lambda^{\mathsf{T}}g(x)+\mu^{\mathsf{T}}h(x)$$

where the dual variables are $\lambda\in\mathbb{R}^m$, $\mu\in\mathbb{R}^l_+$. Let the dual function be

$$d(\lambda,\mu) \;:=\; \min_{x\in\mathbb{R}^n} L(x,\lambda,\mu) \tag{8.95a}$$

and the dual problem be

$$\max_{(\lambda,\mu)\in\mathbb{R}^{m+l}} \quad d(\lambda,\mu) \qquad \text{s.t.} \qquad \mu\geq 0 \tag{8.95b}$$

The steepest descent algorithm (8.78) solves the primal problem (8.94) by iterating on the primal variable x and projects to the feasible set X in each iteration. This is sometimes referred to as a primal algorithm. A dual algorithm iterates on the dual variable (λ,μ) to solve the dual problem (8.95) instead, and a primal-dual algorithm iterates on both the primal and dual variables (x,λ,μ) to seek a saddle point of the Lagrangian L.

In this subsection we describe the dual algorithm and the primal-dual algorithm. Both algorithms produce a saddle point (x^*,λ^*,μ^*) of (8.94) when they converge, provided that the problem is convex (f,h are convex and $g(x)=Ax-b$). The Saddle-point Theorem 8.14 then implies that (x^*,λ^*,μ^*) is primal-dual optimal and strong duality holds.

Dual algorithm.

The key difference between (8.94) and (8.95a) is that the minimization over x is unconstrained in (8.95a). A dual algorithm can be used when the unconstrained minimization in (8.95a) is easy to solve, e.g., a minimizer can be obtained analytically. Given (λ,μ) let $x(\lambda,\mu)$ denote an unconstrained minimizer of L:

$$x(\lambda,\mu) \;\in\; \arg\min_{x\in\mathbb{R}^n} L(x,\lambda,\mu) \tag{8.96a}$$

When L is convex in x, $x(\lambda,\mu)$ is a solution of $\nabla_x L(x,\lambda,\mu)=0$. Then a dual algorithm is a steepest ascent algorithm for solving the dual problem (8.95):

$$\lambda(t+1) \;=\; \lambda(t)+\gamma_\lambda\nabla_\lambda L(x(\lambda(t),\mu(t)),\lambda(t),\mu(t)) \tag{8.96b}$$

$$\mu(t+1) \;=\; \big[\mu(t)+\gamma_\mu\nabla_\mu L(x(\lambda(t),\mu(t)),\lambda(t),\mu(t))\big]^+ \tag{8.96c}$$

where $(\gamma_\lambda,\gamma_\mu)\in\mathbb{R}^2$ are positive constant stepsizes and $[y]^+:=\max\{0,y\}$ componentwise for a vector y. For convex problems, if (8.96) converges and produces a dual optimum (λ^*,μ^*) of (8.95) then $x(\lambda^*,\mu^*)$ will be optimal for (8.94) (Saddle-point Theorem 8.14). Variants of the steepest ascent algorithm (8.96) can be obtained by using iteration-dependent stepsizes $(\gamma_\lambda(t),\gamma_\mu(t))$ or scaling matrices $\gamma_\lambda\in\mathbb{R}^{m\times m}$, $\gamma_\mu\in\mathbb{R}^{l\times l}$.

An important application of the dual algorithm is in distributed computation. When

the problem (8.94) has a certain decentralized structure, e.g., if the cost function $f(x) = \sum_i f_i(x_i)$ is separable in i and the constraints are affine, the dual algorithm decomposes naturally into a distributed method, as the next example shows.

Example 8.15 (Distributed dual algorithm)**.** Consider the utility maximization:

$$\max_{x \in \mathbb{R}^n} \sum_i U_i(x_i) \qquad \text{s.t.} \qquad Rx \le c \tag{8.97}$$

where $U_i : \mathbb{R} \to \mathbb{R}$ for $i = 1, \ldots, n$, are continuously differentiable and strictly concave increasing utility functions, $R \in \{0,1\}^{l \times n}$ and $c \in \mathbb{R}^l$. The Lagrangian is

$$L(x, \mu) := \sum_i U_i(x_i) - \mu^{\mathsf{T}}(Rx - c), \qquad \mu \in \mathbb{R}^l_+$$

and the dual function is

$$d(\mu) := \max_{x \in \mathbb{R}^n} L(x, \mu) = \sum_{i=1}^{n} \max_{x_i \in \mathbb{R}} \left(U_i(x_i) - x_i \sum_{j=1}^{l} R_{ji}\mu_j \right) + \mu^{\mathsf{T}} c$$

i.e., the unconstrained maximization over the vector x decomposes into a distributed maximization over individual components x_i. Given μ, the distributed maximization over x_i can be solved in closed form:

$$x_i(\mu) := U_i'^{-1}\left(\sum_{j=1}^{l} R_{ji}\mu_j \right) =: U_i'^{-1}\left(p_i(\mu)\right), \qquad i = 1, \ldots, n \tag{8.98a}$$

where $p_i(\mu) := \sum_j R_{ji}\mu_j$, U_i' is the derivative of U_i and $U_i'^{-1}$ is its inverse (which exists since U_i is strictly concave). We write this in vector form as

$$x(\mu) := (\nabla_x U)^{-1}\left(R^{\mathsf{T}}\mu\right)$$

The strict concavity of U_i implies that the maximizer $x_i(\mu)$ is unique and hence Danskin's Theorem 8.21 implies that the dual function $d(\mu)$ is differentiable with

$$\nabla_\mu d(\mu) = \nabla_\mu L(x(\mu), \mu) = -(Rx(\mu) - c)$$

Then the dual algorithm for solving the dual problem $\min_{\mu \ge 0} d(\mu)$ is

$$\mu(t+1) = \left[\mu(t) - \gamma \nabla_\mu d(\mu(t))\right]^+ = \left[\mu(t) + \gamma(Rx(\mu(t)) - c)\right]^+$$

Therefore the dual update also decomposes into a distributed computation, given x:

$$\mu_j(t+1) = \left[\mu_j(t) + \gamma\left(y_j(t) - c_j\right)\right]^+, \qquad j = 1, \ldots, l \tag{8.98b}$$

where $y_j(t) := \sum_i R_{ji}x_i(\mu(t))$ and $x_i(\mu(t))$ is given by (8.98a).

Hence the dual algorithm for (8.97) is the distributed algorithm given by (8.98). This is a model of Internet congestion control algorithm. In this application, each i represents a sender that wishes to send its data packets at a rate x_i packets/sec that is as high as possible and each j represents a link (buffer) in the network whose processing speed is limited to c_j packets/sec. The matrix R is a routing matrix that

specifies the path in the network of each sender i from its source node to its destination node, consisting of links j with $R_{ji} = 1$. The optimal sending rate vector x is one that maximizes the aggregate utility $\sum_i U_i(x_i)$ subject to the constraint that the input rates $y_j := \sum_i R_{ji} x_i$ at optimality do not exceed link capacities c_j for every link j.. The Lagrange multiplier $\mu_j \geq 0$ can be interpreted as a congestion price at link j and $p_i := \sum_j R_{ji}\mu_j$ is the end-to-end congestion price (i.e., sum of the link congestion prices μ_j along i's path) observed by sender i. Then the algorithm (8.98) specifies the local decision by each sender i and link j: sender i sets its sending rate to $x_i(t)$ in (8.98a) based on the end-to-end congestion price $p_i(t)$ it observes locally, and link j updates its congestion price $\mu_j(t)$ according to (8.98b) based on the local input flow rate $y_j(t)$ at link j. In particular the congestion price $\mu_j(t)$ is incremented if the input flow rate $y_j(t)$ at link j exceeds the link capacity c_j and decremented otherwise. □

Primal-dual algorithm.

When the unconstrained minimization over x in (8.96a) is difficult to solve, we can replace (8.96a) by iteration on the primal variable x:

$$x(t+1) = x(t) - \gamma_x \nabla_x L(x(t), \lambda(t), \mu(t)) \tag{8.99a}$$

$$\lambda(t+1) = \lambda(t) + \gamma_\lambda \nabla_\lambda L(x(t), \lambda(t), \mu(t)) \tag{8.99b}$$

$$\mu(t+1) = \left[\mu(t) + \gamma_\mu \nabla_\mu L(x(t), \lambda(t), \mu(t))\right]^+ \tag{8.99c}$$

where $[y]^+ := \max\{0, y\}$ componentwise for a vector y. This is called a primal-dual algorithm or a saddle point algorithms. It seeks a saddle point of the Lagrangian L through steepest descent in the primal variable $x(t)$ and steepest ascent in the dual variable $(\lambda(t), \mu(t))$. For convex problems, if (8.99) converges and produces a saddle point (x^*, λ^*, μ^*) of L then it is primal and dual optimal for (8.94)(8.95) and strong duality holds (Saddle-point Theorem 8.14).

For Example 8.15 the primal-dual version of (8.98) replaces (8.98a) by

$$x_i(t+1) = x_i(t) + \gamma_x \left(U_i'(x(t)) - p_i(\mu(t))\right), \qquad i = 1, \ldots, n$$

where $p_i(\mu(t)) := \sum_{j=1}^{l} R_{ji}\mu_j(t)$, i.e., a sender increments its sending rate $x_i(t)$ if its marginal utility $U_i'(x(t))$ exceeds its end-to-end price $p_i(t)$, and decrements it otherwise. It remains a distributed algorithm.

8.5.5 Alternating direction method of multipliers (ADMM)

Consider

$$\min_{x \in \mathbb{R}^{n_1},\ y \in \mathbb{R}^{n_2}} \quad f(x) + g(y) \qquad \text{s.t.} \qquad x \in X,\ y \in Y \tag{8.100a}$$

$$Ax + By = c \tag{8.100b}$$

where $f:\mathbb{R}^{n_1}\to\mathbb{R}, g:\mathbb{R}^{n_2}\to\mathbb{R}, X\subseteq\mathbb{R}^{n_1}, Y\subseteq\mathbb{R}^{n_2}, A\in\mathbb{R}^{m\times n_1}, B\in\mathbb{R}^{m\times n_2}$, and $c\in\mathbb{R}^m$. The key feature of (8.100) is that the cost function and the possibly nonlinear constraints in (8.100a) are separable in x,y. The coupling between x and y is only through the linear constraint (8.100b). This is similar to the problem structure in Example 8.15 and therefore a dual algorithm can be applied to obtain a distributed solution. Dual algorithm however often converges slowly because the Lagrangian is affine, as opposed to strictly concave, in the dual variable.

The alternating direction method of multipliers (ADMM) combines the distributed structure of dual decomposition with better convergence properties of augmented Lagrangian methods. Specifically define the augmented Lagrangian function that relaxes the coupling constraint (8.100b): for $(x,y)\in\mathbb{R}^{n_1+n_2}$ and $\lambda\in\mathbb{R}^m$,

$$L_\rho(x,y,\lambda) \;:=\; f(x)+g(y)+\lambda^{\mathsf{T}}(Ax+By-c)+\frac{\rho}{2}\|Ax+By-c\|_2^2,$$

where $\rho\geq 0$ is a parameter that controls the degree of augmentation (ADMM reduces to dual decomposition when $\rho=0$). The ADMM algorithm is

$$x(t+1)=\arg\min_{x\in X} L_\rho(x,\ y(t),\ \lambda(t)) \tag{8.101a}$$

$$y(t+1)=\arg\min_{y\in Y} L_\rho(x(t+1),\ y,\ \lambda(t)) \tag{8.101b}$$

$$\lambda(t+1)=\lambda(t)\ +\ \rho\,(Ax(t+1)+By(t+1)-c) \tag{8.101c}$$

The update (8.101c) is $\lambda(t+1)=\lambda(t)+\rho\nabla_\lambda L_\rho(x(t+1),y(y+1),\lambda)$. Hence, compared with (8.96), (8.101) is a dual algorithm with stepsize ρ and two differences: it uses an augmented Lagrangian function L_ρ for better convergence properties, and the subproblem (8.101b) and the dual update (8.101c) use the latest available data, $x(t+1)$ and $(x(t+1),y(t+1))$ respectively (this is called one pass of a Gauss-Seidel method).

Suppose f,g are convex and continuously differentiable on $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$ respectively and X,Y are convex sets. If the ADMM algorithm converges to a fixed point $(x^*,y^*,\lambda^*)\in X\times Y\times\mathbb{R}^m$ of (8.101), then

$$L_\rho(x^*,y^*,\lambda^*)\ \leq\ L_\rho(x,y^*,\lambda^*), \qquad x\in X \tag{8.102a}$$

$$L_\rho(x^*,y^*,\lambda^*)\ \leq\ L_\rho(x^*,y,\lambda^*), \qquad y\in Y \tag{8.102b}$$

$$Ax^*+By^*-c\ =\ 0 \tag{8.102c}$$

The (un-augmented) Lagrangian of (8.100) is L_0 with $\rho:=0$. We now show that (x^*,y^*,λ^*) is a saddle point of L_0 and hence is primal-dual optimal for (8.100).

By (8.102c), (x^*,y^*) is feasible and this has two implications. First it means $\nabla_\lambda L_0(x^*,y^*,\lambda^*)=0$ and hence y^* maximizes $L_0(x^*,y^*,\cdot)$ given (x^*,y^*), i.e., $L_0(x^*,y^*,\lambda)\leq L_0(x^*,y^*,\lambda^*)$ for all λ. Second it implies $L_\rho(x^*,y^*,\lambda^*)=L_0(x^*,y^*,\lambda^*)=f(x^*)+g(y^*)$ and hence (8.102a)(8.102b) become:

$$L_0(x^*,y^*,\lambda^*)-L_0(x,y^*,\lambda^*)\ \leq\ 0, \qquad x\in X \tag{8.103a}$$

$$L_0(x^*,y^*,\lambda^*)-L_0(x^*,y,\lambda^*)\ \leq\ 0, \qquad y\in Y \tag{8.103b}$$

This turns out to be equivalent to $L_0(x^*,y^*,\lambda^*) \le L_0(x,y,\lambda^*)$ that is required for (x^*,y^*,λ^*) to be a saddle point. Notice

$$L_0(x,y^*,\lambda^*) - L_0(x,y,\lambda^*) \;=\; L_0(x^*,y^*,\lambda^*) - L_0(x^*,y,\lambda^*) \;\le\; 0, \quad x \in X, y \in Y$$

where the inequality follows from (8.103b), i.e., (8.103b) implies that y^* minimizes $L_0(x,y,\lambda^*)$ over $y \in Y$ for *any* fixed $x \in X$. We therefore have

$$\begin{aligned} L_0(x^*,y^*,\lambda^*) - L_0(x,y,\lambda^*) &= \big(L_0(x^*,y^*,\lambda^*) - L_0(x,y^*,\lambda^*)\big) + \big(L_0(x,y^*,\lambda^*) - L_0(x,y,\tilde{\lambda})\big) \\ &\le 0, \qquad x \in X, y \in Y \end{aligned}$$

We conclude

$$L_0(x^*,y^*,\lambda) \;\le\; L_0(x^*,y^*,\lambda^*) \;\le\; L_0(x,y,\lambda^*), \qquad (x,y) \in X \times Y,\ \lambda \in \mathbb{R}^m$$

i.e., (x^*,y^*,λ^*) is a saddle point of L_0 and hence primal-dual optimal for (8.100) (Saddle-point Theorem 8.14).

8.5.6 Branch and bound

Branch and bound (B&B) methods are algorithms for solving optimization problems that involve integer variables, such as an integer linear program (ILP) where all variables are integers:

$$\min_{x \in \mathbb{N}^n} \; c^\mathsf{T} x \qquad \text{s.t.} \qquad Ax \le b$$

$\mathbb{N}$ being the set of integers, or a mixed integer linear program (MILP) where some of the variables are integers:

$$f^* \;:=\; \min_{x \in \mathbb{R}^n} \; c^\mathsf{T} x \qquad \text{s.t.} \qquad Ax \le b,\ \ x_j \in \mathbb{N}\ \forall j \in J \tag{8.104}$$

$J \subseteq \{1,\dots,n\}$ being a subset of variable indices.[9] Clearly a MILP reduces to a linear program if $J = \emptyset$ and an ILP if $J = \{1,\dots,n\}$.

MILP is generally NP-hard. We present three methods for solving MILP. The security constrained unit commitment problem (6.47) of Chapter 6.4.5 is a mixed integer linear program and can be solved using these methods.

LP relaxation.

The simplest method is to relax the integral constraint and solve the resulting linear program instead of (8.104):

$$f^*_{\text{lp}} \;:=\; \min_{x \in \mathbb{R}^n} \; c^\mathsf{T} x \qquad \text{s.t.} \qquad Ax \le b \tag{8.105}$$

This enlarges the feasible set and therefore provides a lower bound on the original MILP (8.104), i.e., $f^*_{\text{lp}} \le f^*$.

[9] The solution methods of this subsection extends directly mixed integer nonlinear programs.

Dual relaxation.

Another relaxation is to solve the dual of the MILP. The Lagrangian of (8.104) is

$$L(x,\mu) := c^\mathsf{T}x + \mu^\mathsf{T}(Ax-b), \qquad x \in \mathbb{R}^n, \mu \in \mathbb{R}^m$$

the dual function is

$$d(\mu) := \min_{x\in\mathbb{R}^n} L(x,\mu) = \begin{cases} -b^\mathsf{T}\mu & \text{if } (c+A^\mathsf{T}\mu)^\mathsf{T} = 0 \\ -\infty & \text{otherwise} \end{cases}$$

and the dual problem is

$$d^*_{\text{milp}} := \max_{\mu\geq 0} -b^\mathsf{T}\mu \quad \text{s.t.} \quad (c+A^\mathsf{T}\mu)^\mathsf{T} = 0 \tag{8.106}$$

The Lagrangian and hence the dual problem (8.106) are the same as those of the linear program (8.105). If MILP (8.104) is feasible, so is LP (8.105). Theorem 8.23 on LP duality then implies that both the primal and dual optima of (8.105) are attained and strong duality holds, i.e., $d^*_{\text{milp}} = f^*_{\text{lp}} \leq f^*$. Hence dual relaxation also provides a lower bound on the MILP (8.104).

Branch and bound.

While LP relaxation (8.105) and dual relaxation (8.106) are linear programs that provide (the same) lower bound to MILP (8.104), branch and bound methods are exponential algorithms that solve for an optimal mixed integer solution of (8.104). There are many variants and the main idea is as follows.

We can treat MILP (8.104) as searching for a minimum over the feasible set $X_0 := \{x : Ax \leq b,\ x_j \in \mathbb{N}\ \forall j \in J\}$. Branch and bound iteratively divides the feasible set X_0 into subsets and search for a minimum over each of these subsets to eventually arrive at a global optimum $x^* \in X_0$ with $c^\mathsf{T}x^* = f^*$. We can represent the process as iteratively constructing a search tree, starting with X_0 at its root and, in each iteration, either grow the search tree by splitting a node $X_i \subseteq X_0$ into new child nodes (i.e., partition the set X_i into subsets) or prune the node X_i (i.e., stop further partitioning X_i). To determine if X_i will create new branches or if it will be pruned, a LP relaxation of (the subproblem defined by the feasible subset) X_i is solved (bounding), resulting in one of three outcomes:

1 X_i contains no optimal solution of (8.104), in which case X_i will be pruned;
2 A feasible solution in X_0 is found which is a candidate optimal solution of (8.104), in which case X_i will be pruned (i.e., X_i is not further partitioned);
3 Otherwise, X_i may or may not contain an optimal solution of (8.104) amd X_i is further partitioned (branching).

This branch and bound procedure repeats until every leaf node of the search tree is pruned, and the candidate solution with the minimum value is a global optimum of

MILP (8.104). In summary the key components of a branch and bound method are a bounding function that computes a lower bound on a subproblem defined by X_i and a branching function that determines how to split node X_i into child nodes if X_i is not pruned.

This method is illustrated in the following example that uses LP relaxation as the bounding function.

Example 8.16 (Branch and bound)**.** Consider the following integer linear program:

$$f^* := \min_{x \in \mathbb{N}^2} x_1 - 4x_2 \quad \text{s.t.} \quad x \in X_0 \tag{8.107}$$

where $X_0 := \{x \in \mathbb{N}^2 : -x_1 + 3x_2 \leq 0, x_1 + 3x_2 \leq 9, x \geq 0\}$. Let $f(x) := x_1 - 4x_2$.

1 *Initialization.*
- Let the global upper bound be $f^{\max} := f(0) = 0$. As the algorithm proceeds, $f^{\max}$ will be updated but remain a global upper bound throughout, i.e., $f^* \leq f^{\max}$.
- Let Q denote a queue of leaf nodes (feasible sets of subproblems) in the search tree and initialize it to $Q := \{X_0\}$.

2 *Bounding and branching:* X_0.
- Remove X_0 from Q.
- Its LP relaxation is $\min_{x \in \mathbb{R}^2} f(x)$ s.t. $x \in X_0^{\text{lp}} := \{x \in \mathbb{R}^2 : -x_1 + 3x_2 \leq 0, x_1 + 3x_2 \leq 9, x \geq 0\}$ with a unique minimizer $x_0^{\text{lp}} := (4.5, 1.5)$.
- Let a lower bound of X_0 be $f_{X_0}^{\min} := f(x_0^{\text{lp}}) = -1.5$.
- Since $f_{X_0}^{\min} < f^{\max}$ and $x_{0,1}^{\text{lp}} = 4.5$ is fractional, we partition X_0 into two subsets:

$$X_{11} := X_0 \cap \{x \in \mathbb{N}^2 : x_1 \leq 4\}, \qquad X_{12} := X_0 \cap \{x \in \mathbb{N}^2 : x_1 \geq 5\}$$

- Add X_{11} and X_{12} to Q.

3 *Bounding and branching* X_{11}. Remove X_{11} from Q. Its LP relaxation is $\min_{x \in \mathbb{R}^2} f(x)$ s.t. $x \in X_{11}^{\text{lp}} := X_0^{\text{lp}} \cap \{x \in \mathbb{R}^2 : x_1 \leq 4\}$ with a unique minimizer $x_{11}^{\text{lp}} := (4, 4/3)$. Let a lower bound on X_{11} (not necessarily a lower bound on f^*) be $f_{X_{11}}^{\min} := f(x_{11}^{\text{lp}}) = -4/3$.

Since $f_{X_{11}}^{\min} < f^{\max}$ and $x_{11,2}^{\text{lp}} = 4/3$ is fractional, we partition X_{11} into two subsets:

$$X_{21} := X_{11} \cap \{x \in \mathbb{N}^2 : x_2 \leq 1\}, \qquad X_{22} := X_{11} \cap \{x \in \mathbb{N}^2 : x_2 \geq 2\}$$

Add X_{21} and X_{22} to Q.

4 *Bounding and branching:* X_{12}. Remove X_{12} from Q. Its LP relaxation is $\min_{x \in \mathbb{R}^2} f(x)$ s.t. $x \in X_{12}^{\text{lp}} := X_0^{\text{lp}} \cap \{x \in \mathbb{R}^2 : x_1 \geq 5\}$ with a unique minimizer $x_{12}^{\text{lp}} := (5, 4/3)$. Let the local lower bound on X_{12} be $f_{X_{12}}^{\min} := f(x_{12}^{\text{lp}}) = -1/3$. This is an example of the lower bound obtained in a subproblem being local and may not bound f^*: for X_{12}, the lower bound is $f_{X_{12}}^{\min} = -1/3$ but, as we will show below, $f^* = -1$.

Since $f^{\min}_{X_{12}} < f^{\max}$ and $x^{\text{lp}}_{12,2} = 4/3$ is fractional, we partition X_{12} into two subsets:

$$X_{23} := X_{12} \cap \{x \in \mathbb{N}^2 : x_2 \le 1\}, \qquad X_{24} := X_{12} \cap \{x \in \mathbb{N}^2 : x_2 \ge 2\}$$

Add X_{23} and X_{24} to Q.

5 *Bounding and pruning*. Similarly, for each node $X_{21}, X_{22}, X_{23}, X_{24}$ in Q, LP relaxation computes a lower bound, as illustrated in Figure 8.17.

- For X_{21}, the minimizer is an integer solution $x^{\text{lp}}_{21} = (3,1)$ with optimal value $f^{\min}_{X_{21}} := f(x^{\text{lp}}_{21}) = -1$. Reduce the global upper bound to $f^{\max} := -1$. Since $f^{\min}_{X_{12}} = f^{\max}$, x^{lp}_{21} is currently the best candidate optimal solution of (8.107) and X_{21} is pruned (i.e., not further partitioned).
- For X_{23}, the minimizer is an integer solution $x^{\text{lp}}_{23} = (5,1)$ with optimal value $f^{\min}_{X_{23}} := f(x^{\text{lp}}_{21}) = 1$. Since $f^{\min}_{X_{12}} > f^{\max}$, X_{23} contains no optimal solution of (8.107) and it is pruned.
- The subproblems for both X_{22} and X_{24} are infeasible and pruned.

6 Since Q is empty, the global optimum of (8.107) is $x^{\text{lp}}_{21} = (3,1)$ and the optimal value is $f^* = f^{\max} = -1$. □

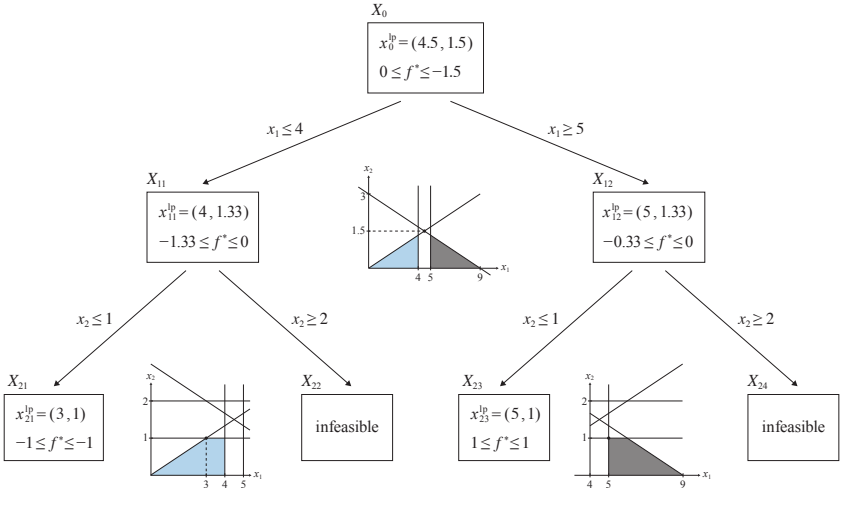


Figure 8.17 Example 8.16. Optimal value $f^* = -1$ which is upper bounded by $f^{\max}$ throughout the algorithm. The shaded areas are the feasible sets of various subproblems. (May 18, 2025: X_2 box: $f^{\min}_0 = -1.5 < 0 = f^{\max} \ge f^*$; X_{11} box: $f^{\min}_{11} = -1.33 < 0 = f^{\max} \ge f^*$; X_{12} box: $f^{\min}_{12} = -0.33 < 0 = f^{\max} \ge f^*$; X_{21} box: $f^{\min}_{21} = -1 = f^{\max} \ge f^*$; X_{23} box: $f^{\min}_{23} = 1 > -1 = f^{\max} \ge f^*$;)

We summarize the branch and bound process illustrated in Example 8.16.

1 *Initialization*.

- Compute the currently best known upper bound $f^{\max}$ on the optimal value f^* of MILP (8.104), e.g., from a known $x_0 \in X_0$ or set $f^{\max} := \infty$. It is a global bound in the sense that $f^* \le f^{\max}$ throughout the algorithm as $f^{\max}$ is updated.
- Initialize the queue of leaf nodes (subproblem feasible sets) in the search tree to $Q := \{X_0\}$.

2 *Bounding*.
- Remove a node X from Q. This defines a MILP subproblem whose feasible set is X.
- Compute a minimizer x^{lp} of the LP relaxation of (the subproblem defined by) X. Denote its optimal value by $f_X^{\min}$. It is a *local* lower bound on X and may not be a lower bound on f^*.

3 *Branching or pruning*.
- If $f_X^{\min} \ge f^{\max}$, then X is pruned. This includes the case where the LP relaxation of X is infeasible ($f_X^{\min} = \infty$).
- If $f_X^{\min} < f^{\max}$ and x^{lp} is a mixed integer solution in X_0, then reduce the global bound to $f^{\max} := f_X^{\min}$. An optimal solution of the subproblem defined by X is found (which is a candidate solution of (8.104)).
- If $f_X^{\min} < f^{\max}$ but x^{lp} is fractional, then a branching rule partitions X into two or more subsets X_i.
- Add each X_i to Q.

4 *Iterate*.
- If Q is empty then the optimal value of MILP (8.104) is $f^{\max}$ and a global optimum is the mixed integer solution found in Step 3 that attained $f^{\max}$.
- Otherwise, goto Step 2.

There are numerous variants of branch and bound methods. They differ on the bounding function in Step 2, rules for pruning and branching in Step 3, and the rule for selecting the next node X in Q to process. In addition, valid inequalities, called cuts, can be added in the branching step to further prune the search space.

8.5.7 Benders decomposition

Consider

$$\min_{x,y} \; f(x,y) \quad \text{s.t.} \quad F(x,y) \le 0,\; x \in X,\; y \in Y \tag{8.108}$$

where $f : \mathbb{R}^n \times \mathbb{R}^l \to \mathbb{R}$, $F : \mathbb{R}^n \times \mathbb{R}^l \to \mathbb{R}^m$, and $X \subseteq \mathbb{R}^n$ and $Y \subseteq \mathbb{R}^l$ are nonempty. The key feature of (8.108) studied in this subsection is that given a feasible y, often called the complicating variable, the minimization over x is a much simpler problem than solving (8.108) directly over (x, y). For example Y is the set of integers, y is a discrete vector in unit commitment, and (8.108) is a mixed integer linear program. An effective

solution approach is then to decompose (8.108) into a master problem that computes a minimizer y:

$$\min_{y} \; g(y) \qquad \text{s.t.} \qquad y \in G \tag{8.109a}$$

where the feasible set G is

$$G \;:=\; \{y \in Y : F(x,y) \le 0 \text{ for some } x \in X\} \tag{8.109b}$$

and the cost function $g : \mathbb{R}^l \to \mathbb{R}$ is the optimal value of minimization over x given y:

$$g(y) \;:=\; \min_{x} \; f(x,y) \quad \text{s.t.} \quad F(x,y) \le 0,\; x \in X \tag{8.109c}$$

The minimization (8.109c) over x may either decompose further into independent subproblems each involving a different subvector of x, or have a simple structure, e.g., is a convex program. In the former case the subproblem (8.109c) is decentralized and can be solved in parallel. In this subsection we study the latter case and present Benders decomposition for its solution when (8.109c) is a linear program over x, given y. See [59] for generalized Benders decomposition when (8.109c) is a convex program, i.e., for each $y \in Y$, $f(x,y)$ and $G(x,y)$ are convex functions in x and X is a convex set.

We start with an example.

Example 8.17 (Unit commitment)**.** The unit commitment problem (6.4) of Chapter 6.2.1 is a mixed integer nonlinear program and can be solved using Benders decomposition. If the constraint functions $g_t, \tilde{g}$ in the real-time dispatch problem (6.4c)(6.4d)(6.4e) are affine and $h_t, \tilde{h}$ are convex, then the subproblem (8.109c) is a convex program. If $h_t, \tilde{h}$ are also affine then it is a linear program.

The security constrained unit commitment problem (6.47) of Chapter 6.4.5 is a mixed integer linear program and can be solved to optimality in finitely (but potentially exponentially) many steps using Benders algorithm, as we now describe. □

Consider

$$\min_{x,y} \; c^{\mathsf{T}}x + f(y) \quad \text{s.t.} \quad Ax + F(y) \le b,\; x \ge 0,\; y \in Y \tag{8.110}$$

where $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $Y \subseteq \mathbb{R}^l$ is nonempty, $f : Y \to \mathbb{R}$, and $F : Y \to \mathbb{R}^m$. Decompose (8.110) into a master problem:

$$\min_{(y_0,y) \in \mathbb{R}^{1+l}} \; y_0 \qquad \text{s.t.} \qquad (y_0,y) \in G' \tag{8.111a}$$

and a linear program over x given a $(y_0,y) \in \mathbb{R}^{1+l}$ with $y \in Y$:

$$g(y_0,y) \;:=\; \min_{x \in \mathbb{R}^n} \; y_0 \qquad \text{s.t.} \qquad \tilde{A}x \le \tilde{b}(y_0,y),\; x \ge 0 \tag{8.111b}$$

where $G' := \{y \in Y : \tilde{A}x \le \tilde{b}(y_0,y) \text{ for some } x \ge 0\}$ and

$$\tilde{A} \;:=\; \begin{bmatrix} c^{\mathsf{T}} \\ A \end{bmatrix}, \qquad \tilde{b}(y_0,y) \;:=\; \begin{bmatrix} y_0 - f(y) \\ b - F(y) \end{bmatrix} \tag{8.111c}$$

We first reformulate the subproblems in (8.111) into a more convenient form.

The linear program (8.111b) is a feasibility problem. A variant of the Farkas lemma from Exercise 8.14 implies that exactly one of the following holds:

1 (8.111b) *is feasible*. There exists an $x \geq 0$ such that $\tilde{A}x \leq \tilde{b}(y_0, y)$, i.e.,
2 (8.111b) *is infeasible*. There exists an $(\mu_0, \mu) \geq 0$ such that $(\mu_0, \mu)^\mathsf{T}\tilde{A} \geq 0$ and $(\mu_0, \mu)^\mathsf{T}\tilde{b}(y_0, y) < 0$, i.e.,

$$\mu_0 c + \mu^\mathsf{T} A \;\geq\; 0 \quad \text{and} \quad \mu_0(y_0 - f(y)) + \mu^\mathsf{T}(b - F(y)) \;<\; 0 \tag{8.112}$$

To shorten notation, define

$$\begin{aligned} C \;&:=\; \{(\mu_0, \mu) \in \mathbb{R}^{1+m} : \mu_0 c + \mu^\mathsf{T} A \geq 0, (\mu_0, \mu) \geq 0\} \\ g(y_0, y; \mu_0, \mu) \;&:=\; u_0(y_0 - f(y)) + \mu^\mathsf{T}(b - F(y)) \end{aligned}$$

Note that C is nonempty (e.g. $0 \in C$); moreover the first condition in (8.112) does not depend on (y_0, y). Therefore, for each $(\mu_0, \mu) \in C$, the linear program (8.111b) defined by each (y_0, y) with $y \in Y$ is infeasible if and only if $g(y_0, y; \mu_0, \mu) < 0$. Consider the two subproblems:

$$\min_{(y_0, y)} \; y_0 \qquad \text{s.t.} \qquad (y_0, y) \in G \tag{8.113a}$$

$$\min_{x \geq 0} \; c^\mathsf{T} x \qquad \text{s.t.} \qquad Ax \leq b - F(y) \tag{8.113b}$$

where

$$G \;:=\; \bigcap_{(\mu_0, \mu) \in C} \{(y_0, y) \in \mathbb{R}^{1+l} : g(y_0, y; \mu_0, \mu) \geq 0,\; y \in Y\} \tag{8.113c}$$

i.e., $(y_0, y) \in G$ if and only if $y \in Y$ and $g(y_0, y; \mu_0, \mu) \geq 0$ for every $(\mu_0, \mu) \in C$. The feasible set G is empty if there exists a $(\mu_0, \mu) \in C$ such that $g(y_0, y; \mu_0, \mu) < 0$ for all (y_0, y) with $y \in Y$.

The following result is the basis of Benders decomposition (its proof is Exercise 8.30). It establishes the equivalence of (8.113) and (8.111), and hence (8.110). It suggests solving (8.110) by iteratively computing a solution (y_0^*, y^*) of (8.113a) and then a solution x^* of the linear program (8.113b). Then (x^*, y^*) is an optimum of (8.110) with optimal value y_0^*.

Theorem 8.31. 1 (8.110) is infeasible if and only if (8.113a) is infeasible.
2 Suppose (x^*, y^*) is an optimal solution of (8.110). Then
- (y_0^*, y^*) is optimal for (8.113a) where $y_0^* := c^\mathsf{T} x^* + f(y^*)$.
- x^* is optimal for the linear program (8.113b) with $y = y^*$.

3 Conversely suppose (y_0^*, y^*) is an optimal solution of (8.113a). Then
- (8.113b) with $y = y^*$ has an optimum x^* with optimal value $c^\mathsf{T} x^* = y_0^* - f(y^*)$.
- (x^*, y^*) is optimal for (8.110) with optimal value y_0^*.

The theorem also implies that (8.110) is feasible but does not attain optimality if and only if (8.113a) is feasible but does not attain optimality.

The challenge of this solution approach is solving (8.113a) which can be nonconvex and/or mixed-integer. Moreover it is not obvious how to compute G. Benders decomposition provides a finite procedure to build up G and solve (8.110) systematically. The basic idea is to start with a relaxation of (8.113a) with a superset $\overline{G}$ of G as its feasible set. The solution of the relaxation defines a linear program (8.113b). Instead of solving (8.113b), we solve its dual. The solution of the dual identifies either an additional constraint to add to $\overline{G}$ and the cycle repeats, or an optimal solution of (8.110) and the procedure terminates. This procedure does not avoid the difficult step of solving a possibly nonconvex and/or mixed-integer program (8.113a) but it solves a sequence of this problem starting from a simple feasible set, adding a constraint in each iteration that *strictly* tightens the relaxation, and terminates after a finitely many iterations. When it terminates, it either identifies a finite optimal solution of (8.110) or determines that none exists (i.e., (8.110) is either infeasible or feasible but unbounded).

We next describe this procedure in more detail under the assumptions:

C8.3: Y is compact.
C8.4: $f(y)$ and $F(y)$ are continuous on an open set $\overline{Y} \subseteq \mathbb{R}^l$ containing Y.

For each $y \in Y$, the (partial) Lagrangian of (8.113a) is

$$L(x,\mu;y) := c^\mathsf{T}x + \mu^\mathsf{T}(Ax - b + F(y)) = (F(y) - b)^\mathsf{T}\mu + (c + A^\mathsf{T}\mu)^\mathsf{T}x, \quad x \geq 0, \mu \geq 0$$

Hence the dual of (8.113a) is

$$\max_{\mu \geq 0} \ (F(y) - b)^\mathsf{T}\mu \qquad \text{s.t.} \qquad c + A^\mathsf{T}\mu \geq 0 \tag{8.114}$$

Note that the feasibility of (8.114) does not depend on y, only its objective function does. It can be shown that, under C8.3 and C8.4, (8.114) is infeasible if and only if y_0 has no lower bound on G, i.e., if and only if the optimal value of (8.113a) is $-\infty$ (Exericse 8.31).

Let Q be any subset of C and define a relaxation $G(Q)$ of G:

$$G(Q) := \bigcap_{(\mu_0,\mu) \in Q} \{(y_0, y) \mid g(y_0, y; \mu_0, \mu) \geq 0, \ y \in Y\} \tag{8.115a}$$

As we will explain below, Benders algorithm identifies a new element (μ_0', μ') to add to Q in each iteration, introducing the additional constraint $g(y_0, y; \mu_0', \mu') \geq 0$ on (y_0, y) that strictly tightens the relaxation $G(Q)$, until an optimal solution of (8.110) is found.

Consider the following subproblems:

$$\textbf{NLP}(Q)\textbf{:} \quad f(Q) := \min_{(y_0,y)} \; y_0 \qquad \text{s.t.} \qquad (y_0,y) \in G(Q) \tag{8.115b}$$

$$\textbf{LP}(y)\textbf{:} \quad c(y) := \min_{x \geq 0} \; c^\mathsf{T} x \qquad \text{s.t.} \qquad Ax \leq b - F(y) \tag{8.115c}$$

$$\textbf{DP}(y)\textbf{:} \quad d(y) := \max_{\mu \geq 0} \; (F(y)-b)^\mathsf{T} \mu \qquad \text{s.t.} \qquad c + A^\mathsf{T}\mu \geq 0 \tag{8.115d}$$

Since NLP(Q) (8.115b) is a relaxation of (8.113a), if (8.115b) is infeasible then (8.113a) is infeasible (and hence (8.110) is infeasible by Theorem 8.31). On the other hand, if (y_0^*, y^*) is an optimal solution of (8.115b), then it is also optimal for (8.113a) if and only if the optimal value of the dual problem DP(y^*) (8.115d) satisfies:

$$d(y^*) = y_0^* - f(y^*) \tag{8.116}$$

This follows from Theorem 8.31 and strong duality between the linear programs (8.115c) and (8.115d) since the optimal value $y_0^* - f(y^*)$ of LP(y^*) is finite.

The Benders algorithm proceeds as follows. Starting with any subset $Q \subseteq C$, the resulting nonlinear program NLP(Q) (8.115b) is solved, with three possible outcomes:

1 unbounded $f(Q) = -\infty$: This happens if and only if the feasible set $P := \{\mu \in \mathbb{R}^m : c + A^\mathsf{T}\mu \geq 0, \mu \geq 0\}$ of DP(y) (8.115d) is empty (Exericse 8.31). It implies that the original problem (8.110) is feasible but unbounded. The algorithm terminates.
2 infeasible $f(Q) = \infty$: This happens only if P is nonempty and implies that (8.110) is infeasible since $f(Q)$ is a relaxation of (8.113a). The algorithm terminates.
3 bounded $-\infty < f(Q) < \infty$: This happens only if P is nonempty. Suppose $(\bar{y}_0, \bar{y}) \in \mathbb{R}^{1+l}$ is a minimizer of NLP(Q). The dual problem DP($\bar{y}$) is then solved.

The solution of DP($\bar{y}$) also leads to three possible next steps, depending on whether DP($\bar{y}$) is bounded:

4 Suppose $d(\bar{y})$ is bounded and attained by $\bar{\mu}$. Then strong duality implies that LP($\bar{y}$) (8.115c) has an optimal solution $\bar{x}$ (Theorem 8.23). The minimizer $(\bar{y}_0, \bar{y})$ of NLP(Q) and the primal-dual optimal solution $(\bar{x}, \bar{\mu})$ of LP($\bar{y}$) and its dual DP($\bar{y}$) satisfy (Exericse 8.32):

$$\bar{y}_0 \leq c^\mathsf{T}\bar{x} + f(\bar{y}) \;=\; (F(\bar{y}) - b)^\mathsf{T}\bar{\mu} + f(\bar{y}) \tag{8.117}$$

with equality if and only if $(\bar{y}_0, \bar{y})$ is feasible and hence optimal for (8.113a). Theorem 8.31 therefore implies that, if equality holds in (8.117), then $(\bar{x}, \bar{y})$ is optimal for (8.110) with optimal value $\bar{y}_0$. The algorithm terminates.
5 On the other hand, suppose the inequality is strict in (8.117). Then $f(Q) > -\infty$ implies that $c + A^\mathsf{T}\bar{\mu} \geq 0$, i.e., $(1, \bar{\mu}) \in C$. Therefore the strict inequality in (8.117) means that $(\bar{y}_0, \bar{y})$ violates the constraint $g(y_0, y; 1, \bar{\mu}) \geq 0$ in the definition of G

in (8.113c), i.e., $(1,\bar{\mu})$ is in C but not Q. We add $(1,\bar{\mu})$ to Q, introducing the additional constraint

$$(y_0 - f(y)) + \bar{\mu}^\mathsf{T}(b - F(y)) \geq 0$$

in the feasible set $G(Q)$ in (8.115a), and solve NLP(Q) again.

6 Suppose DP($\bar{y}$) is unbounded, i.e., $d(\bar{y}) = \infty$. Then LP duality (Theorem 8.23) implies that LP($\bar{y}$) is infeasible, i.e., there is no $x \geq 0$ such that $Ax \leq b - F(\bar{y})$. The Farkas lemma implies that (8.112) is satisfied by some $(\bar{\mu}_0,\bar{\mu}) \in C$ for which $g(\bar{y}_0,\bar{y};\bar{\mu}_0,\bar{\mu}) \geq 0$ is violated in G and which should be added to Q.

To identify such a $(\bar{\mu}_0,\bar{\mu})$, note that, since $d(\bar{y}) = \infty$, there must be a feasible point $\bar{v}$ of DP($\bar{y}$) and a direction $\bar{\mu}$ such that (i) the halfline $\bar{v} + \alpha\bar{\mu}$ is in the feasible set of DP($\bar{y}$) for all $\alpha \geq 0$ ($\bar{\mu}$ is called a direction of recession; see Definition 12.5), and (ii) $(F(\bar{y}) - b)^\mathsf{T}\bar{\mu} > 0$. The first condition implies that $(0,\bar{\mu}) \in C$ because $c + A^\mathsf{T}(\bar{v} + \alpha\bar{\mu}) = (c + A^\mathsf{T}\bar{v}) + \alpha(A^\mathsf{T}\bar{\mu}) \geq 0$ which can hold for all $\alpha \geq 0$ if and only if $A^\mathsf{T}\bar{\mu} \geq 0$. The second condition therefore identifies $(0,\bar{\mu}) \in C$ for which $g(\bar{y}_0,\bar{y};0,\bar{\mu}) \geq 0$ is violated in G, i.e., $(0,\bar{\mu}) \in C \setminus Q$. We add $(0,\bar{\mu})$ to Q and solve NLP(Q) is again with the additional constraint:

$$\bar{\mu}^\mathsf{T}(b - F(y)) \geq 0$$

The overall algorithm for Benders decomposition is summarized in Figure 8.18.

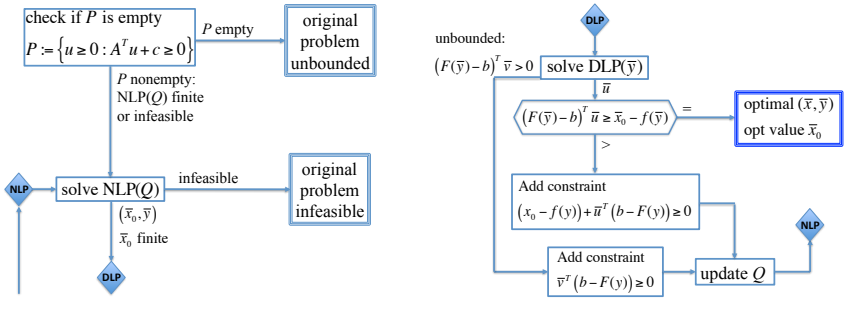


Figure 8.18 Benders decomposition (May 18, 2025: Changes: (i) DLP –> DP; PLP –> LP; (ii) $\bar{x}_0 \to \bar{y}_0$; $\bar{y} \to \bar{y}$; (iii) $\bar{u} \to \bar{\mu}$, $\bar{v} \to \bar{\mu}$; (iii) "unbounded $(F(\bar{y}) - b)^T \bar{v} > 0$" –> "unbounded $\bar{\mu}^\mathsf{T}(b - F(\bar{y})) < 0$"; (iv) $(F(\bar{y}) - b)^T \bar{u} \geq \bar{x}_0 - f(\bar{y}) \to (\bar{y}_0 - f(\bar{y})) + \bar{\mu}^\mathsf{T}(b - F(\bar{y})) \leq 0$; (v) "Add constraint $(y_0 - f(y)) + \bar{\mu}^\mathsf{T}(b - F(y)) \geq 0$ to $G(Q)$"; (vii) "Add constraint $\bar{\mu}^\mathsf{T}(b - F(y)) \geq 0$ to Q". Delete "update $G(Q)$".)

The Benders algorithm terminates in a finite number of steps. This is because C is a polyhedral pointed cone, it is the convex hull of finitely many extreme halflines. In each iteration in which the algorithm does not terminate an extreme halfline, which is not already in Q, is added to Q.

8.6 Convergence analysis

Consider the problem (8.75), reproduced here:

$$f^* \;:=\; \min_x \; f(x) \;\text{ subject to }\; x \in X \tag{8.118}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $X \subseteq \mathbb{R}^n$. Iterative algorithms for solving (8.118) generally take the form (8.77a), reproduced here:

$$x(t+1) \;=\; g(x(t)) \tag{8.119a}$$

where $g : \mathbb{R}^n \to \mathbb{R}^n$. For example a gradient descent algorithm can be interpreted as the following fixed-point iteration

$$x(t+1) \;=\; [x(t) - \gamma G(x(t))\nabla f(x(t)]_X \;=:\; g(x(t)) \tag{8.119b}$$

where $\gamma > 0$ is a stepsize, $G(x) \succ 0$ is a scaling matrix, and $[\cdot]_X$ is the projection to the feasible set X. A fixed point x^* of the gradient algorithm (8.119b) satisfies the optimality condition (8.76). The fixed-point iteration (8.119) can be used not only for solving an optimization problem, but also for solving a system of nonlinear equations $h(x) = 0$ with the corresponding fixed-point iteration $x(t+1) = x(t) + h(x(t))$. Indeed many of the optimization algorithms can be interpreted as solving a system of equations representing the KKT condition.

We assume:

C8.5: The objective function f is lower bounded on X, continuously differentiable and convex. The feasible set X is nonempty, closed and convex.

C8.5 guarantees that (8.118) is feasible and gradient algorithms (8.119b) are well defined.

8.6.1 Convergence theorems

In this subsection we prove some basic results that are widely used for convergence analysis of constrained optimization (8.118).

Since the feasible set X in (8.118) is not necessarily compact (bounded), the optimum may not be attained (e.g., $f(x) = e^{-x}$ on $X = \mathbb{R}$). Moreover the sequence $(x(t), t = 0, 1, \dots)$ generated by the gradient projection algorithm (8.78) may not stay bounded and hence may not have any convergent subsequence (the Bolzano-Weierstrass theorem states that a sequence $(x(t), t = 0, 1, \dots)$ has a convergent subsequence if it is bounded). To guarantee that the gradient projection algorithm makes progress towards minimizing f, we need:

C8.6: The gradient of f is Lipschitz continuous with Lipschitz constant β, i.e.,

$$\|\nabla f(y) - \nabla f(x)\|_2 \leq \beta \|y - x\|_2 \qquad \forall x, y \in \mathbb{R}^n$$

Note that the norm is Euclidean.[10] C8.6 implies the following useful result which is used in Theorem 8.35 to prove the optimality of gradient projection algorithm (8.78).

Lemma 8.32 (Descent Lemma.)**.** If $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and satisfies C8.6 then

$$f(x+y) \leq f(x) + y^{\mathsf{T}} \nabla f(x) + \frac{\beta}{2} \|y\|_2^2 \qquad \forall x, y \in \mathbb{R}^n$$

Proof We estimate the difference $f(x+y) - f(x)$ by considering the *scalar* function $g(s)$ defined by the intersection of the $f(x)$ surface with the vertical plane at x in the direction y. Fix any $x, y \in \mathbb{R}^n$ and define

$$g(s) := f(x + sy) \qquad \text{for } s \in [0, 1]$$

As s ranges from 0 to 1, $x + sy$ moves from x to $x + y$ in a straight line and

$$g'(s) = y^{\mathsf{T}} \nabla f(x + sy)$$

is the directional derivative of f at $x + sy$ in the direction y. Then

$$\begin{aligned} f(x+y) - f(x) = g(1) - g(0) &= \int_0^1 g'(s)\, ds = \int_0^1 y^{\mathsf{T}} \nabla f(x+sy)\, ds \\ &= \int_0^1 \left(y^{\mathsf{T}} \nabla f(x) + y^{\mathsf{T}} (\nabla f(x+sy) - \nabla f(x)) \right) ds \\ &\leq y^{\mathsf{T}} \nabla f(x) + \int_0^1 \|y\|_2 \, \|\nabla f(x+sy) - \nabla f(x)\|_2 \, ds \\ &\leq y^{\mathsf{T}} \nabla f(x) + \|y\|_2 \int_0^1 \beta \|sy\|_2 \, ds \\ &= y^{\mathsf{T}} \nabla f(x) + \frac{\beta}{2} \|y\|_2^2 \end{aligned}$$

where the first inequality follows from the Cauchy-Schwarz inequality and the second inequality follows from condition C8.6. □

As we will see in Chapter 8.6.3, under condition C8.6, the gradient projection algorithm generates a sequence $(x(t), t = 0, 1, \ldots)$ such that $f(x(t))$ monotonically decreases. The sequence $(x(t), t = 0, 1, \ldots)$ may not converge, but any converging subsequence converges to an optimal point (Theorem 8.35). When f is strongly convex (Definition 8.4) then the gradient projection algorithm indeed converges and does so geometrically. This is because strong convexity implies that the gradient projection algorithm is a contraction mapping, as we now explain.

[10] In contrast, the norm that defines a *contraction mapping* can be arbitrary (see Definition 8.10 below).

Definition 8.10 (Contraction)**.** Consider a function $T : X \to X$ from a subset X of $\mathbb{R}^n$ into itself. T is called a *contraction mapping* or simply a *contraction* if there exists an $\alpha \in [0,1)$ such that

$$\|T(y) - T(x)\| \leq \alpha \|y - x\| \qquad \forall x, y \in X$$

for an arbitrary norm $\|\cdot\|$.

A function T can be a contraction under a certain norm, but not under a different norm, so the proper choice of norm is critical.

Theorem 8.33 (Contraction mapping theorem)**.** Suppose $T : X \to X$ is a contraction mapping on a closed subset X of $\mathbb{R}^n$. Then

1 There exists a unique fixed point x^* such that $x^* = T(x^*)$.
2 Starting from any initial point $x(0) \in X$, the contraction iteration $x(t+1) := T(x(t))$ converges geometrically to x^*:

$$\|x(t) - x^*\| \leq \alpha^t \|x(0) - x^*\| \qquad \forall t \geq 0$$

Proof Consider the contraction iteration $x(t+1) := T(x(t))$. Definition 8.10 implies

$$\|x(t+1) - x(t)\| \;\leq\; \alpha \|x(t) - x(t-1)\| \;\leq\; \cdots \;\leq\; \alpha^t \|x(1) - x(0)\|$$

Hence, for all $t \geq 0$ and $s \geq 1$, we have

$$\begin{aligned} \|x(t+s) - x(t)\| &= \left\| \sum_{m=0}^{s-1} (x(t+m+1) - x(t+m)) \right\| \\ &\leq \sum_{m=0}^{s-1} \|x(t+m+1) - x(t+m)\| \;\leq\; \|x(1) - x(0)\| \, \alpha^t \sum_{m=0}^{s-1} \alpha^m \\ &\leq \frac{\alpha^t}{1-\alpha} \|x(1) - x(0)\| \end{aligned}$$

Since $\alpha \in [0,1)$, $x(t)$ is a Cauchy sequence (i.e., given any $\epsilon > 0$, there exists n such that for all $s, t > n$, $\|x(t+s) - x(t)\| < \epsilon$) and hence must converge to a point x^* in $\mathbb{R}^n$. Since X is closed, $x^* \in X$. Since T is continuous,

$$x^* \;=\; \lim_t x(t+1) = \lim_t T(x(t)) \;=\; T(\lim_t x(t)) \;=\; T(x^*)$$

and hence x^* is a fixed point of T. Moreover, the fixed point is unique for, otherwise, if x^* and y^* are both fixed points then

$$\|y^* - x^*\| = \|T(y^*) - T(x^*)\| \;\leq\; \alpha \|y^* - x^*\|$$

implying $y^* = x^*$ since $\alpha \in [0,1)$. This completes the proof of part 1.

For part 2, we have for all $t \geq 1$,

$$\|x(t) - x^*\| = \|T(x(t-1)) - T(x^*)\| \;\leq\; \alpha \|x(t-1) - x^*\|$$

Hence $\|x(t) - x^*\| \;\leq\; \alpha^t \|x(0) - x^*\|$. □

When a function $T : X \to X$ from a subset X of $\mathbb{R}^n$ into itself has a fixed point $x^* \in X$, we call T a *pseudocontraction mapping* or simply a *pseudocontraction* if there exists an $\alpha \in [0,1)$ such that

$$\|T(x) - x^*\| \leq \alpha \|x - x^*\| \qquad \forall x \in X$$

for an arbitrary norm $\|\cdot\|$. Pseudocontraction is a weaker notion than contraction, i.e., if T is a contraction then it is a pseudocontraction, but the converse may not hold. Theorem 8.33 however extends to pseudocontraction, i.e, the fixed point x^* in the definition of pseudocontraction is the unique fixed point in X and the fixed-point iteration converges geometrically to x^*. Note however that the existence of a fixed point x^* is part of the definition of pseudocontraction and x^* is often unavailable in applications.

If a function f is strongly convex on X then Theorem 8.6 implies (see (8.12))

$$\alpha \|y - x\|_2 \;\leq\; \|\nabla f(y) - \nabla f(x)\|_2 \;\leq\; \beta \|y - x\|_2, \qquad x, y \in X$$

In particular it satisfies C8.6. The proof of Theorem 8.6 (see (8.11)) and the Descent Lemma 8.32 show that

$$y^\mathsf{T} \nabla f(x) + \frac{\alpha}{2}\|y\|_2^2 \;\leq\; f(x+y) - f(x) \;\leq\; y^\mathsf{T} \nabla f(x) + \frac{\beta}{2}\|y\|_2^2$$

A consequence is that the gradient projection algorithm (8.78) is a contraction mapping and therefore converges geometrically to the unique optimal point, as explained in Theorem 8.36.

8.6.2 Gauss-Seidel algorithm

The Gauss algorithm introduced in Chapter 4.4.1 is a fixed-point iteration of the form

$$x(t+1) \;=\; g(x(t)) \tag{8.120}$$

where $x \in \mathbb{R}^n$, $g : X \to X$ and X is a nonempty subset of $\mathbb{R}^n$. The goal of (8.120) is to compute a fixed point x^* that satisfies $x^* = g(x^*)$. Almost all iterative algorithms for constrained optimization can be interpreted as a fixed-point iteration (8.120), including gradient algorithms. The advantage of the class of Gauss algorithms however is that gradient is not necessary, simplifying computation, e.g., the backward-forward sweep of Chapter 5.3. We study the convergence of Gauss algorithms in this subsection and that of gradient algorithms in Chapter 8.6.3.

If g is a contraction mapping then the Contraction Mapping Theorem 8.33 implies that the fixed-point iteration (8.120) will converge to a unique fixed point x^* geometrically.

Suppose $X = \prod_{i=1}^m X_i$ where each $X_i \subseteq \mathbb{R}^{n_i}$ is nonempty such that $n_1 + \cdots + n_m = n$.

Decompose $x \in \mathbb{R}^n$ into $x = (x_1, \ldots, x_m)$ where $x_i \in X_i$. We are given a norm $\|\cdot\|_i$ on $\mathbb{R}^{n_i}$ for each i. Define the norm $\|\cdot\|$ on $\mathbb{R}^n$ by

$$\|x\| \;:=\; \max_i \|x_i\|_i \tag{8.121}$$

If $n_i = 1$ and $\|x_i\|_i := |x_i|$ then $\|x\| = \max_i |x_i|$ is the l_∞ norm. The Gauss algorithm (8.120) updates all components x_i simultaneously. A Gauss-Seidel algorithm updates one component at a time and the computation of component $x_i(t+1)$ uses the latest value $x_1(t+1), \ldots, x_{i-1}(t+1)$:

$$x_i(t+1) \;=\; g_i(x_1(t+1), \ldots, x_{i-1}(t+1), x_i(t), \ldots, x_m(t)), \qquad i = 1, \ldots, m$$

We will show that, if the Gauss algorithm (8.120) is a contraction mapping, so is Gauss-Seidel algorithm with the same (unique) fixed point. To this end we define a mapping $h : X \to X$ that represents the Gauss-Seidel update after every m updates.

Let $g_i : X \to X_i$ and $h_i : X \to X_i$ denote the ith block-components of g and h respectively:

$$g(x) \;=\; (g_1(x), \ldots, g_m(x)), \qquad h(x) \;=\; (h_1(x), \ldots, h_m(x))$$

Given a Gauss algorithm $g : X \to X$ in (8.120), the corresponding Gauss-Seidel algorithm $h : X \to X$ is defined recursively through its block-components:

$$h_1(x) \;:=\; g_1(x_1, \ldots, x_m) \tag{8.122a}$$

$$h_i(x) \;:=\; g_i(h_1(x), \ldots, h_{i-1}(x), x_i, \ldots, x_m), \quad i = 2, \ldots, m \tag{8.122b}$$

Theorem 8.34 (Gauss-Seidel algorithm)**.** Suppose $X \subseteq \mathbb{R}^n$ is closed. Suppose $g : X \to X$ is a contraction mapping with a unique fixed point x^* and parameter $\alpha \in [0, 1)$ under the norm (8.121), i.e.,

$$\|g(y) - g(x)\| \;\le\; \alpha \|y - x\|, \qquad \forall x, y \in X$$

Then h in (8.122) is also a contraction with the same (unique) fixed point x^* and parameter α. Hence the sequence $x(t)$ generated by h converges geometrically to x^*.

Proof The assumption of Cartesian product $X = \prod_{i=1}^m X_i$ and the definition of the norm (8.121) imply that the ith block-components of g satisfy

$$\|g_i(y) - g_i(x)\|_i \le \max_j \|g_j(y) - g_j(x)\|_j = \|g(y) - g(x)\| \le \alpha \|y - x\| = \alpha \max_j \|y_j - x_j\|_j$$

Therefore h_i in (8.122) satisfy

$$\|h_i(y) - h_i(x)\|_i \;\le\; \alpha \max \left\{ \max_{j<i} \|h_j(y) - h_j(x)\|_j, \max_{j \ge i} \|y_j - x_j\|_j \right\}, \;\; i = 1, \ldots, m$$

Induction on i then shows that $\|h_i(y) - h_i(x)\|_i \le \alpha \|y - x\|$ for all i, implying that $\|h(y) - h(x)\| \le \alpha \|y - x\|$. It is easy to show that the unique fixed point of h is also x^* and the remaining claim follows from the Contraction Mapping Theorem 8.33. □

Example 8.18 (Backward-forward sweep [60])**.** We analyze the convergence of the backward-forward sweep (BFS) Algorithm 2 in Chapter 5.3.2:

$$\begin{aligned} I^s_{ij}(t) &= \sum_{k:j\to k} I^s_{jk}(t) - \left(\left(\frac{s_j}{V_j(t-1)}\right)^{\mathsf{H}} - y^m_{jj} V_j(t-1)\right), && j \in N \\ V_j(t) &= V_i(t) - (y^s_{ij})^{-1} I^s_{ij}(t), && j \in N \end{aligned}$$

where $i := i(j)$ denotes the unique parent of j and $y^m_{jj} := y^m_{ji} + \sum_{k:j\to k} y^m_{jk}$ is the total shunt admittance incident on bus j. This can be represented compactly using the $(N+1)\times N$ incidence matrix C defined in (5.4) and reproduced here:

$$C_{jl} = \begin{cases} 1 & \text{if } l = j \to k \text{ for some bus } k \\ -1 & \text{if } l = i \to j \text{ for some bus } i \\ 0 & \text{otherwise} \end{cases}$$

The matrix C is of rank N. Decompose C into the $N\times N$ non-singular reduced incidence matrix $\hat{C}$ and the first row c_0^{T} corresponding to the root bus 0:

$$C =: \begin{bmatrix} c_0^{\mathsf{T}} \\ \hat{C} \end{bmatrix}$$

Define $N\times N$ diagonal matrices:

$$\hat{s} := \operatorname{diag}\left(s_j, j \in N\right), \quad \hat{y}^m := \operatorname{diag}\left(y^m_j, j \in N\right), \quad \hat{y}^s := \operatorname{diag}\left(y^s_{ij}, j \in N\right)$$

Then the BFS algorithm consists of the following nonlinear iteration:

$$\hat{C} I^s(t) = \hat{s}^{\mathsf{H}} V^{-\mathsf{H}}(t-1) - \hat{y}^m V(t-1) \tag{8.123a}$$

$$I^s(t) = \hat{y}^s \left(c_0 V_0 + \hat{C}^{\mathsf{T}} V(t)\right) \tag{8.123b}$$

where the column vector $V^{-\mathsf{H}} := (1/V_j^{\mathsf{H}}, j \in N)$; and $\hat{s}^{\mathsf{H}}$ takes the componentwise conjugate of the diagonal matrix $\hat{s}$. Substituting (8.123b) into (8.123a) to eliminate $I^s(t)$ yields a Gauss algorithm in terms of V only:

$$V(t) = \hat{L}^{-1}\left(\hat{s}^{\mathsf{H}} V^{-\mathsf{H}}(t-1) - \hat{y}^m V(t-1) - \hat{C}\hat{y}^s c_0 V_0\right) =: g(V(t-1)) \tag{8.124}$$

where the reduced Laplacian matrix $\hat{L} := \hat{C}\hat{y}^s\hat{C}^{\mathsf{T}}$ is nonsingular and encodes the network topology and series admittances. (Properties of $\hat{L}$ and $\hat{L}^{-1}$ for radial networks are given in Theorem 4.10.) We next prove a sufficient condition for the fixed-point iteration g in (8.124) to be a contraction.

Define column vectors of nonnegative real numbers for non-reference buses:

$$|s| := (|s_1|, \ldots, |s_N|), \quad |y^m| := \left(|y^m_1|, \ldots, |y^m_N|\right), \quad |\bar{I}| := \frac{|s|}{1-\epsilon} + (1+\epsilon)|y^m|$$

and $N\times N$ real matrices:

$$|\hat{y}^s| := \operatorname{diag}\left(|y^s_{ij}|, \forall j \in N\right), \quad |\hat{L}| := \hat{C}|\hat{y}^s|\hat{C}^{\mathsf{T}}.$$

Consider the following set of voltages with magnitudes within ϵ of $|V_0|$:

$$\mathbb{V} := \left\{V \in \mathbb{C}^N : |V_0| - \epsilon \leq |V_j| \leq |V_0| + \epsilon,\ j \in N\right\} \tag{8.125}$$

for a given $\epsilon \in (0,1)$. Assuming $V_0 \approx 1$ pu, $\mathbb{V}$ is a set of voltages of practical interest, one that is closer to 1 pu. The following result provides sufficient conditions under which g in (8.124) maps $\mathbb{V}$ onto $\mathbb{V}$ and, moreover, is a contraction. These conditions are sufficient, but not necessary, for the fixed-point iteration (8.124) to converge to a unique power flow equation $V^* \in \mathbb{V}$. Specifically:

1 Suppose the vector $|\hat{L}|^{-1}|\bar{I}|$ satisfies

$$\frac{1}{\epsilon}\left\||\hat{L}|^{-1}|\bar{I}|\right\|_\infty \leq 1 \tag{8.126}$$

where $\|a\|_\infty := \max_i |a_i|$ for vector a. Then $V \in \mathbb{V}$ implies $g(V) \in \mathbb{V}$.

2 Suppose condition (8.126) holds and

$$\rho := \frac{1}{(|V_0| - \epsilon)^2}\left\|\hat{L}^{-1}\hat{s}^{\mathsf{H}}\right\|_2 + \left\|\hat{L}^{-1}\hat{y}^m\right\|_2 < 1 \tag{8.127}$$

where $\|A\|_2$ is the spectral norm of matrix A. Then g is a contraction with parameter ρ and therefore:

- There is a unique fixed point, i.e., power flow solution, V^* of (8.124) in $\mathbb{V}$.
- Starting from any $V(0) \in \mathbb{V}$, the sequence $(V(t), t \geq 1)$ produced by (8.124) converges geometrically to V^*, i.e., $\|V(t) - V^*\|_2 \leq \rho^t \|V(0) - V^*\|_2$.

We prove these two claims. For the first claim let $\mathbf{1}_N$ and 0_N denote the column vectors of N 1's and 0's respectively. We have

$$C^{\mathsf{T}}\mathbf{1}_{N+1} = c_0 + \hat{C}^{\mathsf{T}}\mathbf{1}_N = 0_N$$

and thus $\hat{L}^{-1}\hat{C}\hat{y}^s c_0 V_0 = (\hat{C}^{-\mathsf{T}} c_0) = -V_0\mathbf{1}_N$. This simplifies the fixed-point iteration g in (8.124) to:

$$g(V) = \hat{L}^{-1}\left(\hat{s}^{\mathsf{H}}V^{-\mathsf{H}} - \hat{y}^m V\right) + V_0\mathbf{1}_N \tag{8.128}$$

If $V \in \mathbb{V}$ in (8.125), then

$$\left|\hat{s}^{\mathsf{H}}V^{-\mathsf{H}} - \hat{y}^m V\right| \leq |\bar{I}| \tag{8.129}$$

where the right-hand side is a nonnegative column vector and the left-hand side takes componentwise magnitudes. Theorem 4.10 of Chapter 4.2.6 implies that, for radial networks, the (i,j)th entry of $\hat{L}^{-1} = \hat{C}^{-\mathsf{T}}(\hat{y}^s)^{-1}\hat{C}^{-1}$ is the sum of $(y_l^s)^{-1}$ over lines l on the common segment of paths from bus i to the root and from bus j to the root. Hence the componentwise magnitudes of $\hat{L}^{-1}$ are upper-bounded by $|\hat{L}|^{-1}$. Then (8.129) implies

$$\left|\hat{L}^{-1}\left(\hat{s}^{\mathsf{H}}V^{-\mathsf{H}} - \hat{y}^m V\right)\right| \leq |\hat{L}|^{-1}|\bar{I}| \tag{8.130}$$

where again the right-hand side is a nonnegative column vector and the left-hand side

takes componentwise magnitudes. Therefore, if condition (8.126) is satisfied, then by (8.128)(8.130), we have $|V_0| - \epsilon \le |g_j(V)| \le |V_0| + \epsilon$ for all $j \in N$, i.e., $V \in \mathbb{V}$ implies $g(V) \in \mathbb{V}$.

For the second claim, by (8.128), for any $U, V \in \mathbb{V}$:

$$\begin{aligned} \|g(U) - g(V)\|_2 &\le \left\|\hat{L}^{-1}\hat{s}^{\mathsf{H}}\right\|_2 \left\|U^{-\mathsf{H}} - V^{-\mathsf{H}}\right\|_2 + \left\|\hat{L}^{-1}\hat{y}^m\right\|_2 \|U - V\|_2 \\ &\le \left(\frac{1}{(|V_0| - \epsilon)^2} \left\|\hat{L}^{-1}\hat{s}^{\mathsf{H}}\right\|_2 + \left\|\hat{L}^{-1}\hat{y}^m\right\|_2 \right) \|U - V\|_2 \;=\; \rho \|U - V\|_2 \end{aligned} \tag{8.131}$$

where the first inequality uses the subadditivity of vector norms and the definition of induced matrix norms. The second inequality is because

$$\begin{aligned} \left\|U^{-\mathsf{H}} - V^{-\mathsf{H}}\right\|_2 &= \sqrt{\sum_{j \in N} \left(\frac{1}{U_j^{\mathsf{H}}} - \frac{1}{V_j^{\mathsf{H}}}\right)\left(\frac{1}{U_j} - \frac{1}{V_j}\right)} = \sqrt{\sum_{j \in N} \frac{|U_j - V_j|^2}{|U_j|^2 |V_j|^2}} \\ &\le \frac{1}{(|V_0| - \epsilon)^2} \sqrt{\sum_{j \in N} |U_j - V_j|^2} \;=\; \frac{1}{(|V_0| - \epsilon)^2} \|U - V\|_2 \end{aligned}$$

where the inequality uses $U, V \in \mathbb{V}$ defined in (8.125). Inequality (8.131), condition (8.127), and part 1 imply that g is a contraction from $\mathbb{V}$ onto $\mathbb{V}$. Since $\mathbb{V}$ is a closed subset of $\mathbb{C}^N$, the second claim follows from Theorem 8.33. [11] □

8.6.3 Steepest descent algorithm

Recall the gradient projection algorithm (8.78) of Chapter 8.5.1, reproduced here:

$$x(t+1) := [\, x(t) - \gamma \nabla f(x(t)) \,]_X \tag{8.132}$$

where $\gamma > 0$ is a constant stepsize, $X \subseteq \mathbb{R}^n$ is nonempty, closed and convex, and $[x]_X$ denotes the projection of x onto X.

Conditions C8.5 and C8.6 do not guarantee that the sequence $(x(t), t = 0, 1, \dots)$ generated by the gradient projection algorithm has any convergent subsequence, but if it does then the subsequence converges to an optimal point x^* of (8.118). Note that $(x(t), t = 0, 1, \dots)$ may have multiple convergent subsequences in which case all their limits points are optimal. This implies that, when f is *strictly* convex so that the optimal point x^* is unique, then $(x(t), t = 0, 1, \dots)$ itself converges to x^*, provided the stepsize γ is sufficiently small. This result does not require the gradient projection algorithm (8.132) to be a contraction and is thus less conservative.

Theorem 8.35 (Optimality of gradient projection algorithm)**.** Suppose conditions C8.5 and C8.6 hold, and suppose $0 < \gamma < 2/\beta$. Let $(x(t), t = 0, 1, \dots)$ denote the sequence

[11] Theorem 8.33 applies to real vector spaces. To apply it here, we can treat $V = (\mathrm{Re}(V), \mathrm{Im}(V))$ as a vector in $\mathbb{R}^{2N}$. The l_2 norm in $\mathbb{C}^N$ naturally extends to the l_2 norm in $\mathbb{R}^{2N}$ and the set $\mathbb{V}$ defined in (8.125) becomes a closed subset of $\mathbb{R}^{2N}$.

produced by the gradient projection algorithm (8.132). Then the limit point x^* of any convergent subsequence $(x(t_k), k = 1,2,\dots)$ is an optimal solution of (8.118).

Proof We prove the theorem in three steps. First we show the sequence $(f(x(t)), t = 0,1,\dots)$ of objective values produced by the gradient projection algorithm (8.132) converges monotonically. Moreover the difference sequence $(x(t+1)-x(t), t = 0,1,\dots)$ converges to zero. Specifically, by the Descent Lemma 8.32, we have

$$f(x(t+1)) \leq f(x(t)) + (x(t+1)-x(t))^{\mathsf{T}} \nabla f(x(t)) + \frac{\beta}{2}\|x(t+1)-x(t)\|_2^2 \qquad (8.133)$$

Theorem 8.9.2 implies that for all t

$$(y - x(t+1))^{\mathsf{T}} \left(x(t) - \gamma \nabla f(x(t)) - x(t+1) \right) \leq 0 \qquad \forall y \in X \qquad (8.134)$$

In particular let $y = x(t)$ and we have, after rearranging,

$$(x(t+1)-x(t))^{\mathsf{T}} \nabla f(x(t)) \leq -\frac{1}{\gamma}\|x(t+1)-x(t)\|_2^2$$

Substituting into (8.133) we have

$$f(x(t+1)) \leq f(x(t)) - \left(\frac{1}{\gamma} - \frac{\beta}{2}\right)\|x(t+1)-x(t)\|_2^2 \qquad (8.135)$$

Hence the sequence $(f(x(t)), t = 0,1,\dots)$ is strictly decreasing as long as $x(t+1) \neq x(t)$ provided $\gamma < 2/\beta$. Since f is lower bounded on X (condition C8.5), the sequence $(f(x(t)), t = 0,1,\dots)$ is bounded and monotone and thus converges. Rearranging (8.135), we also have

$$\|x(t+1)-x(t)\|_2^2 \leq \left(\frac{1}{\gamma} - \frac{\beta}{2}\right)^{-1} (f(x(t)) - f(x(t+1)))$$

Since $f(x(t))$ converges this means that the *differences* $x(t+1)-x(t)$ converge to zero (though this does not guarantee that $x(t)$ itself converges).

Second suppose there is a subsequence $(x(t_k), k = 1,2,\dots)$ that converges to x^*. Consider the sequence $(x(t_k+1), k = 1,2,\dots)$. By Theorem 8.9.3, the iteration $x(t+1) = [x(t) - \gamma \nabla f(x(t))]_X$ defined by (8.132) is a projection and hence a continuous function of $x(t)$. Hence the sequence $(x(t_k+1), k = 1,2,\dots)$, being the image of a continuous function on $x(t_k)$, also converges. We now show that it converges to x^* as $k \to \infty$. Fix any $\epsilon > 0$. We have to show that there exists a K such that

$$\|x(t_k+1) - x^*\|_2 < \epsilon \qquad \forall k > K$$

Since $x(t_k) \to x^*$ there exists an K' such that

$$\|x(t_k) - x^*\|_2 < \frac{\epsilon}{2} \qquad \forall k > K' \qquad (8.136a)$$

Step 1 above shows that $x(t_k+1) - x(t_k)$ converges to zero and hence there exists K'' such that

$$\|x(t_k+1) - x(t_k)\|_2 < \frac{\epsilon}{2} \qquad \forall k > K'' \qquad (8.136b)$$

Combining (8.136) we have for $k > K := \max\{K', K''\}$

$$\|x(t_k+1) - x^*\|_2 \le \|x(t_k+1) - x(t_k)\|_2 + \|x(t_k) - x^*\|_2 \;<\; \epsilon$$

as desired.

Finally note that (8.134) holds for all t. In particular consider $t = t_k, k = 1, 2, \ldots$. Taking $k \to \infty$, (8.134) yields

$$\left(y - \lim_k x(t_k+1)\right)^{\mathsf{T}} \left(\lim_k x(t_k) - \gamma \lim_k \nabla f(x(t_k)) - \lim_k x(t_k+1)\right) \;\le\; 0, \quad \forall y \in X$$

Since f is continuously differentiable and $\lim_k x(t_k) = \lim_k x(t_k+1) = x^*$, we have

$$\gamma\ (y - x^*)^{\mathsf{T}}\ \nabla f(x^*) \ge 0 \qquad \forall y \in X$$

Hence x^* satisfies the optimality condition (8.76) and is globally optimal since f is a convex function over a convex set X. □

Suppose f is strongly convex on $X \subseteq \mathbb{R}^n$ with parameter $\alpha > 0$ (Definition 8.4):

$$(\nabla f(y) - \nabla f(x))^{\mathsf{T}} (y - x) \;\ge\; \alpha \|y - x\|_2^2 \qquad \forall x, y \in X \tag{8.137a}$$

If $\max_{x \in X} f(x) < \infty$ then Theorem 8.6 implies

$$\|\nabla f(y) - \nabla f(x)\|_2 \;\le\; \beta \|y - x\|_2, \qquad x, y \in X \tag{8.137b}$$

where β is a finite bound on the maximum eigenvalue of $\nabla^2 f(x)$ on X, i.e., it satisfies C8.6. The upper bound in (8.137b) guarantees strict descent for sufficiently small stepsize $\gamma > 0$ while the lower bound in (8.137a) guarantees geometric convergence. This implies that the mapping defined by the gradient projection algorithm (8.132) is a contraction. Theorem 8.33 then implies that the algorithm converges geometrically to the unique optimal solution of (8.118). The condition $\sup_{x \in X} f(x) < \infty$ is not restrictive; see Remark 8.1.

Theorem 8.36 (Geometric convergence of gradient projection algorithm)**.** Suppose condition C8.5 holds. Suppose f is twice continuously differentiable, is strongly convex with parameter $\alpha > 0$ and $\max_{x \in X} f(x) < \infty$. If $0 < \gamma < 2\alpha/\beta^2$ then

1 There is a unique optimal solution x^* for (8.118).
2 The gradient projection algorithm (8.132) converges geometrically to x^*.

Proof The gradient project algorithm (8.132) is the iteration $x(t+1) = T(x(t))$ where $T : X \to X$ is defined by $T(x) := [\,x - \gamma \nabla f(x)\,]_X$. We will show that T is a contraction when f is strongly convex. Then the assertions follow from Theorem 8.33.

We have under the Euclidean norm

$$\begin{aligned}
\|T(y) - T(x)\|_2^2 &= \|\,[y - \gamma \nabla f(y)]_X \;-\; [x - \gamma \nabla f(x)]_X\,\|_2^2 \\
&\le \|\,(y - x) - \gamma\,(\nabla f(y) - \nabla f(x))\|_2^2 \\
&= \|y - x\|_2^2 \;-\; 2\gamma\,(\nabla f(y) - \nabla f(x))^{\mathsf{T}}(y - x) \;+\; \gamma^2 \|\nabla f(y) - \nabla f(x))\|_2^2
\end{aligned}$$

where the inequality above follows from the fact that the projection operation is non-expansive (Theorem 8.9.3). Conditions in the theorem imply that (8.137) holds and hence $(\nabla f(y) - \nabla f(x))^{\mathsf{T}}(y-x) \geq \alpha \|y-x\|_2^2$ and $\|\nabla f(y) - \nabla f(x))\|_2^2 \leq \beta^2 \|y-x\|_2^2$. Therefore

$$\|T(y) - T(x)\|_2^2 \leq \left(1 - 2\alpha\gamma + \gamma^2\beta^2\right) \|y-x\|_2^2$$

Hence T is a contraction if and only if $\rho(\gamma) := \gamma^2\beta^2 - 2\alpha\gamma + 1 \in [0,1)$.

Strong convexity of f implies (when $\sup_{x\in X} f(x) < \infty$; see Remark 8.1)

$$\alpha\|y-x\|_2 \;\leq\; \|\nabla f(y) - \nabla f(x)\|_2 \;\leq\; \beta\|y-x\|_2, \qquad x, y \in X$$

and hence $0 < \alpha \leq \beta$. This implies that $\rho(\gamma) \geq 0$ for all γ. Moreover $\rho(\gamma) = 1 - \gamma(2\alpha - \beta^2\gamma) < 1$ if $\gamma < 2\alpha/\beta^2$. Hence T is a contraction if $0 < \gamma < 2\alpha/\beta^2$. Theorem 8.33 then implies that $x(t)$ converges geometrically to a unique fixed point x^* of T and Theorem 8.35 guarantees that x^* is the optimal solution of (8.118). □

The condition number $\|(\nabla^2 f(x))^{-1}\| \|\nabla^2 f(x)\|$ of the Hessian matrix can affect greatly the convergence of gradient algorithms. The bound on the stepsize γ in Theorem 8.35 is $\gamma < 2/\beta$ and that in Theorem 8.36 is $\gamma < (2/\beta)(\alpha/\beta)$. As discussed in Remark 8.1, $\alpha = \min_{x\in X} \lambda_{\min}(x)$ and $\beta = \max_{x\in X} \lambda_{\max}(x)$ (assuming X is closed). If the minimization in α and the maximization in β are attained at the same point $\tilde{x}$, then, since $\nabla^2 f(\tilde{x})$ is symmetric and positive definite, its condition number is β/α under the spectral norm. Hence the bound on the stepsize γ is scaled down by the condition number of the Hessian to ensure (geometric) convergence of the sequence $(x(t), t = 0, 1, \ldots)$.

8.6.4 Interior-point algorithm

Consider the convex program (8.87) with an equality and an inequality constraints, reproduced here:

$$f^* \;:=\; \min_{x\in\mathbb{R}^n} f(x) \qquad \text{s.t.} \qquad Ax = b,\; h(x) \leq 0 \tag{8.138}$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, and $f : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}^l$ are convex and twice continuously differentiable. Recall that interior-point methods approximate (8.138) by an equality constrained problem with the inequality constraint $h(x) \leq 0$ replaced by a penalty term, and then solve a sequence of equality constrained problems (8.89), reproduced here:

$$\min_{x\in\mathbb{R}^n} \; t f(x) + \phi(x) \qquad \text{s.t.} \qquad g(x) = 0 \tag{8.139a}$$

using Newton methods. Here ϕ is the logarithmic barrier function (defined in (8.88)):

$$\phi(x) \;:=\; -\sum_{i=1}^{l} \log(-h_i(x)) \tag{8.139b}$$

defined over $\text{dom}(\phi) := \{x \in \mathbb{R}^n : h_i(x) < 0, i = 1, \ldots, l\}$.

The convergence of the barrier method Algorithm 3 for solving the convex program (8.138) has three components:

1 The solution of (8.93) to compute a strictly feasible point if the barrier method does not start at such a point. This is a one-off computational effort.
2 The convergence of the Newton-Raphson algorithm for (8.139) for each t. This determines the computational effort of each outer iteration in Algorithm 3.
3 How the suboptimality gap in solving (8.138) decreases as a function of the outer iteration t. This determines how many outer iterations are needed to achieve a desired accuracy.

For optimal power flow problems the "flat start" where $V_i = 1\angle 0°$ pu for all nodes i is often a strictly feasible point. If strictly feasible point is not available, the one-off computation effort for solving (8.93) is analyzed, e.g., in [57, Chapter 11.5.4, p,592].

The convergence analysis of the Newton-Raphson algorithm is complicated and out of the scope of this book. The algorithm generally proceeds in two phases. In the first phase, called the damped Newton phase, the gradient $\|\nabla f(x_k)\|_2$ is greater than a threshold $\eta > 0$ and each Newton step k (in the iterative solution of (8.139) for a fixed t) decreases the cost $f(x_k)$ by a least a constant amount. If the optimal objective value f^* is finite then the damped Newton phase will terminate after a finite number of steps. Then the algorithm enters the second phase, called the pure Newton phase where $\|\nabla f(x_k)\|_2 < \eta$. In this phase the algorithm converges extremely rapidly (called quadratic convergence) where the optimality gap $f(x_k) - f^*$ decreases as 2^{-2^k}, i.e., roughly, the number of correct digits doubles every iteration k. For details see e.g. [57, Chapter 9.5.3, p,488] for unconstrained problems and [57, Chapter 10.2.4, p,529] for equality constrained problems.

Finally, Theorem 8.30 shows that the suboptimality gap of the central point $x(t)$ for each problem (8.139) with parameter t is l/t (under conditions C8.1 and C8.2). Hence if the scaling factor in Algorithm 3 is γ and a sequence of problems with parameters $t_0, \gamma t_0, \gamma^2 t_0, \ldots$, are solved, the suboptimality gap decreases at least geometrically as $(l/t_0)(\gamma)^{-t}$. Hence the desired accuracy ϵ is achieved after

$$t_{\max} := \frac{\log(l/\epsilon t_0)}{\log \gamma} \quad \text{iterations.}$$

8.6.5 ADMM

Consider the special case of (8.100) where $X := \mathbb{R}^{n_1}$ and $Y := \mathbb{R}^{n_2}$:

$$p^* := \min_{x \in \mathbb{R}^{n_1},\ y \in \mathbb{R}^{n_2}} f(x) + g(y) \qquad \text{s.t.} \qquad Ax + By = c \tag{8.140}$$

where $f : \mathbb{R}^{n_1} \to \mathbb{R}$ and $g : \mathbb{R}^{n_2} \to \mathbb{R}$ are convex and continuously differentiable, $A \in \mathbb{R}^{m\times n_1}$, $B \in \mathbb{R}^{m\times n_2}$, and $c \in \mathbb{R}^m$. In this subsection we analyze the convergence and optimality properties of the ADMM algorithm (8.101), reproduced here:

$$x(t+1) \;=\; \arg\min_{x\in\mathbb{R}^{n_1}} L_\rho\,(x,\; y(t),\; \lambda(t)) \tag{8.141a}$$

$$y(t+1) \;=\; \arg\min_{y\in\mathbb{R}^{n_2}} L_\rho\,(x(t+1),\; y,\; \lambda(t)) \tag{8.141b}$$

$$\lambda(t+1) \;=\; \lambda(t) \;+\; \rho\,(Ax(t+1)+By(t+1)-c) \tag{8.141c}$$

on the convex problem (8.140).

Recall the augmented Lagrangian of (8.140):

$$L_\rho(x,y,\lambda) \;:=\; f(x)+g(y)+\lambda^{\mathsf{T}}(Ax+By-c)+\frac{\rho}{2}\,\|Ax+By-c\|_2^2,$$

The un-agumented Lagrangian of (8.140) is $L_0(x,y,\lambda)$. A point (x^*,y^*,λ^*) is primal-dual optimal for (8.140) if and only if it satisfies the KKT condition $\nabla_{x,y,\lambda}L_0(x^*,y^*,\lambda^*)=0$ in terms of the un-augmented Lagrangian, i.e.,

$$\nabla_\lambda L_0(x^*,y^*,\lambda^*) \;=\; Ax^*+By^*-c \;=\; 0 \tag{8.142a}$$

$$\nabla_x L_0(x^*,y^*,\lambda^*) \;=\; \nabla f(x^*)+A^{\mathsf{T}}\lambda^{*\mathsf{T}} \;=\; 0 \tag{8.142b}$$

$$\nabla_y L_0(x^*,y^*,\lambda^*) \;=\; \nabla g(y^*)+B^{\mathsf{T}}\lambda^{*\mathsf{T}} \;=\; 0 \tag{8.142c}$$

Such a point (x^*,y^*,λ^*) is also a saddle point of L_0 (Theorem 8.15). Our goal is to show that the iterates $(x(t),y(t),\lambda(t))$ produced by the ADMM algorithm (8.141) will satisfy (8.142) asymptotically. It is in this sense that we interpret the ADMM algorithm as computing a KKT point.

Our analysis will proceed in three steps. First we will show that $(y(t),\lambda(t))$ satisfies condition (8.142c) at every t. Then we will show that $(x(t),y(t),\lambda(t))$ satisfies (8.142a)(8.142b) asymptotically. Finally we show that, as a consequence, the cost $f(x(t))+g(y(t))$ converges to the optimal cost p^*. This does not imply that the sequence $(x(t),y(t),\lambda(t))$ converges, but we will show that the limit point of any convergent subsequence is a saddle point of L_0 and hence primal-dual optimal.

Define the primal residual

$$r(x,y) \;:=\; Ax+By-c, \qquad r(t) \;:=\; r(x(t),y(t)) \;:=\; Ax(t)+By(t)-c$$

Then the derivatives of the augmented Lagrangian L_ρ are:

$$\begin{aligned}
\nabla_\lambda L_\rho(x,y,\lambda) \;&=\; Ax+By-c \;=\; r(x,y)\\
\nabla_x L_\rho(x,y,\lambda) \;&=\; \nabla f(x)+A^{\mathsf{T}}\lambda+\rho A^{\mathsf{T}}(Ax+By-c) \;=\; \nabla f(x)+A^{\mathsf{T}}\lambda+\rho A^{\mathsf{T}}r(x,y)\\
\nabla_y L_\rho(x,y,\lambda) \;&=\; \nabla g(y)+B^{\mathsf{T}}\lambda+\rho B^{\mathsf{T}}(Ax+By-c) \;=\; \nabla g(y)+B^{\mathsf{T}}\lambda+\rho B^{\mathsf{T}}r(x,y)
\end{aligned}$$

where we have used $\nabla_z\|Mz+a\|_2^2 = 2M^{\mathsf{T}}(Mz+a)$. Hence the derivatives of L_ρ equal those of the un-augmented Lagrangian in (8.142) if $r(x,y)=0$, i.e., if (x,y) is primal feasible.

Since the minimizations in (8.141a)(8.141b) are unconstrained, the Gauss-Seidel update means that the minimizers $(x(t+1), y(t+1))$ satisfy[12]

$$\nabla_x L_\rho(x(t+1), y(t), \lambda(t)) \;=\; 0 \tag{8.143a}$$

$$\nabla_y L_\rho(x(t+1), y(t+1), \lambda(t)) \;=\; 0 \tag{8.143b}$$

We examine the implication of each. First (8.143b) implies

$$0 \;=\; \nabla g(y(t+1)) + B^{\mathsf{T}}\lambda(t) + \rho B^{\mathsf{T}} r(t+1) \;=\; \nabla g(y(t+1)) + B^{\mathsf{T}}\lambda(t+1) \tag{8.144a}$$

where the last equality uses $\lambda(t+1) = \lambda(t) + \rho r(t+1)$ from (8.141c). This shows that the ADMM iterates $(y(t), \lambda(t))$ satisfy the stationarity condition (8.142c) at all t.

Then (8.143a) implies

$$\begin{aligned} 0 \;&=\; \nabla f(x(t+1)) + A^{\mathsf{T}}\lambda(t) + \rho A^{\mathsf{T}} r(x(t+1), y(t)) \\ &=\; \nabla f(x(t+1)) + A^{\mathsf{T}}\left(\lambda(t) + \rho r(x(t+1), y(t+1))\right) + \rho A^{\mathsf{T}} B(y(t) - y(t+1)) \end{aligned}$$

where the last equality uses $r(x(t+1), y(t)) - r(x(t+1), y(t+1)) = B(y(t) - y(t+1))$. Hence, since $\lambda(t) + \rho r(t+1) = \lambda(t+1)$, we have

$$\nabla f(x(t+1)) + A^{\mathsf{T}}\lambda(t+1) \;=\; s(t+1) \tag{8.144b}$$

where $s(t)$, called the dual residual, is:

$$s(t) \;:=\; \rho A^{\mathsf{T}} B\,(y(t) - y(t-1))$$

Hence $(x(t), \lambda(t))$ satisfies (8.142b) if the dual residual $s(t) = 0$.

We next show that the primal and dual residuals $(r(t), s(t)) \to 0$ as $t \to \infty$, implying that the other two KKT conditions (8.142a)(8.142b) will be satisfied asymptotically by $(x(t), y(t), \lambda(t))$. Moreover $f(x(t)) + g(y(t)) \to p^*$.

Theorem 8.37 (ADMM convergence)**.** Suppose C8.1 holds and a saddle point (x^*, y^*, λ^*) of the un-augmented Lagrangian L_0 exists. Then as $t \to \infty$

1 $r(t) \to 0$, $s(t) \to 0$.
2 $f(x(t)) + g(y(t)) \to p^*$.

Hence the limit point of any convergent subsequence is a saddle point of L_0 and a primal-dual optimum of (8.140).

Proof Let $p(t) := f(x(t)) + g(y(t))$. We prove the theorem in 3 steps.

Step 1: prove (8.145). We derive upper and lower bounds on $p(t+1) - p^*$:

$$-r^{\mathsf{T}}(t+1)\lambda^* \;\le\; p(t+1) - p^* \;\le\; -r^{\mathsf{T}}(t+1)\lambda(t+1) + s^{\mathsf{T}}(t+1)\,(x(t+1) - x^*) \tag{8.145}$$

We prove $r(t) \to 0$ and $s(t) \to 0$ below and use these bounds to conclude $p(t) \to p^*$.

[12] If $X \subsetneq \mathbb{R}^{n_1}$ or $Y \subsetneq \mathbb{R}^{n_2}$, then the convergence analysis replaces (8.143) by the optimality condition $\nabla^{\mathsf{T}} f\,(x(t+1))(x - x(t+1)) \ge 0$, $\nabla^{\mathsf{T}} g(y(t+1))(y - y(t+1)) \ge 0$ for all $x \in X$, $y \in Y$.

Since (x^*, y^*, λ^*) is a saddle point of L_0, (x^*, y^*) is primal feasible and (Saddle-point Theorem 8.14)

$$p^* \;=\; L_0(x^*, y^*, \lambda^*) \;\leq\; L_0(x(t+1), y(t+1), \lambda^*) \;=\; p(t+1) + \lambda^{*\mathsf{T}} r(t+1)$$

which proves the lower bound in (8.145).

For the upper bound in (8.145) we will use (8.144). Specifically (8.144b) implies that $x(t+1)$ minimizes the function $f(x) + \left(A^\mathsf{T}\lambda(t+1) - s(t+1)\right)^\mathsf{T} x$ and (8.144a) implies that $y(t+1)$ minimizes the function $g(y) + \left(B^\mathsf{T}\lambda(t+1)\right)^\mathsf{T} y$. Hence

$$\begin{aligned} f(x(t+1)) + \left(A^\mathsf{T}\lambda(t+1) - s(t+1)\right)^\mathsf{T} x(t+1) \;&\leq\; f(x^*) + \left(A^\mathsf{T}\lambda(t+1) - s(t+1)\right)^\mathsf{T} x^* \\ g(y(t+1)) + \lambda^\mathsf{T}(t+1) B y(t+1) \;&\leq\; g(y^*) + \lambda^\mathsf{T}(t+1) B y^* \end{aligned}$$

Adding these inequalities and using $Ax^* + By^* = c$, $Ax(t+1) + By(t+1) = r(t+1) + c$, we have

$$p(t+1) + \lambda^\mathsf{T}(t+1)\left(r(t+1) + c\right) - s^\mathsf{T}(t+1)\left(x(t+1) - x^*\right) \;\leq\; p^* + \lambda^\mathsf{T}(t+1) c$$

which proves the upper bound in (8.145).

Step 2: prove (8.146). We will take $(y(t), \lambda(t))$ as the state of the ADMM algorithm and treat $x(t+1)$ as an intermediate quantity as a function of $(y(t), \lambda(t))$. Then (8.141) describes the state evolution from $(y(t), \lambda(t))$ to $(y(t+1), \lambda(t+1))$. Define the Lyapunov function for this dynamical system:

$$V(t) \;:=\; \rho \|B(y(t) - y^*)\|_2^2 + \frac{1}{\rho}\|\lambda(t) - \lambda^*\|_2^2$$

Therefore $V(t) \geq 0$ with equality if and only if $y(t) = y^*$ and $\lambda(t) = \lambda^*$ (It is possible however that $V(t)$ converges to a constant $V^* > 0$ if $(y(t), \lambda(t))$ converges to a different saddle point $(\tilde{y}, \tilde{\lambda})$.) We will lower bound the decrement in $V(t)$ with each iteration:

$$V(t+1) - V(t) \;\leq\; -\rho\|B(y(t+1) - y(t))\|_2^2 - \rho\|r(t+1)\|_2^2 \tag{8.146}$$

This requires a tighter analysis than using triangular inequality of $\|\cdot\|_2$.

The inequalities in (8.145) imply (substituting $s(t) := \rho A^\mathsf{T} B(y(t) - y(t-1))$)

$$\left(\lambda(t+1) - \lambda^*\right)^\mathsf{T} r(t+1) - \rho\left(y(t+1) - y(t)\right)^\mathsf{T} B^\mathsf{T} A\left(x(t+1) - x^*\right) \;\leq\; 0$$

Eliminate x^* and $x(t+1)$ using $Ax^* = c - By^*$ and $Ax(t+1) = c - By(t+1) + r(t+1)$ to get

$$\left(\lambda(t+1) - \lambda^*\right)^\mathsf{T} r(t+1) + \rho\left(y(t+1) - y(t)\right)^\mathsf{T} B^\mathsf{T}\left(B(y(t+1) - y^*) - r(t+1)\right) \leq 0 \tag{8.147}$$

in terms of $(y(t), \lambda(t))$. We now use (8.147) to prove (8.146).

Write

$$V(t+1) - V(t) \;=\; -\rho\Delta V_1(t) - \frac{1}{\rho}\Delta V_2(t) \tag{8.148a}$$

where

$$\Delta V_1(t) \;:=\; \|B(y(t)-y^*)\|_2^2 - \|B(y(t+1)-y^*)\|_2^2 \tag{8.148b}$$
$$\Delta V_2(t) \;:=\; \|\lambda(t)-\lambda^*\|_2^2 - \|\lambda(t+1)-\lambda^*\|_2^2 \tag{8.148c}$$

Substituting $y(t)-y^* = (y(t)-y(t+1))+(y(t+1)-y^*)$ into (8.148b) and expanding $\|B(y(t)-y^*)\|_2^2$ and similarly for $\|\lambda(t)-\lambda^*\|_2^2$ in (8.148c), we have

$$\begin{aligned}\Delta V_1(t) &= \|B(y(t+1)-y(t))\|_2^2 - 2(B(y(t+1)-y(t)))^{\mathsf{T}} B(y(t+1)-y^*)\\ \Delta V_2(t) &= \|\lambda(t+1)-\lambda(t)\|_2^2 - 2(\lambda(t+1)-\lambda(t))^{\mathsf{T}}(\lambda(t+1)-\lambda^*)\\ &= \|\rho r(t+1)\|_2^2 - 2\rho r^{\mathsf{T}}(t+1)(\lambda(t+1)-\lambda^*)\end{aligned}$$

where the last equality follows from $\lambda(t+1)=\lambda(t)+\rho r(t+1)$. Substituting into (8.148a) gives:

$$V(t+1)-V(t) \;=\; -\rho\|B(y(t+1)-y(t))\|_2^2 - \rho\|r(t+1)\|_2^2 + 2Z$$

where

$$Z \;:=\; r^{\mathsf{T}}(t+1)(\lambda(t+1)-\lambda^*) + \rho(y(t+1)-y(t))^{\mathsf{T}} B^{\mathsf{T}} B(y(t+1)-y^*)$$

It therefore suffices to show that $Z \le 0$ to establish (8.146). From (8.147) we have

$$Z \;\le\; \rho(y(t+1)-y(t))^{\mathsf{T}} B^{\mathsf{T}} r(t+1)$$

We now show that $(y(t+1)-y(t))^{\mathsf{T}} B^{\mathsf{T}} r(t+1) \le 0$.

Recall that $y(t+1)$ minimizes $L_\rho(x(t+1), y, \lambda(t))$ over $y \in \mathbb{R}^{n_2}$ and satisfies (8.144a):

$$0 \;=\; \nabla g(y(t+1)) + B^{\mathsf{T}}\lambda(t) + \rho B^{\mathsf{T}} r(t+1)$$

Multiplying both sides by $(y(t+1)-y(t))^{\mathsf{T}}$ and rearranging we have

$$\begin{aligned}\rho(y(t+1)-y(t))^{\mathsf{T}} B^{\mathsf{T}} r(t+1) &= \nabla^{\mathsf{T}} g(y(t+1))(y(t)-y(t+1)) - (y(t+1)-y(t))^{\mathsf{T}} B^{\mathsf{T}}\lambda(t)\\ &\le g(y(t)) - g(y(t+1)) - (y(t+1)-y(t))^{\mathsf{T}} B^{\mathsf{T}}\lambda(t)\\ &= \Big(g(y(t)) + \lambda^{\mathsf{T}}(t) B y(t)\Big) - \Big(g(y(t+1)) + \lambda^{\mathsf{T}}(t) B y(t+1)\Big)\\ &\le 0\end{aligned}$$

where the first inequality follows from the convexity of g and the last inequality follows from the observation above that (8.144a) implies that $y(t)$ minimizes $g(y)+\lambda^{\mathsf{T}}(t)By$ over $y \in \mathbb{R}^{n_2}$. Hence $Z \le 0$. This completes the proof of (8.146).

Step 3: prove $r(t) \to 0$, $s(t) \to 0$, *and* $p(t) \to p^*$. Iterating on (8.146) gives

$$V(t) - V(0) \;\le\; -\rho \sum_{\tau=1}^{t} \Big(\|B(y(\tau)-y(\tau-1))\|_2^2 + \|r(\tau)\|_2^2\Big)$$

Hence $0 \le V(t) \le V(0) - \rho\sum_{\tau=1}^{t}\left(\|B(y(\tau)-y(\tau-1))\|_2^2 + \|r(\tau)\|_2^2\right)$. Taking the limit

we have

$$\sum_{\tau=1}^{\infty}\left(\|B(y(\tau)-y(\tau-1))\|_2^2+\|r(\tau)\|_2^2\right) \;\le\; V(0)$$

implying that $r(t)\to 0$ and $s(t):=\rho A^{\mathsf{T}}B(y(t)-y(t-1))\to 0$. (Note that this does not imply $V(t)\to 0$, nor $(x(t),y(t))\to(x^*,y^*)$, since the series sum may be strictly less than $V(0)$.)

To prove $p(t)\to p^*$, note that $V(t)$ remaining finite as $t\to\infty$ means that $\lambda(t)$ and $y(t)$ remain finite as $t\to\infty$. Since $r(t)=Ax(t)+By(t)-c$ is finite, $Ax(t)$ remains finite as $t\to\infty$. Then, since the second term in the upper bound in (8.145) is

$$s^{\mathsf{T}}(t+1)\,(x(t+1)-x^*) \;=\; \rho\,(B(y(t+1)-y(t)))^{\mathsf{T}}\,A\,(x(t+1)-x^*)$$

$r(t)\to 0$ and $B(y(t)-y(t-1))\to 0$ imply that $p(t)\to p^*$ in view of (8.145).

Finally suppose a subsequence of $(x(t),y(t),\lambda(t))$ converges to $(\tilde{x},\tilde{y},\tilde{\lambda})$. Then it is proved in Chapter 8.5.5 that $(\tilde{x},\tilde{y},\tilde{\lambda})$ is a saddle point of L_0 and hence is primal-dual optimal for (8.140). □

8.7 Bibliographical notes

There are many excellent texts on convex analysis and optimization. We have used materials from [61, 62, 57, 54]. The envelope theorems in Chapter 8.3.6 are from [56] and [61, Proposition A.43, p.649]. See [63, Theorems 1, 2, 3] for envelope theorems that allow nonunique maximizer $x^*(p)$ but requires an upper bound on $|\partial f(x,p)/\partial p_i|$ uniformly in p_i. The main reference for Benders decomposition in Chapter 8.5.7 is [64]. See [59] for generalized Benders decomposition when (8.109c) is a convex program. The convergence analysis in Chapters 8.6.1–8.6.3 mostly follow [61], the analysis on interior-point method in Chapter 8.6.4 is from [57], and that on ADMM in Chapter 8.6.5 is from [65, Appendix A].

Interior-point methods were first employed to solve power system problems in the early 1990s for the purpose of state estimation [66]. See [67] for empirical performance of interior-point methods for large-scale OPF problems.

8.8 Problems

Chapter 8.1.

Exercise 8.1 (Convex sets)**.** Prove that the following sets are convex:

1. *Affine set:* $C=\{x\in\mathbb{R}^n \mid Ax=b\}$ where $A\in\mathbb{R}^{m\times n}$ and $b\in\mathbb{R}^m$, $m,n\ge 1$.
2. *Second-order cone:* $C=\{(x,t)\in\mathbb{R}^{n+1} \mid \|x\|_2\le t\}$, $n\ge 1$. Here $\|x\|_2:=\sqrt{x_1^2+x_2^2+\cdots+x_n^2}$ is the Euclidean norm.

3. *Positive semidefinite matrices:* $C = \{A \in \mathbb{S}^{n\times n} \mid A \succeq 0\}$, $n \geq 1$, where $\mathbb{S}^{n\times n}$ is the set of symmetric $n \times n$ real matrices and $A \succeq 0$ means $x^{\mathsf{T}}Ax \geq 0$ for any $x \in \mathbb{R}^n$.

Exercise 8.2 (Operations preserving set convexity)**.** Let $\mathbb{X}$ and $\mathbb{Y}$ be linear subspaces.

1. *Linear transformation:* Let $f : \mathbb{X} \to \mathbb{Y}$ be linear. Prove:
   1 If $A \subseteq \mathbb{X}$ is convex then $f(A) := \{f(x) : x \in A\}$ is convex.
   2 If $B \subseteq \mathbb{Y}$ is convex then $f^{-1}(B) = \{x \in \mathbb{R}^n : f(x) \in B\}$ is convex.
2. *Arbitrary direct product:* Let $A \subseteq \mathbb{X}$, $B \subseteq \mathbb{Y}$ be convex.
   1. Prove that the product space $\mathbb{X}\times\mathbb{Y} := \{(x,y) : x \in \mathbb{X}, y \in \mathbb{Y}\}$ with $+$ and $\cdot$ defined by
   $$\begin{aligned} (x_1,y_1)+(x_2,y_2) &:= (x_1+x_2, y_1+y_2) && \forall (x_1,y_1),\ (x_2,y_2) \in \mathbb{X}\times\mathbb{Y}; \\ \lambda(x,y) &:= (\lambda x, \lambda y) && \forall \lambda \in \mathbb{R},\ \forall (x,y) \in \mathbb{X}\times\mathbb{Y} \end{aligned}$$
   is also a linear space.
   2. Prove that the direct product $A \times B := \{(x,y) : x \in A, y \in B\}$ is convex. In fact the direct product of an arbitrary number of convex sets is convex.
3. *Finite sum:* Let $A, B \subseteq \mathbb{X}$ be convex. Prove that the set $A+B := \{a+b : a \in A, b \in B\}$ is convex. Therefore the sum of any finite number of convex sets is convex.
4. *Arbitrary intersection:* Let $A, B \subseteq \mathbb{X}$ be convex. Prove that the intersection $A \cap B$ is convex. In fact the intersection of an arbitrary collection of convex sets is convex.
5. *Union can be nonconvex.* Let $A, B \subseteq \mathbb{X}$ be convex. Give an example where the union $A \cup B$ is nonconvex. [Hint: Consider $\mathbb{X} = \mathbb{R}$].

Exercise 8.3 (Directional derivatives and differentiability)**.** Show that

$$f(x,y) := \begin{cases} \frac{x^a y^a}{x^{2a}+y^{2a}} & \text{if} \quad (x,y) \neq (0,0) \\ 0 & \text{if} \quad (x,y) = (0,0) \end{cases}$$

is not continuous, and hence not differentiable, at the origin.

Exercise 8.4 (Convex functions)**.** Prove that the following functions are convex:

1. *Exponential:* $f(x) := e^{ax}$ where $a, x \in \mathbb{R}$.
2. *Entropy:* $f(x) := x \ln x$ defined on $\mathbb{R}_{++} := (0, \infty)$.
3. *Log-exponential:* $f(x_1, x_2) := \ln(e^{x_1} + e^{x_2})$, $x_i \in \mathbb{R}$.

Exercise 8.5 (Convex functions)**.** [57, Exercise 3.6] For each of the following functions determine if it is convex, concave, or neither.

- $f(x) = e^x - 1$ on $\mathbb{R}$.
- $f(x) = x_1 x_2$ on $\{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 > 0, x_2 > 0\}$.
- $f(x) = \frac{1}{x_1 x_2}$ on $\{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 > 0, x_2 > 0\}$.
- $f(x) = x_1 / x_2$ on $\{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 > 0, x_2 > 0\}$.

Exercise 8.6 (Convexity tests)**.** Verify the convexity conditions of Theorem 8.2 on

$$f(x) \;:=\; f(x_1, x_2) := x_1^2 - 4x_1x_2 + 4x_2^2 \;=\; (x_1 - 2x_2)^2$$

Exercise 8.7 (Strict convexity)**.** Prove Corollary 8.3.

Exercise 8.8 (Operations preserving function convexity)**.** Suppose f_1 and f_2 are two convex functions on the same domain. Prove that:

1. $f := \alpha f_1 + \beta f_2$, $\alpha, \beta \geq 0$, is convex.
2. $f := \max\{f_1, f_2\}$ is convex.
3. $f(x, y) := |x| + |y|$ defined on $\mathbb{R}^2$ is convex. [Hint: use part 2.]
4. $f(g(x))$ is convex if $g : \mathbb{R}^n \to \mathbb{R}^m$ is convex (componentwise) and $f : \mathbb{R}^m \to \mathbb{R}$ is convex and nondecreasing (componentwise), i.e., $f(y_1) \leq f(y_2)$ for $y_1, y_2 \in \mathbb{R}^m$ with $y_1 \leq y_2$.

Exercise 8.9 (Level set and convex problem)**.** 1 *Level set.* Let $f : C \to \mathbb{R}$ where $C \subseteq \mathbb{R}^n$. Prove that the level set $\{x \in C \mid f(x) \leq \alpha\}$ is convex for any $\alpha \in \mathbb{R}$ provided that C is a convex set and f is a convex function.

2 *Convex problem.* Consider

$$\min_x \;\; f(x) \qquad \text{s.t.} \qquad Ax = b, \;\; g_i(x) \leq 0, \qquad i = 1, \ldots, k$$

where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $k \geq 1$, and $f, g_1, \ldots, g_k$ are scalar functions defined on $\mathbb{R}^n$. Prove that if $f, g_1, g_2, \ldots, g_k$ are convex then the feasible set

$$X \;:=\; \{x \in \mathbb{R}^n \mid Ax = b, \; g_i(x) \leq 0, i = 1, \ldots, k\}$$

is convex.

Chapter 8.2.

Exercise 8.10 (Carathéodory theorem)**.** Prove Theorem 8.7.

Exercise 8.11 (Second-order cone)**.** 1 The second-order cone $K_{\text{soc}} = \tilde{K} \cap H$ where $\tilde{K} := \{(x,t) \in \mathbb{R}^{n+1} : \|x\|_2^2 \le t^2\}$ and $H := \{(x,t) : t \ge 0\}$ is a halfspace. Show that while K_{soc} is a convex cone, $\tilde{K}$ is a cone but nonconvex.

2 Show that $h_1(x,t) := \|x\|_2 - t$ is a convex function while $h_2(x,t) := \|x\|_2^2 - t^2$ is nonconvex.

Exercise 8.12 (Rotated second-order cone)**.** Show that the rotated second-order cone

$$K_{\text{rsoc}} \;:=\; \left\{(x,y,z) \in \mathbb{R}^n \times \mathbb{R}^2 : \|x\|_2^2 \le yz,\; y \ge 0,\; z \ge 0\right\}$$

is a linear transformation of the standard second-order cone

$$K_{\text{soc}} \;:=\; \left\{(w,t) \in \mathbb{R}^{n+1} \times \mathbb{R} : \|w\| \le t\right\}$$

i.e., $(w,t) = A(x,y,z) \in K_{\text{soc}} \subseteq \mathbb{R}^{n+2}$ if and only if $(x,y,z) \in K_{\text{rsoc}}$ for a $(n+2)\times(n+2)$ nonsingular matrix A. Derive A and its inverse.

Exercise 8.13 (SOC constraint)**.** Consider the second-order cone K_{soc} in Exercise 8.11 and the set defined in terms of K_{soc}:

$$C \;:=\; \{x : (Ax+b, c^{\mathsf{T}}x+d) \in K_{\text{soc}}\} \;=\; \{x : \|Ax+b\|_2 \le c^{\mathsf{T}}x+d\} \;\subseteq\; \mathbb{R}^m$$

where $A \in \mathbb{R}^{n\times m}$, $b \in \mathbb{R}^n$, $c \in \mathbb{R}^m$, and $d \in \mathbb{R}$. Since C is the pre-image of an affine function on K_{soc}, it is convex.

1 Verify directly the convexity of C using the definition of convex sets.

2 Write $C = \tilde{C} \cap H$ where $\tilde{C} := \{x : \|Ax+b\|_2^2 \le (c^{\mathsf{T}}x+d)^2\}$ and $H := \{x : c^{\mathsf{T}}x+d \ge 0\}$ is a halfspace. Give an example where $\tilde{C}$ is not convex and illustrate how the intersection with H yields a convex set.

Exercise 8.14 (Farkas Lemma)**.** Prove the following variant of Theorem 8.12: Exactly one of the following holds:

1 There exists an $x \ge 0$ such that $Ax \le b$.

2 There exists an $y \ge 0$ such that $y^{\mathsf{T}}A \ge 0$ and $y^{\mathsf{T}}b < 0$.

where $A \in \mathbb{R}^{m\times n}$, $b, y \in \mathbb{R}^m$ and $x \in \mathbb{R}^n$. (Hint: Consider $Y := \{y \in \mathbb{R}^m : Ax \le y, x \ge 0\} = \{Ax+s : x \ge 0, s \ge 0\}$.)

Chapter 8.3

Exercise 8.15 (Equivalent property of saddle point)**.** Consider the primal problem and its partial dual (8.34) with the undualized constraint set X', the dualized constraint set $X := \{x \in \mathbb{R}^n : g(x) = 0, h(x) \leq 0\}$ and dual feasible set $Y := \{(\lambda, \mu) \in \mathbb{R}^{m+l} : \mu \geq 0\}$. Show that $(x^*, \lambda^*, \mu^*) \in X' \times Y$ is a saddle point, i.e.,

$$\max_{(\lambda,\mu)\in Y} L(x^*, \lambda, \mu) \;=\; L(x^*, \lambda^*, \mu^*) \;=\; \min_{x\in X'} L(x, \lambda^*, \mu^*)$$

if and only if

$$L(x^*, \lambda^*, \mu^*) \;=\; \min_{x\in X'} L(x, \lambda^*, \mu^*), \qquad x^* \;\in\; X, \qquad \mu^{*\mathsf{T}} h(x^*) \;=\; 0$$

Exercise 8.16 (KKT condition)**.** This problem derives the KKT condition for the constrained optimization problem:

$$\textbf{(P)} \;:\quad \min_{x\in\mathbb{R}^n} \; f(x) \quad \text{s.t.} \quad Ax = b, \;\; h_i(x) \leq 0, \quad i = 1, \ldots, l$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $k \geq 1$, and $f, h_1, \ldots, h_l$ are scalar functions defined on $\mathbb{R}^n$. Let $\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^l_+ = [0, \infty)^l$, and define

$$L(x, \lambda, \mu) := f(x) + \lambda^{\mathsf{T}}(Ax - b) + \mu^{\mathsf{T}} h(x)$$

where $h(x) = (h_1(x), h_2(x), \ldots, h_l(x))^{\mathsf{T}}$.

1 *Unconstrained optimization.* Let $d(\lambda, \mu) := \min_{x\in\mathbb{R}^n} L(x, \lambda, \mu)$ denote the unconstrained optimization over x for fixed (λ, μ). Assume that Problem (P) has an optimal solution and denote it by x^*. Show that $d(\lambda, \mu) \leq f(x^*)$ for any $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^l_+$.

2 *Dual problem.* Consider the dual problem

$$\textbf{(D)} \;:\quad \max_{(\lambda,\mu)\in\mathbb{R}^{m+l}} \; d(\lambda, \mu) \quad \text{s.t.} \quad \mu \geq 0$$

Assume (D) has an optimal solution (λ^*, μ^*).

1. Show that $d(\lambda^*, \mu^*) - f(x^*) \leq \sum_{i=1}^l \mu_i^* h_i(x^*) \leq 0$. It implies that Problem (D) provides a lower bound for Problem (P). Note that this holds whether or not $f, h_1, \ldots, h_l$ are convex.
2. Assume now $f, h_1, \ldots, h_l$ are convex and differentiable. Show that the equality is attained, i.e., $d(\lambda^*, \mu^*) = f(x^*) + \sum_{i=1}^l \mu_i^* h_i(x^*)$, if and only if
$$\nabla_x L(x^*, \lambda^*, \mu^*) = 0$$
3. Show that if there exists (x, λ, μ) such that x is feasible for (P), (λ, μ) is feasible for (D), $\nabla_x L(x, \lambda, \mu) = 0$, and $\mu_i h_i(x) = 0$ for $i = 1, \ldots, l$, then x solves (P) and (λ, μ) solves (D). These are the KKT conditions.

Exercise 8.17 (LICQ implies MFCQ)**.** Suppose x^* is a local optimal of the constrained optimization problem (8.25). Let $\bar{Y}(x^*)$ be the set of Lagrange multipliers associated with x^*:

$$\bar{Y}(x^*) := \left\{ (\lambda, \mu) \in \mathbb{R}^{m+l} : \frac{\partial L}{\partial x}(x^*, \lambda, \mu) = 0, g(x^*) = 0, h(x^*) \leq 0, \mu \geq 0, \mu^\mathsf{T} h(x^*) = 0 \right\}$$

Prove that the linear independence constraint qualification (8.43) implies the Mangasarian-Fromovitz constraint qualification (8.42). (Hint: Use the Farkas Lemma 8.12.)

Exercise 8.18 (Slater Theorem)**.** For

$$f^* := \inf_{x \in \mathbb{R}} f(x) := e^{-x} \quad \text{s.t.} \quad x = 0$$

check that the conditions in the Slater Theorem 8.17 are satisfied and derive the primal-dual optimal solution (x^*, λ^*).

Exercise 8.19 (Slater Theorem: dual optimal set)**.** [55, Lemma 1] Consider the following primal problem with only the inequality constraint and its dual:[13]

$$f^* := \min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t.} \quad h(x) \leq 0$$

$$d^* := \max_{\mu \geq 0} d(\mu) := \max_{\mu \geq 0} \left(\inf_{x \in \mathbb{R}^n} f(x) + \mu^\mathsf{T} h(x) \right)$$

Suppose:

- *Convexity*: f, h are convex.
- *Finite primal value*: f^* is finite, i.e., $-\infty < f^* < \infty$.
- *Strict Slater condition*: there exists $\bar{x}$ such that $h(\bar{x}) < 0$.

Then Theorem 8.17 says that strong duality holds and dual optimal solutions μ^* exist. Fix any $\bar{\mu} \in \mathbb{R}^l$ with $\bar{\mu} \geq 0$ and let $\bar{D} := \{\mu \in \mathbb{R}^l : d(\mu) \geq d(\bar{\mu})\}$ be the level set defined by $\bar{\mu}$. Let D^* denote the set of dual optimal solutions. Show that [55, Lemma 1]:

1 The level set $\bar{D}$ is compact and convex.
2 The dual optimal set D^* is compact and convex. In particular D^* is bounded by the weak duality gap at the strict Slater point $\bar{x}$ divided by the worst-case "constraint gap":

$$\max_{\mu \in D^*} \|\mu\|_2 \leq \max_{\mu \in D^*} \|\mu\|_1 \leq \frac{f(\bar{x}) - d^*}{\min_i(-h_i(\bar{x}))} = \frac{f(\bar{x}) - f^*}{\min_i(-h_i(\bar{x}))}$$

The boundedness of the dual optimal set D^* is also proved in Lemma 12.30 in

[13] The absence of equality constraint is only important for the upper bound on $\|\mu\|$ below. That D^* is bounded can be proved without this assumption as in Lemma 12.30.

the context of MC/MC problem where $h(\bar{x}) < 0$ corresponds to the condition $0 \in \operatorname{int}\left(D_{\overline{M}}\right)$ (not just $0 \in \operatorname{ri}\left(D_{\overline{M}}\right)$). The argument there is by contradiction and does not provide an explicit bound on $\|\mu\|$.

The following problem studies Theorem 8.19 when the feasible set X_p depends on p. It shows that the theorem generally no longer holds.

Exercise 8.20 (Saddle-point envelope theorem)**.** Consider the master problem:

$$\min_x \; f(x) := (x-p)^2 \;\text{ s. t. }\; \frac{p}{4} \le x \le \frac{p}{2} \tag{8.149}$$

for $p \in P := (0,2)$. Clearly the unique minimizer $x^*(p) = p/2$. We study three ways to dualize, resulting in different Lagrangian functions, X_p, and saddle points.

1 Dualize both constraints with dual variables $y := (y_1, y_2) \ge 0$ and the Lagrangian

$$L(x,y;p) := f(x) + y_1\left(\frac{p}{4} - x\right) + y_2\left(x - \frac{p}{2}\right)$$

Exhibit that Theorem 8.19 holds.

2 Consider the form of (8.149)

$$\min_{x \in X_p} \; f(x) := (x-p)^2 \;\text{ s. t. }\; x \ge \frac{p}{4} \tag{8.150}$$

with $X_p := \{x : x \le p/2\}$, and Lagrangian

$$L_1(x,y_1;p) := f(x) + y_1\left(\frac{p}{4} - x\right)$$

Show that Theorem 8.19 does not hold because of the reason explained in Remark 8.7 (even though all other conditions in Theorem 8.19 hold).

3 Consider the following form of (8.149)

$$\min_{x \in X_p} \; f(x) := (x-p)^2 \;\text{ s. t. }\; x \le \frac{p}{2} \tag{8.151}$$

with $X_p := \{x : x \ge p/4\}$, and Lagrangian

$$L_2(x,y_2;p) := f(x) + y_2\left(x - \frac{p}{2}\right)$$

Show that Theorem 8.19 holds because $x^*(p) \in X_q$ for all $p, q \in P$.

Chapter 8.4.

Exercise 8.21 (Convex programs)**.** Show how the different classes of convex problems in Figure 8.14 reduce to each other.

Exercise 8.22 (LP duality)**.** Consider the linear program (8.56a) and suppose $-\infty < f^* < \infty$. Lemma 8.22 then implies the existence of an optimal primal solution $x^* \in X$. Use Farkas Lemma (Theorem 8.12) to show that there exists a dual optimal solution $\mu^* \in Y$ that closes the duality gap, i.e., $f^* = d^* = b^\mathsf{T}\mu^*$.

Exercise 8.23 (Unconstrained quadratic program)**.** This exercise proves step by step Theorem 8.24 on unconstrained convex QP:

$$f_1^* \;:=\; \min_{x\in\mathbb{R}^n} f(x) \;:=\; x^\mathsf{T}Qx \,+\, 2c^\mathsf{T}x$$

where $Q \succeq 0$ and $c \in \mathbb{R}^n$.

1 Suppose $Q \succ 0$ is positive definite. Show that the unique minimizer x^* and the minimum value f_1^* are respectively

$$x^* \;=\; -Q^{-1}c, \qquad\qquad f_1^* = -c^\mathsf{T}Q^{-1}c$$

2 Suppose $Q \succeq 0$ but not positive definite. Let the spectral decomposition of Q be

$$Q \;=\; U\Lambda U^\mathsf{T} \;=\; \begin{bmatrix} U_r & U_{n-r}\end{bmatrix}\begin{bmatrix}\Lambda_r & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix} U_r^\mathsf{T}\\ U_{n-r}^\mathsf{T}\end{bmatrix} \;=\; U_r\Lambda_r U_r^\mathsf{T}$$

Write $Q = R^\mathsf{T}R$ where $R := \Lambda_r^{1/2}U_r^\mathsf{T} \in \mathbb{R}^{r\times n}$.

1 Show that it is possible to complete the square, i.e., write

$$f(x) \;=\; x^\mathsf{T}R^\mathsf{T}Rx + 2c^\mathsf{T}x \;=\; \|Rx+\tilde{c}\|_2^2 - \|\tilde{c}\|_2^2$$

if and only if $c \in \mathrm{range}(Q)$. Determine $\tilde{c}$.

2 Show that if $c \in \mathrm{range}(Q)$ then the set of minimizers x^* and the minimum value f_1^* are respectively

$$x^* \;=\; -Q^\dagger c + \mathrm{null}(Q), \qquad\qquad f_1^* = -c^\mathsf{T}Q^\dagger c$$

where $Q^\dagger := U_r\Lambda_r^{-1}U_r^\mathsf{T}$ is the pseudo-inverse of Q.

3 Show that if $c \notin \mathrm{range}(Q)$ then $f_1^* = -\infty$. (Hint: Transform to the coordinate defined by the basis U.)

Exercise 8.24 (Constrained quadratic program)**.** This exercise proves a slightly more general version of Theorem 8.25 step by step for the affinely constrained convex QP:

$$f_2^* \;:=\; \min_{x\in\mathbb{R}^n} f(x) \;:=\; x^\mathsf{T}Qx \,+\, 2c^\mathsf{T}x \quad \text{s.t.} \quad Ax = b,\; Bx + d \geq 0$$

where $Q \succeq 0$, $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $B \in \mathbb{R}^{l\times n}$ and $d \in \mathbb{R}^l$. Here we replace the condition $Q \succ 0$ by the weaker condition $f_2^* > -\infty$.

1 *Dual problem*. Show that the Lagrangian dual problem is:

$$d^* \;:=\; -c^\mathsf{T} Q^\dagger c \;-\; \min_{\lambda\in\mathbb{R}^m,\,\mu\in\mathbb{R}^k_+} \left(\begin{bmatrix}\lambda^\mathsf{T} & \mu^\mathsf{T}\end{bmatrix} \hat{Q} \begin{bmatrix}\lambda\\ \mu\end{bmatrix} + 2\hat{c}^\mathsf{T} \begin{bmatrix}\lambda\\ \mu\end{bmatrix} \right)$$

where $\mathbb{R}^l_+ := \{\mu \in \mathbb{R}^l : \mu \geq 0\}$ and

$$\hat{Q} \;:=\; \begin{bmatrix}A\\ B\end{bmatrix} Q^\dagger \begin{bmatrix}A^\mathsf{T} & B^\mathsf{T}\end{bmatrix}, \qquad \hat{c} \;:=\; \begin{bmatrix}-b\\ +d\end{bmatrix} - \begin{bmatrix}A\\ B\end{bmatrix} Q^\dagger c \tag{8.152}$$

2 *Strong duality, dual optimality, KKT condition*. Show that strong duality holds and dual optimality is attained. Moreover a feasible x^* is optimal if and only if there exists $(\lambda^*, \mu^*) \in \mathbb{R}^{m+l}$ such that $\mu^* \geq 0$ and

$$A^\mathsf{T}\lambda^* + B^\mathsf{T}\mu^* - Qx^* \;=\; c, \qquad\qquad \mu^{*\mathsf{T}}(Bx^* + d) \;=\; 0$$

Exercise 8.25 (QCQP)**.** Consider the convex quadratically constrained quadratic program (QCQP):

$$f^* \;:=\; \min_{x\in\mathbb{R}^n} \; f(x) \;:=\; x^\mathsf{T} Q_0 x \;+\; 2c_0^\mathsf{T} x \quad \text{s.t.} \quad x^\mathsf{T} Q_1 x \;+\; 2c_1^\mathsf{T} x \leq d$$

where $Q_0 \succ 0$ is positive definite, $Q_1 \succeq 0$ is positive semidefinite, $c_0, c_1 \in \mathbb{R}^n$ and $d \in \mathbb{R}$.

1 *Dual problem*. Show that the Lagrangian dual problem is:

$$d^* \;:=\; -\min_{\mu\in\mathbb{R}_+} \; d\mu + (c_0 + \mu c_1)^\mathsf{T} (Q_0 + \mu Q_1)^{-1} (c_0 + \mu c_1)$$

2 *Strong duality, dual optimality, KKT condition*. Suppose f^* is finite and there exists $\bar{x}$ such that $\bar{x}^\mathsf{T} Q_1 \bar{x} \;+\; 2c_1^\mathsf{T}\bar{x} < d$. Show that strong duality holds and dual optimality is attained. Moreover a feasible x^* is optimal if and only if there exists $\mu^* \in \mathbb{R}$ such that $\mu^* \geq 0$ and

$$(Q_0 + \mu^* Q_1)x^* + (c_0 + \mu^* c_1) \;=\; 0, \qquad \mu^*(x^{*\mathsf{T}} Q_1 x^* \;+\; 2c_1^\mathsf{T} x^* - d) \;=\; 0$$

Exercise 8.26 (Dual problem of SOCP)**.** For the second-order constraint problem (8.69):

1 Derive the dual problem. (Hint: Use (8.154): $\min_{x\in\mathbb{R}^n} (a\|x\|_2 - bx) = 0$ if $\|b\|_2 \leq a$ and $-\infty$ otherwise.)
2 When the cost function is linear $f(x) := c^\mathsf{T} x$, show that the dual problem is

$$d^* \;:=\; \max_{(\lambda,\gamma)\in\mathbb{R}^{m+l}} \; b^\mathsf{T}\lambda - \tilde{d}^\mathsf{T}\gamma \qquad \text{s.t.} \qquad A^\mathsf{T}\lambda + \tilde{B}^\mathsf{T}\gamma = c, \;\; \|\gamma^{l-1}\|_2 \leq \gamma_l$$

Exercise 8.27 (Equivalent representations: SOCP)**.** Consider SOCP (8.66) and an alternative representation (**??**) of SOCP, reproduced here

$$f_1^* := \min_{x\in\mathbb{R}^n} c^\mathsf{T}x \qquad \text{s.t.} \qquad Ax=b,\ \|x^{n-1}\|_2 \le x_n \tag{8.153a}$$

$$f_2^* := \min_{x\in\mathbb{R}^n} c^\mathsf{T}x \qquad \text{s.t.} \qquad Ax=b,\ \|x^{n-1}\|_2^2 \le x_n^2,\ x_n \ge 0 \tag{8.153b}$$

They are equivalent representations in the sense that they have the same cost function and feasible set. The constraint function $h_1(x) := \|x^{n-1}\|_2 - x_n$ in (8.153a) is nondifferentiable at $x=0$ and the constraint function $h_2(x) := \|x^{n-1}\|_2^2 - x_n^2$ in (8.153b) is nonconvex. In this exercise we show that they have different duality and optimality properties.

Separate the first $n-1$ columns of A from the last column and the first $n-1$ entries of $c - A^\mathsf{T}\lambda$ from the last:

$$A =: \begin{bmatrix} A^{n-1} & a_n \end{bmatrix}, \qquad \rho := \begin{bmatrix} \rho^{n-1} \\ \rho_n \end{bmatrix} := \begin{bmatrix} c^{n-1} - (A^{n-1})^\mathsf{T}\lambda \\ c_n - a_n^\mathsf{T}\lambda \end{bmatrix} := c - A^\mathsf{T}\lambda$$

1 Consider the SOCP (8.153a).

1 Show that, if $g(x) := a\|x\|_2 - b^\mathsf{T}x$, then

$$\min_{x\in\mathbb{R}^n} g(x) = \begin{cases} 0 & \text{if } \|b\|_2 \le a \\ -\infty & \text{otherwise} \end{cases} \tag{8.154}$$

2 Use (8.154) to show that the Lagrangian dual function of (8.153a) is

$$d_1(\lambda,\mu) := \min_{x\in\mathbb{R}^n} L(x,\lambda,\mu) = \begin{cases} \lambda^\mathsf{T}b & \text{if } \|\rho^{n-1}\|_2 \le \rho_n = \mu \\ -\infty & \text{otherwise} \end{cases}$$

and hence the dual problem is an SOCP:

$$d_1^* := \max_{\lambda\in\mathbb{R}^m} \lambda^\mathsf{T}b \qquad \text{s.t.} \qquad \|c^{n-1} - (A^{n-1})^\mathsf{T}\lambda\|_2 \le c_n - a_n^\mathsf{T}\lambda \tag{8.155}$$

2 Consider the SOCP (8.153b). Show that the Lagrangian dual function is:

$$d_2(\lambda,\mu) := \min_{x\in\mathbb{R}^n} L(x,\lambda,\mu) = \begin{cases} \lambda^\mathsf{T}b & \text{if } \rho^{n-1}=0,\ \rho_n=\mu\ge 0,\ \mu_1 = 0 \\ -\infty & \text{otherwise} \end{cases}$$

and hence the dual problem is a LP:

$$d_2^* := \max_{\lambda\in\mathbb{R}^m} \lambda^\mathsf{T}b \qquad \text{s.t.} \qquad (A^{n-1})^\mathsf{T}\lambda = c^{n-1},\ a_n^\mathsf{T}\lambda \le c_n$$

whose feasible set is a subset of that of (8.155).

3 *Strong duality and dual optimality*. Consider now the case where the constraint $Ax=b$ is absent in SOCP (8.153):

$$f_1^* := \min_{x\in\mathbb{R}^n} c^\mathsf{T}x \qquad \text{s.t.} \qquad \|x^{n-1}\|_2 \le x_n \tag{8.156}$$

$$f_2^* := \min_{x\in\mathbb{R}^n} c^\mathsf{T}x \qquad \text{s.t.} \qquad \|x^{n-1}\|_2^2 \le x_n^2,\ x_n \ge 0 \tag{8.157}$$

Show that, if $\|c^{n-1}\|_2 \le c_n$, then strong duality holds and dual optimality is attained

for (8.156), but, as long as $0 \neq \|c^{n-1}\|_2 \leq c_n$, $f_2^* = 0 > -\infty = d_2^*$, i.e., the duality gap is unbounded and the dual problem is infeasible, for (8.157).

Chapter 8.5.

Exercise 8.28 (Linear equality constraint)**.** For the quadratic program (8.81) over $\Delta x(t)$, Derive its KKT condition is given by (8.82a).

Exercise 8.29 (Interior-point method - strictly feasible point)**.** Consider the following problem to compute a strictly feasible point for (8.87):

$$\min_{(x,s)\in\mathbb{R}^{n+1}} \quad s \qquad \text{s.t.} \qquad g(x) = 0,\ h_i(x) \ \leq\ s,\ \ i = 1,\ldots,l \tag{8.158}$$

Assume (8.158) is feasible. Show that a strictly feasible point for (8.87) exists if and only if the optimal value s^{opt} of (8.158) is strictly negative (possibly $-\infty$), whether or not the minimum of (8.158) is attained.

Exercise 8.30 (Benders decomposition)**.** Prove Theorem 8.31.

Exercise 8.31 (Benders decomposition)**.** Suppose C8.3 and C8.4 hold. Then y_0 has no lower bound on G if and only if (8.114) is infeasible, provided $G \neq \emptyset$.

Exercise 8.32 (Benders decomposition)**.** Prove (8.117) reproduced here:

$$\bar{y}_0 \leq c^{\mathsf{T}}\bar{x} + f(\bar{y}) \ = \ (F(\bar{y}) - b)^{\mathsf{T}}\bar{\mu} + f(\bar{y}) \tag{8.159}$$

with equality if and only if $(\bar{y}_0, \bar{y}) \in G$. In particular, if strict inequality holds in (8.159), then there is a $(\mu_0', \mu') \in C \setminus Q$. (Hint: use LP duality Theorem 8.23 and Theorem 8.31).

Chapter 8.6.

9 Optimal power flow

As we see in Chapter 6 optimal power flow (OPF) is a fundamental problem that underlies numerous applications in power system operation and planning. In this chapter we study computational issues of OPF as a general constrained optimization that takes the form

$$\min_{u,x} \; c(u,x) \qquad \text{subject to} \qquad f(u,x)=0, \;\; g(u,x)\leq 0$$

The cost function c may represent generation cost, voltage deviation, power loss, or user disutility. The variable u collects control decisions such as generator commitment, generation setpoints, transformer taps, capacitor switch status, electric vehicle charging levels, thermostatic settings, or inverter reactive power. The variable x collects network state such as voltage levels, line currents, or power flows. The constraint functions f, g describe current or power balance, generation or consumption limits, voltage or line limits, and stability and security constraints, as well as other operational requirements.

In Chapter 9.1 we use the single-phase models of Part I to formulate OPF in the bus injection model. In Chapter 9.2 we formulate OPF in the branch flow model for radial networks and show that it is equivalent to OPF in the bus injection model. In Chapter 9.3 we prove that OPF is NP-hard and in Chapter 9.4 we prove that a subclass characterized by a Lyapunov-like condition can be solved efficiently to global optimality. In Chapter 9.5 we describe techniques for scaling OPF solutions. Popular algorithms for solving OPF problems are studied in Chapter 8.5 and example applications are discussed in Chapter 6.

9.1 Bus injection model

In Chapter 9.1.1 we describe how to represent different devices in terms of their nodal power injections and voltages (s_j, V_j). The interaction of these terminal variables over the network is described by power flow equations. We formulate in Chapter 9.1.2 OPF in the bus injection model and then express it in Chapter 9.1.3 as a standard quadratically constrained quadratic program.

9.1.1 Single-phase devices

For simplicity we will assume voltages are defined with respect to the ground and every single-phase device is connected between its terminal (bus) and the ground. We will model the devices we encounter by one of the following:

1 *Voltage source V_j:* An ideal voltage source j fixes its voltage $V_j \in \mathbb{C}$ if it is uncontrollable and it adjusts V_j if it is controllable.
2 *Current source I_j:* An ideal current source fixes its current $I_j \in \mathbb{C}$ if it is uncontrollable and it adjusts I_j if it is controllable. An example current source is a load model for an electric vehicle charger whose charging current is controllable.
3 *Power source s_j:* An ideal power source fixes its power injection $s_j \in \mathbb{C}$ if it is uncontrollable and adjusts s_j if it is controllable.
4 *Impedance z_j:* An impedance z_j connected between the terminal and the ground fixes the relationship between the nodal voltage and current $V_j = -z_j I_j$ where the negative sign indicates that I_j is defined in the direction of ground-to-terminal.

The bus injection model studied in Chapter 4 focuses on the nodal power injections and voltages (V_j, s_j) of these devices. The relation among the nodal variables at each bus j is $s_j = V_j \bar{I}_j$. The nodal variables at different buses interact with each other over the network through current balance equation $I = YV$ or power flow equations $s_j = f_j(V)$. We now formulate OPF for single-phase systems.

9.1.2 Single-phase OPF

Consider a single-phase network modeled as an undirected graph $G := (\overline{N}, E)$ where there are $N+1$ buses $j \in \overline{N} := \{0, 1, \ldots, N\}$ and M lines in E. Each line $(j,k) \in E$ is characterized by admittances $\left(y^s_{jk}, y^m_{jk}\right) \in \mathbb{C}^2$ and $\left(y^s_{kj}, y^m_{kj}\right) \in \mathbb{C}^2$. We now explain the variables, power flow equations, cost function, and constraints that define an OPF problem. As we will see the OPF formulation (9.5) below does not require assumption C4.1 that $y^s_{jk} = y^s_{kj}$. It can therefore accommodate single-phase transformers that have complex turns ratios.

OPF.

Without loss of generality we first make the following assumptions and present a simple OPF formulation:

1 The OPF involves only voltage sources and power sources.
2 There is exactly one single-phase device (voltage or power source) at each bus j. We therefore interchangeably refer to j as a bus, a node, a terminal or a device.

We will explain below how to relax these assumptions.

Under these assumptions, associated with each bus j is its bus (nodal) power injection s_j and voltage V_j. The vectors $s := (s_j, j \in \overline{N})$ and $V := (V_j, j \in \overline{N})$ are the optimization variables. The cost function $C(s,V)$ may represent the cost of generation (e.g. in economic dispatch), estimation error (e.g. in state estimation), line loss (e.g. in volt/var control in distribution systems), and user disutility (e.g., in demand response). For instance to minimize a weighted sum of real power generations we can use

$$C(s,V) \;:=\; \sum_{j:\text{gens}} c_j \operatorname{Re}(s_j)$$

To minimize the total real power loss over the network we can use

$$C(s,V) \;:=\; \sum_{j} \operatorname{Re}\left(s_j\right)$$

There are two type constraints on (s,V). The first is power flow equations, the complex form of which is derived in Chapter 4.2 as follows. The sending-end line currents from buses j to k in terms of V and those from buses k to j are given in (4.1a) and reproduced here:

$$I_{jk}(V) \;=\; y^s_{jk}(V_j - V_k) \;+\; y^m_{jk} V_j, \qquad I_{kj}(V) \;=\; y^s_{kj}(V_k - V_j) \;+\; y^m_{kj} V_k, \qquad (j,k) \in E \tag{9.1}$$

The sending-end complex power flow from buses j to k and that from buses k to j are respectively (from (4.2)):

$$S_{jk}(V) \;:=\; V_j \bar{I}_{jk}(V) \;=\; \bar{y}^s_{jk}\left(|V_j|^2 - V_j \bar{V}_k\right) \;+\; \bar{y}^m_{jk}|V_j|^2, \qquad (j,k) \in E \tag{9.2a}$$

$$S_{kj}(V) \;:=\; V_k \bar{I}_{kj}(V) \;=\; \bar{y}^s_{kj}\left(|V_k|^2 - V_k \bar{V}_j\right) \;+\; \bar{y}^m_{kj}|V_k|^2, \qquad (j,k) \in E \tag{9.2b}$$

The bus injection model in complex form is therefore (from (4.26a)):

$$s_j \;=\; \sum_{k:j\sim k} S_{jk}(V) \;:=\; \sum_{k:j\sim k} \bar{y}^s_{jk}\left(|V_j|^2 - V_j \bar{V}_k\right) \;+\; \bar{y}^m_{jj}|V_j|^2, \qquad j \in \overline{N} \tag{9.3}$$

where $y^m_{jj} := \sum_{k:j\sim k} y^m_{jk}$ are the total shunt admittances incident on buses j. Instead of the complex form (9.3), we can also use the polar form or the Cartesian form of power flow equations.

The second type of constraints on (s,V) is operational constraints. We will consider only three constraints:

1 *Injection limits*: These can represent generation or load capacity limits and take the form:

$$s^{\min}_j \;\le\; s_j \;\le\; s^{\max}_j, \qquad j \in \overline{N} \tag{9.4a}$$

where $s^{\min}_j, s^{\max}_j \in \mathbb{C}$ are given bounds on the injections at buses j. Recall that $a_1 + \mathbf{i}b_1 \le a_2 + \mathbf{i}b_2$ is a shorthand for two real inequalities $a_1 \le a_2$ and $b_1 \le b_2$.

2 *Voltage limits*: These are limits on voltage magnitudes:

$$v_j^{\min} \le |V_j|^2 \le v_j^{\max}, \qquad j \in \overline{N} \tag{9.4b}$$

where $v_j^{\min}, v_j^{\max} \in \mathbb{R}$ are given lower and upper bounds on the squared voltage magnitudes. We assume $v_j^{\min} > 0$ to avoid triviality.

3 *Line limits*: Thermal limits can be expressed in terms of line currents $\big(I_{jk}(V), I_{kj}(V)\big)$ in (9.1):

$$\left|y_{jk}^s(V_j - V_k) + y_{jk}^m V_j\right|^2 \le \ell_{jk}^{\max}, \quad \left|y_{kj}^s(V_k - V_j) + y_{kj}^m V_k\right|^2 \le \ell_{kj}^{\max}, \qquad (j,k) \in E \tag{9.4c}$$

which are quadratic inequalities in V.

Alternatively line limits can be expressed in terms of complex line power:

$$S_{jk}^{\min} \le S_{jk}(V) \le S_{jk}^{\max}, \quad S_{kj}^{\min} \le S_{kj}(V) \le S_{kj}^{\max}, \qquad (j,k) \in E$$

or in terms of apparent power:

$$|S_{jk}(V)| \le S_{jk}^{\max}, \qquad |S_{kj}(V)| \le S_{kj}^{\max}, \qquad (j,k) \in E$$

where $\big(S_{jk}(V), S_{kj}(V)\big)$ are given by (9.2). The limits on apparent power can be expressed in terms of a degree four polynomial in V which can be converted into quadratic constraints with additional variables (see Exercise 9.2).

Depending on the application there can be many more constraints, e.g., stability and security constraints, ramp limits, limits on battery state of charge and charging rates. For illustration purpose we will mostly restrict ourselves to these three types of constraints.

A simple OPF problem in the bus injection model is then

$$\min_{(s,V)} \quad C(s,V) \text{ s.t. } (9.3)(9.4) \tag{9.5}$$

Since the constraints (9.3)(9.4c) do not require assumption C4.1 that $y_{jk}^s = y_{kj}^s$, the OPF formulation (9.5) can accommodate single-phase transformers that have complex turns ratios.

Remark 9.1 (Uncontrollable parameters and reference voltage)**.** This is a general formulation that allows the power injection s_j and voltages V_j at every bus j to be optimization variables. If there is practically no bound on the injection at bus j then $s_j^{\min} := -\infty - \mathbf{i}\infty$ or $s_j^{\max} := \infty + \mathbf{i}\infty$ which removes the lower or upper bound on the function $s_j(V)$ of V. On the other hand the inequality constraints also allow the case where a quantity is not an optimization variable but a parameter, by setting $s_j^{\min} = s_j^{\max}$ to the specified value. For instance $s_j(V) = s_j^{\min} = s_j^{\max}$ may represent a given uncontrollable constant-power load or a given renewable generation. For the slack bus 0, unless otherwise specified, we always assume $V_0 := 1\angle 0°$ pu so that $v_0^{\min} = v_0^{\max} = 1$ and $s_0^{\min} = -\infty - \mathbf{i}\infty$, $s_0^{\max} = \infty + \mathbf{i}\infty$. Therefore we sometimes replace $j \in \overline{N}$ in (9.3)(9.4) by $j \in N$. □

Other devices.

Single-phase devices other than voltage and power sources can also be included in the OPF formulation. For instance an electric vehicle charger can be modeled by a current source. If it is controllable then its current I_j is an additional optimization variable and it imposes a quadratic equality constraint on (s_j, V_j, I_j):

$$s_j = V_j I_j^{\mathsf{H}}$$

If the current source is uncontrollable with a fixed I_j, then the constraint above is a linear constraint on (s_j, V_j). A nodal impedance z_j introduces a quadratic equality constraint on (s_j, V_j):

$$s_j = -\frac{|V_j|^2}{\bar{z}_j}$$

where the negative sign indicates that the direction of s_j is ground-to-terminal through the impedance. A nodal admittance y_j, such as a capacitor tap, can be incorporated by including the the variable y_j and quadratic equality constraint on (s_j, V_j, y_j):

$$s_j = -\bar{y}_j |V_j|^2$$

where the negative sign indicates that the direction of s_j is ground-to-terminal through the admittance.

We assume in the OPF formulation (9.5) that each bus j has a single device with the nodal variable (s_j, V_j). If multiple devices are connected to bus j in parallel with power injections $s_{jk}, k = 1, \ldots, K_j$, they introduce additional variables $(s_{jk}, k = 1, \ldots, K_j)$ and impose the linear constraint

$$s_j = \sum_k s_{jk}$$

Hence other devices can be incorporated and they impose local constraints at each bus j. If a devices at bus j is controllable, it introduces an additional optimization variable u_j (e.g., I_j of a controllable current source) and a local constraint of the form

$$f_j(u_j, s_j, V_j) = 0, \qquad j \in \overline{N} \tag{9.6a}$$

Otherwise, it does not introduce additional variable at bus j (e.g., impedance z_j) and (9.6a) reduces to a local constraint of the form $f_j(s_j, V_j) = 0$ where the local device (e.g, z_j) is a parameter of the constraint function f_j. When an additional optimization variable u_j is introduced, there may also be an operational constraint on u_j of the form

$$g_j(u_j) \leq 0, \qquad j \in \overline{N} \tag{9.6b}$$

Most applications indeed involve other variables in addition to (s_j, V_j). For example, the unit commitment problem in Chapter 6.2.1 includes binary variables to indicate if a unit will be on or off. In distributed energy resource optimization, battery charging

rates and their states of charge as well as the temperature setpoint of a thermostat may be additional variables. In volt/var control that optimizes over the reactive power output of an inverter given its real power input, the reactive power needs to satisfy a sector constraint. For single-phase networks, however, we will focus on the simple OPF (9.5) and study its computational properties. In particular we will omit variables u_j and the associated local constraints (9.6).

OPF in terms of V only.

We can treat the power flow equation (9.3) as defining $s_j(V)$ as a function of V:

$$s_j(V) = \sum_{k:j\sim k} S_{jk}(V) \;:=\; \sum_{k:j\sim k} \bar{y}^s_{jk}\left(|V_j|^2 - V_j\bar{V}_k\right) + \bar{y}^m_{jj}|V_j|^2, \qquad j \in \overline{N} \quad (9.7)$$

where $y^m_{jj} := \sum_{k:j\sim k} y^m_{jk}$ are the total shunt admittances incident on buses j. Using (9.1)(9.2)(9.7) for single-phase networks, we can express powers and currents $\left(s_j, S_{jk}, I_{jk}\right)$ in terms of voltages V and formulate OPF as an optimization over V only.

For instance the cost function to minimize a weighted sum of real power generations is:

$$C(V) \;:=\; \sum_{j:\text{gens}} c_j \operatorname{Re}(s_j(V)) = \sum_{j:\text{gens}} c_j \operatorname{Re}\left(\sum_{k:j\sim k} \bar{y}^s_{jk}\left(|V_j|^2 - V_j\bar{V}_k\right) + \bar{y}^m_{jj}|V_j|^2\right)$$

The cost function to minimize the total real power loss over the network is:

$$C(V) \;:=\; \sum_{j} \operatorname{Re}\left(\sum_{k:j\sim k} \bar{y}^s_{jk}\left(|V_j|^2 - V_j\bar{V}_k\right) + \bar{y}^m_{jj}|V_j|^2\right)$$

which can be shown to be a quadratic form $C(V) = V^{\mathsf{H}}\operatorname{Re}(Y)V$ in terms of the admittance matrix Y (Exercise 9.1). The total real power loss equals the total thermal $(r|I|^2)$ loss in the network lines if line shunt admittances are reactive, i.e., if y^m_{jk} and y^m_{kj} are pure imaginary:

$$C(V) \;:=\; \sum_{(j,k)\in E} r_{jk}\,|I^s_{jk}(V)|^2$$

where $r_{jk} := \operatorname{Re}\left(z^s_{jk}\right) = \operatorname{Re}\left(\left(y^s_{jk}\right)^{-1}\right)$ is the series resistance of the line and $I^s_{jk}(V) := y^s_{jk}(V_j - V_k)$ is the current through the series impedance of the line. All these costs are quadratic functions of V (Exercise 9.1).

For operational constraints, the voltage limits (9.4b) and the line limits (9.4c) are already quadratic inequalities in V. We can use (9.7) to express the injection limits $s^{\min}_j \le s_j(V) \le s^{\max}_j$ also as quadratic inequalities in V:

$$s^{\min}_j \;\le\; \sum_{k:j\sim k} \bar{y}^s_{jk}\left(|V_j|^2 - V_j\bar{V}_k\right) + \bar{y}^m_{jj}|V_j|^2 \;\le\; s^{\max}_j, \qquad j \in \overline{N} \quad (9.8)$$

If we use the polar form (4.27) BIM then the injection limits become:

$$p_j^{\min} \le \sum_{k:k\sim j}\left(g_{jk}^s + g_{jk}^m\right)|V_j|^2 - \sum_{k:k\sim j}|V_j||V_k|\left(g_{jk}^s\cos\theta_{jk} + b_{jk}^s\sin\theta_{jk}\right) \le p_j^{\max}, \qquad j \in \overline{N}$$

$$q_j^{\min} \le -\sum_{k:k\sim j}\left(b_{jk}^s + b_{jk}^m\right)|V_j|^2 - \sum_{k:k\sim j}|V_j||V_k|\left(g_{jk}^s\sin\theta_{jk} - b_{jk}^s\cos\theta_{jk}\right) \le q_j^{\max}, \qquad j \in \overline{N}$$

For notational simplicity only, we will mostly use the complex form (9.8) as injection limits.

The simple OPF (9.5) can be equivalently formulated in terms of V only:

$$\min_V \quad C(V) \quad \text{s.t. (9.8)(9.4b)(9.4c)} \tag{9.9}$$

As mentioned before, this formulation does not require assumption C4.1 that $y_{jk}^s = y_{kj}^s$ and hence can accommodate single-phase transformers that have complex turns ratios. To avoid triviality we will assume unless otherwise specified that OPF (9.9) is feasible.

9.1.3 OPF as QCQP

As we have seen above the constraints in OPF (9.9) are quadratic in V. We now explain how to express (9.9) as a quadratically constrained quadratic program (QCQP).

QCQP.

A QCQP is the following problem:

$$\min_{x\in\mathbb{C}^n} \; x^{\mathsf{H}}C_0x \tag{9.10a}$$

$$\text{s.t. } x^{\mathsf{H}}C_lx \le b_l, \qquad l = 1,\dots,L \tag{9.10b}$$

where $x \in \mathbb{C}^n$ is a vector, $C_l \in \mathbb{S}^n$ for $l = 0,\dots,L$, are Hermitian matrices so that $x^{\mathsf{H}}C_lx$ are real values, and $b_l \in \mathbb{R}$ are given scalars. If C_l, $l = 0,\dots,L$, are positive semidefinite (psd) then (9.10) is a convex QCQP. Otherwise it is nonconvex. If x^{opt} is optimal for (9.10), so is $-x^{\text{opt}}$.

The inequality constraints (9.10b) can include equality constraints ($a = b \Leftrightarrow a \le b, b \le a$). Sometimes equality constraints are specified explicitly as in

$$\begin{aligned}\min_{x\in\mathbb{C}^n} \; & x^{\mathsf{H}}C_0x \\ \text{s.t. } & x^{\mathsf{H}}C_lx \le b_l, \qquad l = 1,\dots,L \\ & x^{\mathsf{H}}\tilde{C}_lx = \tilde{b}_l, \qquad l = 1,\dots,\tilde{L}\end{aligned}$$

Remark 9.2 (Equivalent real QCQP)**.** In computing a solution of (9.10), the QCQP is first converted into a problem in the real domain. Indeed the complex QCQP (9.10) is

equivalent to the following QCQP in the real domain of twice the dimension (Exercise 9.5):

$$\min_{y\in\mathbb{R}^{2n}} \quad y^{\mathsf{T}} D_0 y \qquad \text{s.t.} \qquad y^{\mathsf{T}} D_l y \;\le\; b_l, \qquad l = 1,\dots,L \tag{9.11a}$$

where

$$y \;:=\; \begin{bmatrix} \mathrm{Re}(x) \\ \mathrm{Im}(x) \end{bmatrix}, \qquad D_l \;:=\; \begin{bmatrix} \mathrm{Re}(C_l) & -\mathrm{Im}(C_l) \\ \mathrm{Im}(C_l) & \mathrm{Re}(C_l) \end{bmatrix}, \qquad l = 0,1,\dots,L \tag{9.11b}$$

Note that D_l are symmetric matrices. □

The problem (9.10) is called a homogeneous QCQP because each term, called a monomial, in the polynomial $x^{\mathsf{H}} C_l x$ is of degree 2. An inhomogeneous QCQP contains monomials with degree 1 and takes the form

$$\min_{x\in\mathbb{C}^n} \; x^{\mathsf{H}} C_0 x \;+\; \left(c_0^{\mathsf{H}} x + x^{\mathsf{H}} c_0\right) \tag{9.12a}$$

$$\text{s.t.} \; x^{\mathsf{H}} C_l x \;+\; \left(c_l^{\mathsf{H}} x + x^{\mathsf{H}} c_l\right) \;\le\; b_l, \qquad l = 1,\dots,L \tag{9.12b}$$

Note that $\left(c_l^{\mathsf{H}} x + x^{\mathsf{H}} c_l\right)$ are real numbers. This problem can be homogenized by introducing a scalar complex variable $t \in \mathbb{C}$ because, if we set $x := \hat{x}\bar{t}$ and require $|t|^2 = 1$ (i.e., $t = e^{\mathbf{i}\theta}$ for some θ), then

$$x^{\mathsf{H}} C_l x + c_l^{\mathsf{H}} x + x^{\mathsf{H}} c_l \;=\; \hat{x}^{\mathsf{H}} C_l \hat{x} + c_l^{\mathsf{H}} (\hat{x}\bar{t}) + (\hat{x}\bar{t})^{\mathsf{H}} c_l \;=\; \begin{bmatrix} \hat{x}^{\mathsf{H}} & t^{\mathsf{H}} \end{bmatrix} \begin{bmatrix} C_l & c_l \\ c_l^{\mathsf{H}} & 0 \end{bmatrix} \begin{bmatrix} \hat{x} \\ t \end{bmatrix}$$

Hence the inhomogeneous QCQP (9.12) is equivalent to the following homogeneous QCQP with equality and inequality constraints:

$$\min_{\hat{x}\in\mathbb{C}^n, t\in\mathbb{C}} \; \begin{bmatrix} \hat{x}^{\mathsf{H}} & t^{\mathsf{H}} \end{bmatrix} \begin{bmatrix} C_0 & c_0 \\ c_0^{\mathsf{H}} & 0 \end{bmatrix} \begin{bmatrix} \hat{x} \\ t \end{bmatrix} \tag{9.13a}$$

$$\text{s.t.} \; \begin{bmatrix} \hat{x}^{\mathsf{H}} & t^{\mathsf{H}} \end{bmatrix} \begin{bmatrix} C_l & c_l \\ c_l^{\mathsf{H}} & 0 \end{bmatrix} \begin{bmatrix} \hat{x} \\ t \end{bmatrix} \;\le\; b_l, \qquad l = 1,\dots,L \tag{9.13b}$$

$$\begin{bmatrix} \hat{x}^{\mathsf{H}} & t^{\mathsf{H}} \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{x} \\ t \end{bmatrix} \;=\; 1 \tag{9.13c}$$

If $(\hat{x}^{\mathrm{opt}}, t^{\mathrm{opt}}) \in \mathbb{C}^{n+1}$ is optimal for (9.13), then the product $x^{\mathrm{opf}} := \hat{x}^{\mathrm{opt}} \bar{t}^{\mathrm{opt}} = \hat{x}^{\mathrm{opt}} e^{-\mathbf{i}\theta^{\mathrm{opt}}}$ is optimal for (9.12).

We will hence study, without loss of generality, homogeneous QCQP (9.10) with inequality constraints.

Remark 9.3 (Real QCQP)**.** If the variable x is in $\mathbb{R}^n$ instead of $\mathbb{C}^n$ and C_l are $n \times n$ real symmetric matrices, $l = 0,\dots,L$, then (9.10) is a real homogeneous QCQP:

$$\min_{x\in\mathbb{R}^n} \quad x^{\mathsf{T}} C_0 x \qquad \text{s.t.} \qquad x^{\mathsf{T}} C_l x \;\le\; b_l, \qquad l = 1,\dots,L$$

A real inhomogeneous QCQP

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \; & x^{\mathsf{T}} C_0 x + \left(c_0^{\mathsf{T}} x + x^{\mathsf{T}} c_0\right) \\ \text{s.t.} \; & x^{\mathsf{T}} C_l x + \left(c_l^{\mathsf{T}} x + x^{\mathsf{T}} c_l\right) \le b_l, \qquad l = 1,\ldots,L \end{aligned}$$

is equivalent to the following real homogeneous QCQP

$$\begin{aligned} \min_{\hat{x}\in\mathbb{R}^n, t\in\mathbb{R}} \; & \begin{bmatrix}\hat{x}^{\mathsf{T}} & t\end{bmatrix} \begin{bmatrix} C_0 & c_0 \\ c_0^{\mathsf{T}} & 0 \end{bmatrix} \begin{bmatrix}\hat{x} \\ t\end{bmatrix} \\ \text{s.t.} \; & \begin{bmatrix}\hat{x}^{\mathsf{T}} & t\end{bmatrix} \begin{bmatrix} C_l & c_l \\ c_l^{\mathsf{T}} & 0 \end{bmatrix} \begin{bmatrix}\hat{x} \\ t\end{bmatrix} \le b_l, \qquad l = 1,\ldots,L \\ & \begin{bmatrix}\hat{x}^{\mathsf{T}} & t\end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix}\hat{x} \\ t\end{bmatrix} = 1 \end{aligned}$$

in that, if $(\hat{x}^{\text{opt}}, t^{\text{opt}}) \in \mathbb{R}^{n+1}$ is optimal for the homogeneous QCQP, then $x^{\text{opt}} := \hat{x}^{\text{opt}} t^{\text{opt}}$ is optimal for the original nonhomogeneous QCPQ ($t^{\text{opt}} \in \{-1, 1\}$). □

Remark 9.4 (Linear and bilinear cost or constraints)**.** For any $l \ge 0$, $C_l = 0$ corresponds to a linear cost or constraint. It can be homogenized in exactly the same way above, i.e., (9.13) allows any of the matrices C_l to be zero. For example, in the scalar case $n = 1$, a linear constraint can be homogenized by setting $x := \hat{x}\bar{t}$ and requiring $|t|^2 = 1$, so that

$$c^{\mathsf{H}} x + x^{\mathsf{H}} c = c^{\mathsf{H}} (\hat{x}\bar{t}) + c(\hat{x}\bar{t})^{\mathsf{H}} = \begin{bmatrix}\hat{x}^{\mathsf{H}} & t^{\mathsf{H}}\end{bmatrix} \begin{bmatrix} 0 & c_l \\ c_l^{\mathsf{H}} & 0 \end{bmatrix} \begin{bmatrix}\hat{x} \\ t\end{bmatrix}, \qquad |t|^2 = 1 \tag{9.14a}$$

Note that the two linear terms must be complex conjugates of each other so that they sum to a real number. For a linear inequality $d^{\mathsf{H}} x \le b$ where $b := b_r + \mathbf{i} b_i$ is complex, we can rewrite it as two real inequalities:

$$\frac{1}{2}\left(d^{\mathsf{H}} x + x^{\mathsf{H}} d\right) \le b_r, \qquad \frac{1}{2\mathbf{i}}\left(d^{\mathsf{H}} x - x^{\mathsf{H}} d\right) \le b_i \tag{9.14b}$$

The first inequality takes the form of (9.14a) with $c := d/2$. The second inequality takes the form of (9.14a) with $c := \mathbf{i}d/2$.

A block bilinear term of the form $x^{\mathsf{H}} C y$ can be homogenized as follows. For any variables $(x, y) \in \mathbb{C}^{2n}$ and any square matrices $C, D \in \mathbb{C}^{n\times n}$

$$x^{\mathsf{H}} C y + y^{\mathsf{H}} D x = \begin{bmatrix}x^{\mathsf{H}} & y^{\mathsf{H}}\end{bmatrix} \begin{bmatrix} 0 & C \\ D & 0 \end{bmatrix} \begin{bmatrix}x \\ y\end{bmatrix} \tag{9.15}$$

Note that C and D may not be Hermitian of each other so that the product $x^{\mathsf{H}} C y + y^{\mathsf{H}} D x$ may be a complex number. Its real and imaginary parts can be written as quadratic forms of (x, y) in terms of the following Hermtian matrices respectively:

$$\Phi := \frac{1}{2}\begin{bmatrix} 0 & C + D^{\mathsf{H}} \\ C^{\mathsf{H}} + D & 0 \end{bmatrix}, \qquad \Psi := \frac{1}{2\mathbf{i}}\begin{bmatrix} 0 & C - D^{\mathsf{H}} \\ -C^{\mathsf{H}} + D & 0 \end{bmatrix}$$

We emphasize that we convert QCQPs to their homogenized form mainly so that

we can focus only on homogeneous QCQP in our study of structural properties. In computation, one may not convert an inhomogeneous constraint, especially a linear constraint, into a homogeneous quadratic constraint. □

Example 9.1 (Polynomial cost or constraints)**.** A polynomial can be expressed as a quadratic with auxiliary variables. Write the following as quadratic constraints:

1 $\left(|V_j|^2 - 1\right)^2 \le \epsilon$.
2 $a_0 x^3 + a_1 x^2 + a_2 x \le \alpha$ with $a_i, x, \alpha \in \mathbb{C}$.

Solution.

1 We have $\left(|V_j|^2 - 1\right)^2 \le \epsilon$ if and only if there exist $t_j \in \mathbb{C}$ such that $\left(V_j, t_j\right)$ satisfies

$$\left|t_j - 1\right|^2 \;\le\; \epsilon, \qquad\qquad t_j \;=\; |V_j|^2$$

which are quadratic equality and inequality constraints that can be homogenized as discussed above. Note that $t_j = V_j^2$ is not a quadratic form when $\left(V_j, t_j\right)$ are complex.

2 Let $x =: y + \mathbf{i}z$ with $y, z \in \mathbb{R}$. First convert the constraint into two real polynomial constraints in y and z, each of the form

$$\sum_{(i,j):i+j=3} b_{ij} y^i z^j \;+\; \sum_{(i,j):i+j=2} c_{ij} y^i z^j \;+\; \sum_{(i,j):i+j=1} d_{ij} y^i z^j \;\le\; \beta$$

for some real coefficients b_{ij}, c_{ij}, d_{ij} and real β. To write this as a quadratic constraint in $(y, z) \in \mathbb{R}^2$, introduce auxiliary variables $t = y^2$, $u = z^2$. Then write $y^3 = ty$, $y^2 z = tz$, $yz^2 = yu$, $z^3 = uz$. These quadratic expressions can then be homogenized as discussed above.

□

OPF as QCQP.

We now assume the cost function $C(V) := V^{\mathsf{H}} C_0 V$ is a quadratic form in V for some positive semidefinite matrix C_0. We can then express OPF (9.9) as a QCQP, by deriving the cost matrices C_l underlying the quadratic constraints (9.4b)(9.4c)(9.8).

1 *Injection limits*: To express the injection s_j in (9.8) as a quadratic form, use $I = YV$ to write

$$s_j = V_j I_j^{\mathsf{H}} \;=\; \left(e_j^{\mathsf{H}} V\right)\left(e_j^{\mathsf{H}} I\right)^{\mathsf{H}} \;=\; e_j^{\mathsf{H}}\, V V^{\mathsf{H}} Y^{\mathsf{H}} e_j$$

where e_j is the $(N+1)$-dimensional vector with 1 in the jth entry and 0 elsewhere.

Since $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, we have[1]

$$s_j \;=\; \mathrm{tr}\left(e_j^{\mathsf{H}} V V^{\mathsf{H}} Y^{\mathsf{H}} e_j\right) = \mathrm{tr}\left(\left(Y^{\mathsf{H}} e_j e_j^{\mathsf{H}}\right) V V^{\mathsf{H}}\right) \;=:\; V^{\mathsf{H}} Y_j^{\mathsf{H}} V$$

where $Y_j := e_j e_j^{\mathsf{H}} Y$ is an $(N+1)\times(N+1)$ matrix with its jth row equal to the jth row of the admittance matrix Y and all other rows equal to the zero vector. Y_j is not Hermitian so that $V^{\mathsf{H}} Y_j^{\mathsf{H}} V$ is in general a complex number. Its real and imaginary parts can be expressed in terms of the Hermitian and skew Hermitian components of Y_j^{H} defined as:

$$\Phi_j \;:=\; \frac{1}{2}\left(Y_j^{\mathsf{H}} + Y_j\right) \quad \text{and} \quad \Psi_j \;:=\; \frac{1}{2\mathbf{i}}\left(Y_j^{\mathsf{H}} - Y_j\right)$$

Then Φ_j and Ψ_j are Hermitian matrices and (Exercise 9.4)

$$\mathrm{Re}(s_j) \;=\; V^{\mathsf{H}} \Phi_j V \quad \text{and} \quad \mathrm{Im}(s_j) \;=\; V^{\mathsf{H}} \Psi_j V$$

They will be upper and lower bounded by

$$\begin{aligned} p_j^{\min} \;&:=\; \mathrm{Re}\; s_j^{\min} \quad \text{and} \quad p_j^{\max} \;:=\; \mathrm{Re}\; s_j^{\max} \\ q_j^{\min} \;&:=\; \mathrm{Im}\; s_j^{\min} \quad \text{and} \quad q_j^{\max} \;:=\; \mathrm{Im}\; s_j^{\max} \end{aligned}$$

These quantities will be used to rewrite below OPF as a standard QCQP of the form (9.10).

2 *Voltage limits*: Let $E_j := e_j e_j^{\mathsf{H}}$ denote the Hermitian matrix with a single 1 in the (j,j)th entry and 0 everywhere else. Then squared voltage magnitude $|V_j|^2 = V^{\mathsf{H}} E_j V$ is a quadratic form. It will be lower and upper bounded by $v_j^{\min}$ and $v_j^{\max}$ in (9.4b) respectively.

3 *Line limits*: For the first set of constraints in (9.4c), use (9.1) to write

$$I_{jk} = y_{jk}^s (V_j - V_k) \;+\; y_{jk}^m V_j \;\;=\;\; \left(y_{jk}^s (e_j - e_k)^{\mathsf{T}} \;+\; y_{jk}^m\, e_j^{\mathsf{T}}\right) V$$

Hence $|I_{jk}|^2 = V^{\mathsf{H}} \hat{Y}_{jk} V$, which will be upper bounded by $\ell_{jk}^{\max}$, where

$$\hat{Y}_{jk} := \left(\bar{y}_{jk}^s (e_j - e_k) \;+\; \bar{y}_{jk}^m\, e_j\right)\left(y_{jk}^s (e_j - e_k)^{\mathsf{T}} \;+\; y_{jk}^m\, e_j^{\mathsf{T}}\right)$$

The matrix $\hat{Y}_{jk}$ is Hermitian and hence $V^{\mathsf{H}} \hat{Y}_{jk} V$ is indeed a real number. Similarly for bounds on $|I_{kj}|^2$.

[1] The inner product of two complex matrices is defined to be $A \cdot B := \mathrm{tr}(A^{\mathsf{H}} B) = \sum_{i,j} \overline{A}_{ij} B_{ij}$ and is not equal to $\mathrm{tr}(AB) = \sum_{i,j} A_{ij} B_{ji}$ unless A is Hermitian; see Exercise 9.3.

Putting all this together, OPF (9.9) can be written as a standard QCQP

$$\textbf{OPF}: \quad \min_{V\in\mathbb{C}^{N+1}} \; V^{\mathsf{H}} C_0 V \tag{9.16a}$$

$$\text{s.t. } p_j^{\min} \le V^{\mathsf{H}}\Phi_j V \le p_j^{\max}, \qquad j\in\overline{N} \tag{9.16b}$$

$$q_j^{\min} \le V^{\mathsf{H}}\Psi_j V \le q_j^{\max}, \qquad j\in\overline{N} \tag{9.16c}$$

$$v_j^{\min} \le V^{\mathsf{H}} E_j V \le v_j^{\max}, \qquad j\in\overline{N} \tag{9.16d}$$

$$V^{\mathsf{H}}\hat{Y}_{jk} V \le \ell_{jk}^{\max}, \qquad (j,k)\in E \tag{9.16e}$$

$$V^{\mathsf{H}}\hat{Y}_{kj} V \le \ell_{kj}^{\max}, \qquad (j,k)\in E \tag{9.16f}$$

This form will be used to derive a convex relaxation in Chapter 10.2. As mentioned above the OPF formulation here does not require assumption C4.1 that $y^s_{jk} = y^s_{kj}$, and hence can accommodate single-phase transformers that have complex turns ratios. To avoid triviality we will assume unless otherwise specified that OPF (9.16) is feasible.

Instead of (9.16e)(9.16f), line limits are sometimes expressed in terms of line power flows. The next example shows how to express such limits on real and reactive line flows as quadratic constraints. See Exercise 9.2 on how to express limits on apparent powers $|S_{jk}(V)|$, $|S_{kj}(V)|$ as inhomogeneous quadratic constraints.

Example 9.2 (Quadratic line power limit)**.** Use (9.2) to write the line limit

$$S_{jk}^{\min} \le S_{jk}(V) \le S_{jk}^{\max}, \;\; S_{kj}^{\min} \le S_{kj}(V) \le S_{kj}^{\max}, \qquad (j,k)\in E \tag{9.17}$$

as quadratic forms in V.

Solution. We will rewrite the first constraint in (9.17) on $S_{jk}(V)$ as a quadratic constraint; the constraint on $S_{kj}(V)$ can be similarly converted. Using the expression of I_{jk}, $S_{jk}(V)$ in quadratic form is:

$$\begin{aligned} S_{jk}(V) = V_j I_{jk}^{\mathsf{H}} &= \left(e_j^{\mathsf{H}} V\right)\left(y^s_{jk}(e_j - e_k)^{\mathsf{T}} V + y^m_{jk} e_j^{\mathsf{T}} V\right)^{\mathsf{H}} \\ &= e_j^{\mathsf{H}}\left(VV^{\mathsf{H}}\right)\left(\left(\bar{y}^s_{jk} + \bar{y}^m_{jk}\right) e_j - \bar{y}^s_{jk} e_k\right) \\ &= \operatorname{tr}\left(\tilde{Y}_{jk}^{\mathsf{H}}\left(VV^{\mathsf{H}}\right)\right) =: V^{\mathsf{H}}\tilde{Y}_{jk}^{\mathsf{H}} V \end{aligned}$$

where

$$\tilde{Y}_{jk} := e_j\left(\left(y^s_{jk} + y^m_{jk}\right) e_j^{\mathsf{H}} - y^s_{jk} e_k^{\mathsf{H}}\right) \tag{9.18a}$$

or explicitly

$$\left[\tilde{Y}_{jk}\right]_{mn} := \begin{cases} \left(y^s_{jk} + y^m_{jk}\right) & m = n = j \\ \left(-y^s_{jk}\right) & m = j,\; n = k \\ 0 & \text{otherwise} \end{cases}$$

which is symmetric if and only if $y^s_{jk} = y^s_{kj}$. $\tilde{Y}_{jk}$ is not Hermitian and hence $V^{\mathsf{H}}\tilde{Y}_{jk}V$ is a complex number. Define the Hermitian and skewed Hermitian components of $\tilde{Y}_{jk}$:

$$\tilde{\Phi}_{jk} := \frac{1}{2}\left(\tilde{Y}^{\mathsf{H}}_{jk} + \tilde{Y}_{jk}\right) \quad \text{and} \quad \tilde{\Psi}_{jk} := \frac{1}{2\mathbf{i}}\left(\tilde{Y}^{\mathsf{H}}_{jk} - \tilde{Y}_{jk}\right) \tag{9.18b}$$

so that

$$\mathrm{Re}\left(S_{jk}\right) = V^{\mathsf{H}}\tilde{\Phi}_{jk}V \quad \text{and} \quad \mathrm{Im}\left(S_{jk}\right) = V^{\mathsf{H}}\tilde{\Psi}_{jk}V \tag{9.18c}$$

Hence the constraint $S^{\min}_{jk} \le S_{jk}(V) \le S^{\max}_{jk}$ becomes a pair of quadratic constraints:

$$\begin{aligned} \mathrm{Re}\left(S^{\min}_{jk}\right) &\le V^{\mathsf{H}}\tilde{\Phi}_{jk}V \le \mathrm{Re}\left(S^{\max}_{jk}\right) \\ \mathrm{Im}\left(S^{\min}_{jk}\right) &\le V^{\mathsf{H}}\tilde{\Psi}_{jk}V \le \mathrm{Im}\left(S^{\max}_{jk}\right) \end{aligned}$$

□

9.2 Branch flow model: radial networks

DistFlow model.

Since the branch flow model is most useful for radial networks, we first formulate OPF in the DistFlow model that assumes:

- $z^s_{jk} = z^s_{kj}$, or equivalently $y^s_{jk} = y^s_{kj}$, for every line (j,k) (assumption C5.1).
- $y^m_{jk} = y^m_{kj} = 0$ for every line (j,k). This is a reasonable assumption on distribution lines where y^m_{jk} and y^m_{kj} are typically much smaller in magnitude than the series admittance y^s_{jk}.

Consider a single-phase radial network $G = (\overline{N}, E)$ with $N+1$ buses and $M = N$ lines. The assumptions allow us to adopt a directed graph $G = (\overline{N}, E)$ and include branch variables in only one direction. We denote a line in E from bus j to bus k either by $(j,k) \in E$ or $j \to k$. It is characterized by its series impedance $z_{jk} := z^s_{jk}$ (we sometimes omit the superscript when there is no danger of confusion). Without loss of generality we take bus 0 as the root of the tree.

The device models are the same as those for the bus injection model described in Chapter 9.1.1. OPF in the branch flow model differs only in the terminal variables and power flow equations that relate them. We use the DistFlow model (5.8) with down orientation (all lines point away from bus 0), reproduced here:

$$\sum_{k:j\to k} S_{jk} = S_{ij} - z^s_{ij}\ell_{ij} + s_j, \qquad j \in \overline{N} \tag{9.19a}$$

$$v_j - v_k = 2\,\mathrm{Re}\left(\bar{z}^s_{jk}S_{jk}\right) - |z^s_{jk}|^2\ell_{jk}, \qquad j \to k \in E \tag{9.19b}$$

$$v_j\ell_{jk} = |S_{jk}|^2, \qquad j \to k \in E \tag{9.19c}$$

where, in (9.19a), bus $i := i(j)$ denotes the unique adjacent node of j on the path from node 0 to node j, with the understanding that when $j = 0$ then $S_{i0} := 0$ and $\ell_{i0} := 0$. The injection, voltage and line limits are:

$$s_j^{\min} \le s_j \le s_j^{\max}, \quad v_j^{\min} \le v_j \le v_j^{\max}, \quad \ell_{jk} \le \ell_{jk}^{\max}, \quad j \in \overline{N}, \ (j,k) \in E \tag{9.19d}$$

Denote by $(s,v) := (s_j, v_j, j \in \overline{N}) \in \mathbb{R}^{3(N+1)}$ the bus injections and squared voltage magnitudes, and by $(\ell, S) := (\ell_{jk}, S_{jk}, j \to k \in E) \in \mathbb{R}^{3M}$ the squared line current magnitudes and line powers. The vector v includes v_0 and s includes s_0. Let $x := (s,v,\ell,S)$ in $\mathbb{R}^{3(2N+1)}$ since G is a tree. Let the cost function in the branch flow model be $C(x)$. Let the feasible set be

$$\mathbb{X}_{\text{df}} := \{x := (s,v,\ell,S) \in \mathbb{R}^{6N+3} \mid x \text{ satisfies (9.19)}\} \tag{9.20a}$$

Then the optimal power flow problem in the branch flow model is:
OPF:

$$\min_x \ C(x) \ \text{ subject to } \ x \in \mathbb{X}_{\text{df}} \tag{9.20b}$$

To avoid triviality we will assume unless otherwise specified that OPF (9.20) is feasible. We assume the cost functions $C(x)$ here and $C(V)$ in the single-phase OPF problem (9.9) or (9.16) in the bus injection model represent the same function but in terms of different variables. Since $\mathbb{X}_{\text{df}} \equiv \mathbb{V}$ by Theorem 5.2, the single-phase OPF problem (9.20) in the branch flow model is equivalent to (9.9) or (9.16) in the bus injection model. (See the proof of Theorem 11.2 in Chapter 11.1.2 for an explicit construction of a bijection between $\mathbb{X}_{\text{df}}$ and a set equivalent to the feasible set $\mathbb{V}$ of (9.9).)

Remark 9.5 (Current sources and impedances)**.** The model (9.19) includes only voltage and power sources whose controllable variables are v_j and s_j respectively. A current source will introduce its current $I_j \in \mathbb{C}$ as an additional variable and an equality constraint $|s_j|^2 = v_j |I_j|^2$ that relate I_j to (s_j, v_j). An impedance z_j will introduce an equality constraint $s_j = -v_j / z_j^{\mathsf{H}}$ on (s_j, v_j). If z_j is controllable, e.g., representing a switched capacitor, then z_j is an additional variable. For simplicity we restrict ourselves to voltage and power sources only. (See Chapter 9.1.2 for more discussions.) □

General radial network.

The feasible set $\mathbb{X}_{\text{df}}$ is based on the DistFlow equations (9.19a)–(9.19c) that assume $z_{jk}^s = z_{kj}^s$ and $y_{jk}^m = y_{kj}^m = 0$. OPF can also be formulated without these assumptions, based on the branch flow model (5.1) that includes branch variables $\ell := \big(\ell_{jk}, \ell_{kj}, (j,k) \in E\big)$, $S := \big(S_{jk}, S_{kj}, (j,k) \in E\big)$ in both directions, reproduced

here:

$$s_j = \sum_{k:j\sim k} S_{jk}, \qquad j \in \overline{N} \tag{9.21a}$$

$$|\alpha_{jk}|^2 v_j - v_k = 2\mathrm{Re}\left(\alpha_{jk}\bar{z}^s_{jk}S_{jk}\right) - |z^s_{jk}|^2\ell_{jk}, \qquad (j,k)\in E \tag{9.21b}$$

$$|\alpha_{kj}|^2 v_k - v_j = 2\mathrm{Re}\left(\alpha_{kj}\bar{z}^s_{kj}S_{kj}\right) - |z^s_{kj}|^2\ell_{kj}, \qquad (j,k)\in E \tag{9.21c}$$

$$\left|S_{jk}\right|^2 = v_j\ell_{jk}, \qquad \left|S_{kj}\right|^2 = v_k\ell_{kj}, \qquad (j,k)\in E \tag{9.21d}$$

$$\bar{\alpha}_{jk}v_j - \bar{z}^s_{jk}S_{jk} = \left(\bar{\alpha}_{kj}v_k - \bar{z}^s_{kj}S_{kj}\right)^{\mathsf{H}}, \qquad (j,k)\in E \tag{9.21e}$$

where

$$\alpha_{jk} := 1 + z^s_{jk}\,y^m_{jk}, \qquad \alpha_{kj} := 1 + z^s_{kj}\,y^m_{kj}$$

The operation limits are the same as (9.19d) but include line limits in both directions:

$$s_j^{\min} \le s_j \le s_j^{\max}, \quad v_j^{\min} \le v_j \le v_j^{\max}, \quad \ell_{jk} \le \ell_{jk}^{\max}, \quad \ell_{kj} \le \ell_{kj}^{\max}, \quad j \in \overline{N}, \ (j,k)\in E \tag{9.21f}$$

The feasible set is

$$\mathbb{X}_{\text{tree}} := \{x : (s,v,\ell,S) \in \mathbb{R}^{9N+3} \mid x \text{ satisfies (9.21)}\} \tag{9.22a}$$

and the OPF problem is:

OPF:

$$\min_x \ C(x) \quad \text{subject to} \quad x \in \mathbb{X}_{\text{tree}} \tag{9.22b}$$

Since $\mathbb{X}_{\text{tree}} \equiv \mathbb{V}$ by Theorem 5.2, the single-phase OPF problem (9.22) for a general radial network is equivalent to (9.9) or (9.16) in the bus injection model, provided the cost functions $C(x)$ here and $C(V)$ in the bus injection model are the same.

9.3 NP-hardness

Since the feasible set of OPF is generally nonconvex (see e.g. (9.16)), OPF is a nonconvex problem. Moreover OPF has been shown to be NP-hard in [68, 69, 70, 71, 72, 73, 74]. We present the result of [70] that shows that even determining the feasibility of an OPF on a tree network is NP-hard. As hardness results describe worst-case complexity this suggests that there are OPF instances that are hard to scale.

9.3.1 OPF feasibility on a tree network

Consider a tree network represented by a graph $(\overline{N}, E)$ with $N+1$ buses and $M = N$ lines described by the polar-form power flow equations (4.27):

$$p_j = \sum_{k:k\sim j} \left(g^s_{jk} + g^m_{jk}\right)|V_j|^2 - \sum_{k:k\sim j} |V_j||V_k|\left(g^s_{jk}\cos\theta_{jk} + b^s_{jk}\sin\theta_{jk}\right), \qquad j \in \overline{N}$$

$$q_j = -\sum_{k:k\sim j} \left(b^s_{jk} + b^m_{jk}\right)|V_j|^2 - \sum_{k:k\sim j} |V_j||V_k|\left(g^s_{jk}\sin\theta_{jk} - b^s_{jk}\cos\theta_{jk}\right), \quad j \in \overline{N}$$

We make the following assumptions:

- fixed voltage magnitudes $|V_j| := 1$ pu for all $j \in \overline{N}$;
- $y^s_{jk} = y^s_{kj}$ and $y^m_{jk} = y^m_{kj} = 0$;
- $y^s_{jk} = g + \mathbf{i}b$ for all $(j,k) \in E$ with $g \geq 0$, $b \leq 0$.

Assume also that each bus either has a fixed and given injection (p_j, q_j) or a dispatchable generation (p_j, q_j) with the requirement $p_j \geq 0$ (no constraint on q_j). Let $N_L \subset \overline{N}$ denote the set of fixed injections and $N_G \subset \overline{N}$ the set of generations. We are to determine if there are generations $(p_j, q_j, j \in N_G)$ to balance the given injections $(p_j, q_j, j \in N_L)$ subject to the inequality constraints that $p_j \geq 0$ for $j \in N_G$ and a common line limit of the form:

$$|\theta_j - \theta_k| \leq \overline{\theta}, \qquad (j,k) \in E$$

for a given $\overline{\theta} \in (0, \pi/2]$. Exercise 9.8 shows that this constraint is equivalent to a limit on the squared apparent line flow $P^2_{jk} + Q^2_{jk}$ over $\theta \in (0, \overline{\theta})$. Hence the OPF feasibility problem is to find nonnegative real power injections $(p_j, j \in N_G) \geq 0$ at generation buses, voltage angles $(\theta_j, j \in \overline{N})$ at all buses, and line flows $(P_{jk}, Q_{jk}, (j,k) \in E)$ that satisfy the following constraints

OPF feasibility:

$$p_j = \sum_{k:j\sim k} P_{jk}, \qquad q_j = \sum_{k:j\sim k} Q_{jk}, \qquad j \in N_L \tag{9.23a}$$

$$p_j \geq 0, \qquad j \in N_G \tag{9.23b}$$

$$P_{jk} = g_{jk}(1 - \cos\theta_{jk}) - b_{jk}\sin\theta_{jk}, \qquad (j,k) \in E \tag{9.23c}$$

$$Q_{jk} = -b_{jk}(1 - \cos\theta_{jk}) - g_{jk}\sin\theta_{jk}, \qquad (j,k) \in E \tag{9.23d}$$

$$|\theta_j - \theta_k| \leq \overline{\theta}, \qquad (j,k) \in E \tag{9.23e}$$

These constraints define the feasible set of an OPF on the tree network. An instance of the OPF feasibility problem is specified by $(N_G \cup N_L, E)$, $(g_{jk}, b_{jk}, (j,k) \in E)$, $\overline{\theta} \in (0, \pi/2]$, and $(p_j, q_j, j \in N_L)$.

9.3.2 OPF is NP-hard

We often say a function is computable in polynomial time (tractable) or a problem is NP hard (intractable). We first describe these notions more precisely by summarizing basic concepts of complexity theory (see e.g. [75] for more details). We then state the theorem that OPF feasibility problem is NP-hard.

P and NP.

NP-hardness is formally defined first for language problems. Let Σ be a finite set of symbols called an *alphabet* and Σ^* denote the set of all finite strings of symbols in Σ. A *language* L over Σ is any subset of Σ^*. A deterministic Turing machine (DTM) is a model for computation that takes an input σ from Σ^*, performs computation (e.g., read, write, state transition), and then either halts in one of a set of designated states or does not halt. We will focus on classes of languages $L \subseteq \Sigma^*$ for which a DTM always halts in one of two states, "yes" or "no" (these are called decidable decision problems). Given a DTM M, the time complexity function $c_M : \mathbb{N}_+ \to \mathbb{N}_+$ of M ($\mathbb{N}_+$ is the set of positive integers) is:

$$c_M(n) := \max\{m : \exists \sigma \in \Sigma^* \text{ with } |\sigma| = n \text{ s.t. } M \text{ takes } m \text{ steps to halt on } \sigma\}$$

A DTM M is called a *polynomial time DTM* if there exists a polynomial p such that $c_M(n) \leq p(n)$ for all $n \in \mathbb{N}_+$. The set

$$L_M := \{\sigma \in \Sigma^* : M \text{ halts on } \sigma \text{ in "yes" state }\} \tag{9.24}$$

is called the *language recognized by* M. The *class P* of languages is

$$\mathrm{P} := \{L \subseteq \Sigma^* : \exists \text{ polynomial time DTM } M \text{ for which } L = L_M\}$$

Informally the class P consists of all languages over Σ that are recognizable by a DTM in time upper bounded by a polynomial in the length of the input string.

While P is meant to capture the "solvability" of a problem, NP is meant to capture the "verifiability" of a problem, i.e., given a guess, verify if it is a solution. For many problems, it is much easier to verify if a given candidate is a solution than computing a solution. For instance, it is difficult (NP-complete) to find a cycle in an arbitrary graph that visits every node exactly once, but much easier to verify if a candidate path is a solution. This is called the Hamiltonian circuit/cycle problem and is a special case of traveling salesman problem where the distances between adjacent cities are 1. Formally, given a nondeterministic Turing machine (NDTM) M, the time complexity function of M is:

$$c_M(n) := \max\{m : \exists \sigma \in \Sigma^* \text{ with } |\sigma| = n \text{ s.t. } M \text{ takes } m \text{ steps to halt on } \sigma \textit{ in "yes" state}\}$$

If M does not halt in "yes" state for any σ with $|\sigma| = n$, then $c_M(n) := 1$. (For decidable problems, which are what we focus on, M will halt in "no" state on σ.) Then M is called a *polynomial time NDTM* if there exists a polynomial p such that $c_M(n) \leq p(n)$

for all $n \in \mathbb{N}_+$. The language recognized by a NDTM M is L_M as defined in (9.24) except for a NDTM M. Then

$$\text{NP} := \{L \subseteq \Sigma^* : \exists \text{ polynomial time NDTM } M \text{ for which } L = L_M\}$$

Informally the class NP consists of all languages over Σ that are recognizable by a NDTM (or equivalently, verifiable by a DTM) in time upper bounded by a polynomial in the length of the input string. NP contains P as a subclass.

A function $f : \Sigma_1^* \to \Sigma_2^*$ is a language $L_f := \{(\sigma, f(\sigma)) : \sigma \in \Sigma_1^*\} \subseteq \Sigma_1^* \times \Sigma_2^*$. We say a DTM M *computes* f if $L_M = L_f$. Let $L_1 \subseteq \Sigma_1^*$ and $L_2 \subseteq \Sigma_2^*$ be two languages. A *polynomial transformation* or *polynomial reduction* from L_1 to L_2 is a function $f : \Sigma_1^* \to \Sigma_2^*$ which can be computed by a polynomial time DTM such that, for all $\sigma \in \Sigma_1$, $\sigma \in L_1$ if and only if $f(\sigma) \in L_2$. Note the asymmetry between L_1 and L_2. A language L is *NP-hard* if for every $L' \in$ NP there exists a polynomial reduction from L' to L. It is *NP complete* if L is NP-hard and $L \in$ NP. NP-complete languages are in a sense the "hardest" languages in NP.

A decision problem is a problem whose solution is either "yes" or "no". Such a problem is defined by a (possibly countably infinite) set Π of finite *instances*, usually described in terms of sets, graphs, functions, real numbers, etc. These instances are finite in the sense that each instance in Π can be represented by a finite number of symbols. Even though the specification of an instance can involve real numbers such as $\sqrt{7/3}, \cos(\pi/3)$, they are typically described symbolically in terms of integers. We consider decision problems Π that can be "encoded" into language problems defined over some alphabet Σ. Informally, an encoding is a mapping from Π to Σ^*. For any instance $y \in \Pi$ let $\sigma(y) \in \Sigma^*$ denote the result of the mapping, i.e., the encoding of the instance y. We sometimes refer to y or $\sigma(y)$ interchangeably as a decision problem instance when the underlying encoding is understood.

Let $Y \subseteq \Pi$ be the subset of instances of the decision problem Π whose solutions are "yes". We will refer to Y either as a set of problem instances (from Π) or simply as a problem by itself. Let $L_Y := \{\sigma(y) : y \in Y\} \subseteq \Sigma^*$ be the language defined by the instances in Y, i.e., the solution of an instance y is "yes" if and only if its encoding $\sigma(y) \in L_Y$. Hardness properties of the problem (instances) Y are then defined in terms of the hardness properties of its encoding L_Y. For example the problem (instances) Y is said to be in P if $L_Y \in$ P and it is said to be NP-complete if L_Y is NP-complete. The OPF feasibility problem (9.23) is such a decision problem.

Computation problems such as solving a system of equations or optimization problems can likewise be encoded into a language L for which hardness properties can be formally defined. The hardness properties of L then endow a computation or optimization problem with the corresponding hardness properties. A large number of prototypical problems have been proved to be NP-complete. No polynomial time algorithms are known for solving these problems. Moreover a polynomial time algorithm

for solving one of these problems will lead to polynomial time algorithms for all of them. It is in this sense that NP-complete problems are the "hardest" problems.

Hardness result.

We can now state the hardness result.

Theorem 9.1 (OPF NP-hardness [70])**.** The OPF feasibility problem (9.23) is NP-hard.

Remark 9.6. 1 The OPF feasibility problem is not proved to be in the class of NP (and hence NP-complete) because solutions of (9.23) can be irrational.

2 Consider a decision problem defined by the set Π of all its instances. Each instance σ (or more precisely, the encoding $\sigma(y)$ of each $y \in \Pi$) has $\text{len}(\sigma)$, which is a measure of the size of the specification of σ, and $\max(\sigma)$, which is a measure of the magnitude of numerical parameters of σ ($(N_G, N_L, E, g, b, \overline{\theta}, p_j, q_j, j \in N_L)$ in our case). Let p be a polynomial over integers and $Y_p \subseteq \Pi$ be the set of all problem instances with $\max(\sigma) \leq p(\text{len}(\sigma))$, i.e., Y_p is the subset of Π instances for which all numerical parameters are bounded by the single polynomial p in the size of the input instance. The problem Y_p is called *strongly NP-hard* if there exists a polynomial p such that Y_p is NP-hard. It is called *strongly NP-complete* if Y_p is strongly NP-hard and $Y_p \in$ NP [76].

We will prove Theorem 9.1 below by reducing the NP-complete subset sum problem to our OPF feasibility problem. The theorem does not imply that (9.23) is strongly NP-hard because the subset sum problem is NP-complete but not strongly NP-complete. See [73] for a proof that determining OPF feasibility is strongly NP-hard by a polynomial reduction of the strongly NP-complete one-in-three 3SAT problem.

3 The more restrictive the class of OPF instances to which all instances of the subset sum problem can be reduced, the stronger the hardness result because computation complexity is about the performance on worst-case instances. For example the constraints in (9.23) apply to networks with meshed topology, but the NP-hardness proof reduces any instance of the subset sum problem to OPF feasibility instances that use only star networks. Theorem 9.1 says that even the OPF feasibility problem Π_1 in which all instances are restricted to star networks is NP-hard. The larger class of OPF feasibility problem $\Pi_2 \supset \Pi_1$ in which instances may be meshed networks is therefore also NP-hard. It suggests that OPF as an optimization problem is NP-hard.

4 Theorem 9.1 does not mean that all instances of OPF are hard to solve. Indeed we study in Chapters 10 and 11 subclasses of OPF on tree networks that are polynomial time solvable. These subclasses fall outside the subclass defined by the OPF feasibility problem (9.23) ((9.23) does not satisfy the sufficient conditions in Chapter 10 and 11 for exact convex relaxations). We will also study in Chapter 9.4 another class of OPF that can be solve efficiently.

5 Besides nonconvexity another source of hardness is involvement of discrete variables in OPF such as in unit commitment. The hardness of approximation and

approximation ratios of such problems are studied in [72, 74]. See also [77, Chapter 5] for a collection of hardness and approximation results. □

9.3.3 Proof of Theorem 9.1

To show that the OPF feasibility problem is NP-hard we will reduce an arbitrary instance of the NP-complete subset sum problem to an instance of (9.23).

Subset sum problem (A,σ):
Problem instance: a set A of positive integers and a positive integer σ.
Decision: whether there is a subset $A_0 \subseteq A$ such that $\sum_{a\in A_0} a = \sigma$.

OPF feasibility $(N_G \cup N_L, E)$, $(g_{jk}, b_{jk}, (j,k) \in E)$, $\overline{\theta} \in (0,\pi/2]$, *and* $(p_j, q_j, j \in N_L)$:
Problem instance: a graph (star) $(N_G \cup N_L, E)$, $|N_L|$ rational numbers $(p_j, q_j, j \in N_L)$, $|E|$ rational numbers $(g_{jk}, b_{jk}, (j,k) \in E)$, and rational number $\overline{\theta} \in (0,\pi/2]$ that define an instance of the OPF feasibility problem (9.23).
Decision: whether there exist nonnegative real power injections $(p_j, j \in N_G) \geq 0$ at generation buses, voltage angles $(\theta_j, j \in \overline{N})$ at all buses, and line flows $(P_{jk}, Q_{jk}, (j,k) \in E)$ that satisfy (9.23).

An instance of the subset sum problem specified by (A,σ) is said to be *solvable* if a solution A_0 exists. In the following we will describe a polynomial reduction of an arbitrary instance (A,σ) to an instance of the OPF feasibility problem, and show that (A,σ) is solvable if and only if the corresponding instance of the OPF feasibility problem is feasible. Let

$$\hat{P}(\theta) \;:=\; g(1-\cos\theta) - b\sin\theta, \qquad \hat{Q}(\theta) \;=\; -b(1-\cos\theta) - g\sin\theta \tag{9.25}$$

for some (g,b) to be chosen later. We now prove Theorem 9.1 in three steps.

Step 1: Polynomial reduction. Fix an arbitrary subset sum instance (A,σ). We specify the parameters $(N_G \cup N_L, E)$, $(p_j, q_j, j \in N_L)$, $(g_{jk}, b_{jk}, (j,k) \in E)$ and $\overline{\theta} \in (0,\pi/2]$ that defines an instance of the OPF feasibility problem (9.23). Choose $(g,b,\overline{\theta})$ such that $b < 0 < g$, $\hat{P}(-\overline{\theta}) < 0$ in (9.25), and $\overline{\theta} := (0,\pi/2]$. Construct the following star network $(N_G \cup N_L, E)$ with $|A|$ generator buses connected to a single load bus where

- $N_G := A$, $N_L := \{0\}$ with $p_0 := \sigma\hat{P}(-\overline{\theta})$ and $q_0 := \sigma\hat{Q}(-\overline{\theta})$ at the load bus $j = 0$.
- For all lines $(a,0) \in E$ and all $a \in A$, $g_{a0} := ag$ and $b_{a0} := ab$.

Denote this OPF feasibility problem instance as $T(A,\sigma)$. This reduction is polynomial in the size of (A,σ) since the construction only uses rational numbers and finitely many real numbers constructed from integers $a \in A$, basic arithmetic operations, sin and cos. We next show that (A,σ) is solvable if and only if $T(A,\sigma)$ has a feasible solution $x := (p_j, j \in N_G;\ \theta_j, j \in N_G \cup N_L;\ P_{jk}, Q_{jk}, (j,k) \in E)$ for (9.23).

Step 2: (A,σ) is solvable $\Rightarrow$ $T(A,\sigma)$ is feasible. Let $A_0 \subseteq A$ be a solution of (A,σ). Define x by (recall that the only load bus is $j=0$):

$$\begin{aligned} \theta_0 &:= 0, & \theta_a &:= \overline{\theta}, & \forall a \in A_0 \\ P_{0a} &:= a\hat{P}(-\overline{\theta}), & Q_{0a} &:= a\hat{Q}(-\overline{\theta}), & \forall a \in A_0 \\ p_a := P_{a0} &:= a\hat{P}(\overline{\theta}), & Q_{a0} &:= a\hat{Q}(\overline{\theta}), & \forall a \in A_0 \end{aligned}$$

and for all buses outside the solution set,

$$p_a := \theta_a := P_{0a} := Q_{0a} := P_{a0} := Q_{a0} := 0, \qquad \forall a \in A \setminus A_0$$

We show that x satisfies (9.23). Clearly the line flows (9.23c)(9.23d) and the line limits (9.23e) are satisfied by construction. The injection at each generator bus $a \in A$ is

$$p_a := P_{a0} = a(g(1-\cos\overline{\theta}) - b\sin\overline{\theta}) \geq 0$$

where the inequality follows from a being a positive integer, $b < 0 < g$ and $\overline{\theta} \in (0, \pi/2]$, which is (9.23b). Finally the power balance (9.23a) at the load bus $j=0$ is:

$$\sum_{a\in A} P_{0a} = \hat{P}(-\overline{\theta}) \sum_{a\in A_0} a = \sigma\hat{P}(-\overline{\theta}) = p_0$$

where the second equality follows because A_0 is a solution of (A,σ) and the other equalities are due to construction. Similarly

$$\sum_{a\in A} Q_{0a} = \hat{Q}(-\overline{\theta}) \sum_{a\in A_0} a = \sigma\hat{Q}(-\overline{\theta}) = q_0$$

Hence x is a feasible solution of (9.23).

Step 3: $T(A,\sigma)$ is feasible $\Rightarrow$ (A,σ) is solvable. Let $x := (p_j, j \in N_G;\ \theta_j, j \in N_G \cup N_L;\ P_{jk}, Q_{jk}, (j,k) \in E)$ be a solution of (9.23). Consider the line flow $P_{a0} = a\hat{P}(\theta_{a0}) := a\,(g(1-\cos\theta_{a0}) - b\sin\theta_{a0})$ on each line $(a,0) \in E$. Suppose $\theta_{a0} < 0$. Then it can be shown that $\hat{P}(\theta_{a0}) < 0$ (Exercise 9.9). But this implies $p_a = P_{a0} = a\hat{P}(\theta_{a0}) < 0$, contradicting (9.23b). Therefore $\theta_{a0} \geq 0$ for all $a \in A$.

Let $A_0 := \{a \in A : \theta_{a0} > 0\}$. We now show that $\sum_{a\in A_0} a = \sigma$. From (9.23a) we have at bus 0,

$$\begin{aligned} \sigma\hat{P}(-\overline{\theta}) =: p_0 &= \sum_{a\in A} P_{0a} = \sum_{a\in A} a\hat{P}(\theta_{0a}) \\ \sigma\hat{Q}(-\overline{\theta}) =: q_0 &= \sum_{a\in A} Q_{0a} = \sum_{a\in A} a\hat{Q}(\theta_{0a}) \end{aligned} \tag{9.26}$$

Since $\hat{P}(-\overline{\theta}) < 0$ by construction, we have $\hat{Q}(-\overline{\theta}) > 0$ (Exercise 9.9), and hence we can divide both sides by $\hat{P}(-\overline{\theta})$ and $\hat{Q}(-\overline{\theta})$ to obtain:

$$\sum_{a\in A} a\left(\frac{\hat{P}(\theta_{0a})}{\hat{P}(-\overline{\theta})} - \frac{\hat{Q}(\theta_{0a})}{\hat{Q}(-\overline{\theta})}\right) = 0$$

or

$$\sum_{a\in A} a\left(\hat{Q}(-\overline{\theta})\hat{P}(\theta_{0a}) - \hat{P}(-\overline{\theta})\hat{Q}(\theta_{0a})\right) \;=\; 0$$

It can be shown that $\hat{P}(-\overline{\theta})\hat{Q}(\theta_{0a}) \geq \hat{Q}(-\overline{\theta})\hat{P}(\theta_{0a})$ with equality if and only if $\theta_{a0} \in \{0, \overline{\theta}\}$ (Exercise 9.9). Therefore $\theta_{a0} = 0$ or $\overline{\theta}$ for all $a \in A$. Note that $\hat{P}(0) = 0$ and $\theta_{0a} = -\theta_{a0}$, and hence we have from (9.26)

$$\sigma\hat{P}(-\overline{\theta}) \;=\; \sum_{a\in A_0} a\hat{P}(-\overline{\theta})$$

i.e., A_0 is a solution of (A, σ). □

9.4 Global optimality: Lyapunov-like condition

OPF is NP-hard in theory, but seems easy in practice in that polynomial time algorithms often produce globally optimal solutions. In this subsection we study Lyapunov-like conditions for global optimality. Sufficient conditions for global optimality through semidefinite relaxation are studied in Chapters 10 and 11.

9.4.1 Convex relaxation

Consider

$$\min_x \quad f(x) \qquad \text{s.t.} \qquad x \in X \subseteq \mathbb{R}^n \tag{9.27}$$

and

$$\min_x \quad f(x) \qquad \text{s.t.} \qquad x \in \hat{X} \subseteq \mathbb{R}^n \tag{9.28}$$

where X is a nonempty compact set (not necessarily convex), $\hat{X}$ is an arbitrary compact and convex superset of X, and $f : \mathbb{R}^n \to \mathbb{R}$ is a convex (and hence continuous) function. Hence optimal solutions exist for both (9.27) and (9.28) according to Theorem 8.16. Problem (9.27) is a nonlinear program and generally NP-hard (Exercise 9.10). Problem (9.28) is called a *convex relaxation* of (9.27). Since it is a convex problem it is polynomial time solvable (assuming $\hat{X}$ is efficiently represented). If an optimal solution x^* of (9.28) is feasible for (9.27) then x^* is optimal for (9.27). In Chapters 10 and 11 we study the semidefinite relaxations of OPF where $\hat{X}$ is restricted to be a semidefinite cone or a second-order cone, but in this section we allow any convex relaxation.

The cost function of OPF is typically convex but its feasible set is nonconvex due to nonlinear power flow equations. Most algorithms used for solving OPF are local algorithms such as Newton-Raphson or interior-point methods (studied in Chapter 8.5). First order conditions are available to guarantee that these algorithms converge

to produce a global optimum for convex problems. Since OPF is nonconvex there is usually no guarantee that a local algorithm will converge or will produce a global (or local) optimum when it does. Solving convex relaxations of OPF is also widely studied, and in general, there is no guarantee that relaxations will be exact. Yet there is significant evidence that, in practice, local algorithms and convex relaxations tend to produce globally optimal solutions, e.g., [78].

9.4.2 Conditions for global optimality

We now present conditions from [79, 80] for the nonlinear program (9.27) to simultaneously have exact convex relaxation and no spurious local optima. These conditions help explain the empirical experience that local algorithms and convex relaxations for OPF tend to work well in practice.

Definition 9.1 (Exact relaxation)**.** 1 A point $x^* \in X$ is called a *local optimum* of (9.27) if there exists a $\delta > 0$ such that $f(x^*) \leq f(x)$ for all $x \in X$ with $\|x - x^*\| < \delta$. It is called a *global optimum* or an *optimum* if $f(x^*) \leq f(x)$ for all $x \in X$.

2 If *every* optimal solution x^* of (9.28) is feasible for, and therefore a global optimum of, (9.27) then we say that the convex relaxation (9.28) is *exact* with respect to (9.27).

The optimality conditions rely on, for every infeasible point $x \in \hat{X} \setminus X$, finding a path that takes x back to the feasible set X along which the cost is nonincreasing.

Definition 9.2 (Path)**.** 1 A *path in* $Y \subseteq \mathbb{R}^n$ connecting point a to point b is a continuous function $h : [0,1] \to Y$ such that $h(0) = a$ and $h(1) = b$.

2 An arbitrary set $\{h_i : i \in I\}$ of paths in Y is called

1 *uniformly bounded* if there exists a finite number H such that $\|h_i(t)\|_\infty \leq H$ for all $t \in [0,1]$ and all $i \in I$;

2 *uniformly equicontinuous* if for any $\epsilon > 0$, there exists $\delta > 0$ such that $\|h_i(t_2) - h_i(t_1)\|_\infty < \epsilon$ for all $i \in I$ whenever $|t_2 - t_1| < \delta$.

As an example, if all paths in $\{h_i : i \in I\}$ consist of at most m linear segments for some finite m, then (the arc-length reparametrized version of) $\{h_i : i \in I\}$ is both uniformly bounded and uniformly equicontinuous; see [79].

Definition 9.3 (Lyapunov-like function)**.** A *Lyapunov-like function* associated with (9.27) and (9.28) is a continuous function $V : \hat{X} \to \mathbb{R}_+$ such that $V(x) = 0$ if $x \in X$ and $V(x) > 0$ if $x \in \hat{X} \setminus X$.

We can now state a sufficient condition and a necessary condition for (9.27) to simultaneously have exact convex relaxation and no spurious local optima. The first condition C9.1 says that every infeasible point x can be brought back to the feasible set X with a strictly lower cost along a path on which neither the cost f nor the

Lyapunov function V increases. Condition C9.3(b) requires that the cost decreases sufficiently along the path, not just nonincreasing, in order to eliminate the possibility of pseudo local optimum (see Definition 9.4 below). C9.2 is a regularity condition on the set of paths for all infeasible points. It is needed for the Arzelà-Ascoli Theorem that guarantees that this set of paths has a uniformly convergent subsequence in order to prove that all local optima are global optima.

C9.1: There is a Lyapunov-like function V associated with (9.27) and (9.28) and, for every infeasible point $x \in \hat{X} \setminus X$, there is a path h_x in $\hat{X}$ such that
1 $h_x(0) = x$, $h_x(1) \in X$, and $f(h_x(1)) < f(x)$.
2 Both $f(h_x(t))$ and $V(h_x(t))$ are nonincreasing for $t \in [0,1]$.

C9.2: The set $\{h_x : x \in \hat{X} \setminus X\}$ of paths in C9.1 is uniformly bounded and uniformly equicontinuous.

C9.3: At least one of the following holds:
1 All local optima of (9.27) are isolated, i.e., every local optimum has an open neighborhood that contains no other local optimum.
2 For the set $\{h_x : x \in \hat{X} \setminus X\}$ of paths in C9.1, there exists $\alpha > 0$ such that for all infeasible points $x \in \hat{X} \setminus X$ and all $0 \le s < t \le 1$, we have $f(h_x(s)) - f(h_x(t)) \ge \alpha \|h_x(s) - h_x(t)\|$ for some norm $\|\cdot\|$.

Theorem 9.2 (Sufficiency)**.** Suppose conditions C9.1, C9.2, C9.3 hold. Then

1 The convex relaxation (9.28) is exact with respect to (9.27).
2 Every local optimum of (9.27) is a global optimum.

Moreover if C9.3(a) holds then the optimal point is unique.

A set $Y \subseteq \mathbb{R}^n$ is semianalytic if every $x \in \mathbb{R}^n$ has a neighborhood U such that $Y \cap U$ can be represented as a finite Boolean combination of sets $\{x : g(x) = 0\}$ and $\{x : h(x) < 0\}$ for some analytic functions g, h (i.e., for every x_0, $g(x) = \sum_{n=0}^{\infty} a_n (x - x_0)^n$ for some real coefficients a_n in a neighborhood of x_0, and similarly for h). Engineering problems are often specified in terms of analytic functions and semianalytic sets.

Theorem 9.3 (Necessity)**.** Suppose the feasible set X is semianalytic and the cost function f is analytic. If (9.28) is exact with respect to (9.27) and every local optimum of (9.27) is a global optimum, then there exists Lyapunov-like function V and a family of paths $\{h_x : x \in \hat{X} \setminus X\}$ that satisfy C9.1, C9.2.

Remark 9.7 (Sufficiency)**.** 1 Conditions C9.1 and C9.2 imply that the feasible set X of (9.27) is connected. For OPF however it is possible that the feasible set is disconnected. In that case convex relaxation may not be exact in the strong sense of Definition 9.1 that *all* optimal points of (9.28) are optimal for (9.27). Theorems 9.2 and 9.3 hold however for X restricted to a connected component of the feasible set. We can also consider a weaker notion of exactness that requires at least one global optimum of (9.28) to be feasible and hence optimal for (9.27). See [79,

Theorem 4] for a similar sufficient condition that guarantees weak exactness of (9.28) and no spurious local optimum for (9.27).

2 As we will show in Lemma 9.4 below the exactness of (9.28) with respect to (9.27) is equivalent to the existence of a path h_x for each infeasible point $x \in \hat{X} \setminus X$ that satisfies C9.1. Indeed proofs of exact relaxations in Chapters 10 and 11 can be interpreted as constructing such a path. The existence of a Lyapunov-like function and all other conditions in Theorem 9.2 are needed to prove the global optimality of every local optimum.

3 Consider the dynamical system

$$\dot{x} = f(x(t)), \qquad t \geq 0,\ x(0) = x_0 \tag{9.29}$$

and suppose x^* is an equilibrium point where $f(x^*) = 0$. The equilibrium point x^* is said to be globally asymptotically stable if the trajectory $x(t)$ of (9.29) stays close to x^* whenever the initial point x_0 is close to x^* and $x(t) \to x^*$ for any initial point x_0. The standard Lyapunov stability theory says that x^* is globally asymptotically stable if there exists a continuously differentiable Lyapunov function $V(x)$ such that $V(x) > V(x^*)$ and $\dot{V}(x) < 0$ for all $x \neq x^*$. In this case (9.29) specifies the trajectory (path) that $x(t)$ takes starting from a given x_0 and the Lyapunov function V certifies a stability property of the equilibrium point x^*. There is no general method to construct V except on a case-by-case basis.

In our case, the Lyapunov-like function V in Theorem 9.2 certifies that a local optimum $x^* \in X$ of (9.27) is a global optimum. Since there is no dynamics, there is no requirement on the differentiability of V. We however have to construct a path h_x for every infeasible point $x \in \hat{X} \setminus X$ that takes x back to a feasible point in X with a strictly lower cost. No general methods to construct V or h_x are known (see an example in Chapter 9.4.4). □

Figure 9.1 illustrates the NP hardness of OPF and the set of problem instances that both have exact convex relaxation and no spurious local optimum characterized by Theorems 9.2 and 9.3.

9.4.3 Proof of Theorem 9.2

We next prove the sufficiency condition taken from [80]; see [79] for the proof of Theorem 9.3.

Lemma 9.4. The convex relaxation (9.28) is exact with respect to (9.27) if and only if, for every infeasible point $x \in \hat{X} \setminus X$, there exists a path h_x that satisfies C9.1.

Proof of Lemma 9.4. Suppose (9.28) is exact and let $x^* \in X$ be a global optimum of (9.27), which exists due to Theorem 8.16. Given any infeasible point $x \in \hat{X} \setminus X$, let h_x be the line segment connecting x to x^*. Then h_x is in $\hat{X}$ since $\hat{X}$ is convex. Moreover $f(x) > f(x^*)$ since (9.28) is exact, and hence C9.1 for h_x follows from the convexity

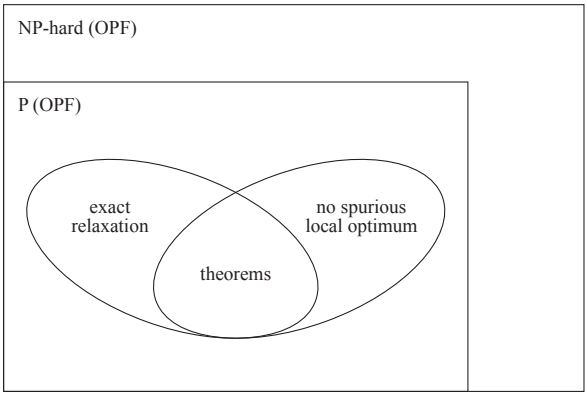


Figure 9.1 Problem instances of OPF. Theorems 9.2 and 9.3 provide a sufficient condition and a necessary condition respectively that characterize the intersection.

of f. Conversely suppose every $x \in \hat{X} \setminus X$ has a path h_x in $\hat{X}$ that satisfies C9.1. If a global optimum x^* of (9.28) is not in X then $f(h_{x^*}(1)) < f(x^*)$, contradicting the optimality of x^*. Hence $x^* \in X$ and is a global optimum of (9.27). □

Lemma 9.4 says that, for exact relaxation, it is sufficient if every infeasible point $x \in \hat{X} \setminus X$ has a path h_x that satisfies just Condition C9.1. For global optimality of local optima of (9.27), we need to differentiate between two types of local optima that are not global optima; see Figure 9.2.

Definition 9.4 (Pseudo local optimum)**.** A local optimum $x^* \in X$ that is not a global optimum is called

1 a *pseudo local optimum* if there is a path $h : [0,1] \to X$ that starts at $h(0) = x^*$ and ends at a point $h(1)$ that is not a local optimum, such that $f(h(t)) \equiv f(x^*)$ for all $t \in [0,1]$.
2 a *genuine local optimum* if it is a local optimum but neither a global optimum nor a pseudo local optimum.

A local optimum x^* is a pseudo local optimum if it can be strictly improved without incurring a higher cost in the process.

Definition 9.5 (Improvable)**.** A point $x \in X$ is called *improvable* in X if there is a path $h_x : [0,1] \to X$ with $h_x(0) = x$ such that

1 $f(h(t))$ is nonincreasing for $t \in [0,1]$;
2 $f(h(1)) < f(x)$ or $h(1)$ is not a local optimum.

A local optimum is a pseudo local optimum if and only if it is improvable in X.

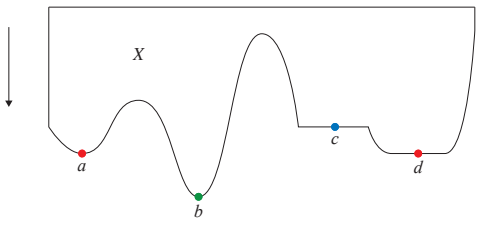


Figure 9.2 Three types of local optima. The cost function decreases in the direction of the arrow. Point b is a global optimum, c a pseudo local optimum, and a, d are genuine local optima.

The next lemma, together with Lemma 9.4, says that conditions C9.1 and C9.2 almost imply Theorem 9.2, except for the possibility of pseudo local optima.

Lemma 9.5. Suppose conditions C9.1 and C9.2 hold. Then every local optimum of (9.27) is either a global optimum or a pseudo local optimum.

Proof of Lemma 9.5. Fix an $x \in X$ that is a local but not global optimum of (9.27). We will show that x is improvable in X and hence a pseudo local optimum.

Let $x^* \neq x$ be a global optimum of (9.27) with $f(x^*) < f(x)$. Let $\ell : [0,1] \to \hat{X}$ be the line segment connecting x to x^*, $\ell(t) := (1-t)x + tx^*$ for $t \in [0,1]$. The convexity of f implies that $f(\ell(t))$ is nonincreasing in t because, for any $0 \leq \tau < t \leq 1$, $\ell(t) = \alpha\ell(\tau) + (1-\alpha)x^*$ for some $\alpha \in [0,1]$ and hence

$$f(\ell(t)) \;=\; \alpha f(\ell(\tau)) + (1-\alpha) f(x^*) \;\leq\; f(\ell(\tau))$$

If $\ell(t) \in X$ for all $t \in [0,1]$, i.e., the line segment is in X, then ℓ defines the path h_x in Definition 9.5 with $f(h_x(1)) < f(x)$. Therefore x is improvable in X and hence a pseudo local optimum.

Suppose then part of ℓ lies in $\hat{X} \setminus X$ and define the first time the line segment ℓ leaves X (see Figure 9.3):

$$t^\dagger \;:=\; \sup_{\tau \in [0,1]} t \qquad \text{s.t.} \qquad \ell(\tau) \in X \quad \forall \tau \leq t$$

Since X is closed, $t^\dagger \in X$. First note that $f(\ell(t))$ is strictly decreasing in t until $f(\ell(s)) = f(x^*)$ for some s and $f(\ell(t)) \equiv f(x^*)$ over $t \in [s,1]$. To see this suppose $f(\ell(\tau)) \equiv f(\ell(\tau_1))$ over any interval $\tau \in [\tau_1, \tau_2]$ with $0 \leq \tau_1 < \tau_2 \leq 1$. Then, since $\ell(\tau_2) = \alpha\ell(\tau_1) + (1-\alpha)x^*$ for some $\alpha \in [0,1)$, we have

$$f(\ell(\tau_1)) \;=\; f(\ell(\tau_2)) \;\leq\; \alpha f(\ell(\tau_1)) + (1-\alpha) f(x^*)$$

implying $f(\ell(\tau)) = f(x^*)$ over $[\tau_1, \tau_2]$ and that $f(\ell(t))$ is strictly decreasing in t until $f(\ell(s)) = f(x^*)$ for some s. Then, since $f(x) > f(x^*)$, the convexity and hence continuity of f imply that $f(\ell(t))$ is strictly decreasing in t until at least $\ell(t)$ is at the

boundary of X. Therefore x can only be on the boundary of X to be a local optimum, i.e., $t^\dagger = 0$ and $x = \ell(t^\dagger) = \ell(0)$.

We will show that $x = \ell(0)$ is improvable in X by constructing the path h in Definition 9.5, i.e., constructing an $h : [0,1] \to X$ with $h(0) = x$ such that $f(h(t))$ is nonincreasing for $t \in [0,1]$ and either $f(h(1)) < f(x)$ or $h(1)$ is not a local optimum. Conditions C9.1 and C9.2 are required for this construction because x is on the boundary of X but the path h must lie entirely in X (as opposed to be in $\hat{X}$ as ℓ is). The notation for the rest of the proof is illustrated in Figure 9.3. The basic idea is as follows. For

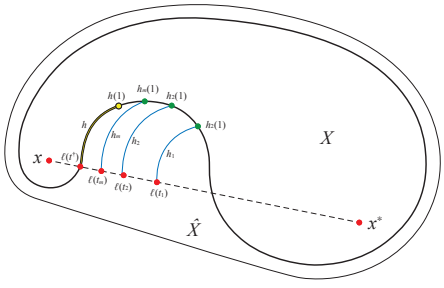


Figure 9.3 Notation for the proof of Lemma 9.5. Point x and $\ell(t^\dagger)$ will be proved to be identical.

a sequence $t_1, t_2, \ldots$ that converges to $t^\dagger = 0$, C9.1 provides a sequence $h_1, h_2, \ldots$ of paths in $\hat{X}$ (not X) that takes the infeasible points $\ell(t_1), \ell(t_2), \ldots$ to some feasible points $h_1(1), h_2(1), \ldots$ in X with strictly lower costs. C9.2 implies that the sequence $h_1, h_2, \ldots$ of paths has a convergent subsequence that converges to a limit function $h : [0,1] \to \hat{X}$ which is then proved to satisfy Definition 9.5. In particular, even though each h_m is a path in $\hat{X}$, their limit h will be the required path in X; see Figure 9.3 (the Lyapunov function V is needed in this step to certify that $h(t) \in X$.). We now make this precise.

By the definition of $t^\dagger$ there exists a decreasing sequence $t_1 > t_2 > \cdots > t^\dagger$ such that $\lim_m t_m = t^\dagger = 0$ and $\ell(t_m) \in \hat{X} \setminus X$ for all m. Since $f(x) > f(x^*)$ and $f(\ell(t))$ is strictly decreasing in t until a certain $s \in (0,1]$, we have for $t_m \le s$

$$f(\ell(t_m)) \;<\; f(\ell(0)) \;=\; f(x) \tag{9.30}$$

Moreover $\lim_m f(\ell(t_m)) = f(x)$ monotonically because $f(\ell(t_m))$ is a nondecreasing sequence in m. C9.1 therefore guarantees a sequence h_m of paths in $\hat{X}$ (not X) with $h_m(0) = \ell(t_m)$ and $h_m(1) \in X$ with a strictly lower cost. Since the sequence $\{h_m : m = 1, 2, \ldots\}$ of paths are uniformly bounded and uniformly equicontinuous, the Arzelà-Ascoli Theorem implies that it has a uniformly convergent subsequence with a limit point $h : [0,1] \to \hat{X}$. Without loss of generality we denote the convergent subsequence by $\{h_m : m = 1, 2, \ldots, \}$.

We now show that h is in X (not just in $\hat{X}$) and satisfies Definition 9.5:

1 $h(t) \in X$ *for* $t \in [0,1]$: Fix any $t \in [0,1]$ and consider the convergent (sub)sequence $\{h_m(t) : m = 1,2,\dots,\}$. The continuity of the Lyapunov-like function V implies

$$V(h(t)) \;=\; V\Big(\lim_m h_m(t)\Big) \;=\; \lim_m V(h_m(t)) \;\le\; \lim_m V(h_m(0))$$

where the inequality follows because $V(h_m(t))$ is nonincreasing in t due to C9.1. Substituting $V(h_m(0)) = V(\ell(t_m))$ from the definition of h_m we have

$$V(h(t)) \;\le\; \lim_m V(\ell(t_m)) \;=\; V\Big(\lim_m \ell(t_m)\Big) \;=\; V(\ell(0)) \;=\; V(x) \;=\; 0$$

since $\lim_m \ell(t_m) = \ell(\lim_m t_m) = x \in X$. Hence $V(h(t)) = 0$ and $h(t) \in X$.

2 $h(0) = x$: We have $h(0) = \lim_m h_m(0) = \lim_m \ell(t_m) = \ell(\lim_m t_m) = \ell(0) = x$.

3 $f(h(t))$ *nonincreasing in* t: This follows from $f(h(t)) = f(\lim_m h_m(t)) = \lim_m f(h_m(t))$ and $f(h_m(t))$ is nonincreasing in t by C9.1.

4 $f(h(1)) < f(x)$ *or* $h(1)$ *is not a local optimum*: Suppose $f(h(1)) = f(x)$. We will show that $h(1)$ cannot be a local optimum. For each m we have

$$f(h_m(1)) \;<\; f(h_m(0)) \;=\; f(\ell(t_m)) \;<\; f(\ell(0)) \;=\; f(x) \;=\; f(h(1))$$

where the first inequality and the first equality follow from C9.1, and the second inequality follows from (9.30). This means that there are infinitely many m such that $f(h_m(1)) < f(h(1))$ and $h_m(1) \to h(1)$. Therefore there is no neighborhood of $h(1)$ in which f attains minimum.

This shows that x, which is a local but not global optimum of (9.27), is improvable in X and hence a pseudo local optimum. □

Finally we show that C9.3 eliminates the possibility of pseudo local optimum. This, together with Lemmas 9.4 and 9.5, proves Theorem 9.2.

Lemma 9.6. Suppose conditions C9.1, C9.2 and C9.3 hold. Then every local optimum of (9.27) is a global optimum.

Proof of Lemma 9.6. If C9.3(a) holds, then clearly a local optimum x cannot be a pseudo local optimum. Lemma 9.5 then implies that x is a global optimum. Moreover if there are multiple local (and hence global) optima x and $\hat{x}$ then, since f is convex, any convex combination of x and $\hat{x}$ is optimal, contradicting that x and $\hat{x}$ are isolated optima. (For the DistFlow model, this uniqueness properties is Theorem 11.1.)

Suppose C9.3(b) holds and x is a local but not a global optimum of (9.27). Following the proof of Lemma 9.5 we have a uniformly convergent (sub)sequence $\{h_m : m = 1,2,\dots\}$ whose limit point is the path $h : [0,1] \to X$ with $h(0) = x$. Since $f(h_m(s)) - f(h_m(t)) \ge \alpha\|h_m(s) - h_m(t)\|$ for any $s < t$ by C9.3(b), taking limit as $m \to \infty$ we have,

$$f(h(s)) - f(h(t)) \;\ge\; \alpha\|h(s) - h(t)\| \;>\; 0 \quad \text{whenever } h(s) \neq h(t) \tag{9.31}$$

Let $s_0 := \inf\{s \in (0,1] : h(s) \neq x\}$. Then $h(s_0) = x$ since h is continuous. The proof of

Theorem 9.2 shows that $f(h(1)) < f(x)$ or $h(1)$ is not a local optimum. This means that $h(1) \neq x$, and hence $0 \leq s_0 < 1$. We claim that $h(s_0+\epsilon) \neq x$ for any $\epsilon \in (0, 1-s_0]$, because

$$f(h(s_0+\epsilon)) \;<\; f(h(s_0)) \;=\; f(x)$$

where the first inequality follows from substituting $s := s_0$ and $t := s_0+\epsilon$ into (9.31). Therefore $f(x) > f(h(t))$ for all $t \in (s_0, 1]$, contradicting the local optimality of x. □

9.4.4 Application to OPF on radial network

Consider the single-phase OPF (9.20) formulated in Chapter 9.2 on a radial network $G = (\overline{N}, E)$ with $N+1$ buses and $M = N$ lines modeled by the DistFlow equation (9.19), reproduced here (all lines pointing away from bus 0):

$$\sum_{k:j\to k} S_{jk} \;=\; S_{ij} - z_{ij}^s \ell_{ij} + s_j, \qquad j \in \overline{N} \tag{9.32a}$$

$$v_j - v_k \;=\; 2\,\mathrm{Re}\left(\bar{z}_{jk}^s S_{jk}\right) - |z_{jk}^s|^2 \ell_{jk}, \qquad j \to k \in E \tag{9.32b}$$

$$v_j \ell_{jk} \;=\; |S_{jk}|^2, \qquad j \to k \in E \tag{9.32c}$$

and operational constraints:

$$s_j^{\min} \;\leq\; s_j \;\leq\; s_j^{\max}, \quad v_j^{\min} \;\leq\; v_j \;\leq\; v_j^{\max}, \quad \ell_{jk} \;\leq\; \ell_{jk}^{\max}, \quad j \in \overline{N},\ (j,k) \in E \tag{9.32d}$$

Denote by $(s, v) := (s_j, v_j, j \in \overline{N}) \in \mathbb{R}^{3(N+1)}$ the bus injections and squared voltage magnitudes, and by $(\ell, S) := (\ell_{jk}, S_{jk}, j \to k \in E) \in \mathbb{R}^{3M}$ the squared line current magnitudes and line powers. The vector v includes v_0 and s includes s_0. Let $x := (s, v, \ell, S)$ in $\mathbb{R}^{3(2N+1)}$ since G is a tree. Let

$$X \;:=\; \{x := (s, v, \ell, S) \in \mathbb{R}^{3(2N+1)} \mid x \text{ satisfies (9.32)}\} \tag{9.33a}$$

Let the cost function be a real-valued function $f(x)$. Then OPF formulated in (9.20) and reproduced here is:

$$\min_x \; f(x) \qquad \text{s.t.} \qquad x \in X \tag{9.33b}$$

The feasible set X is nonconvex because of the nonlinear constraint (9.32c). Relax it to a second-order constraint (studied in Chapter 8.2.1):

$$v_j \ell_{jk} \;\geq\; |S_{jk}|^2, \qquad j \to k \in E \tag{9.34}$$

Consider the relaxed convex feasible set

$$\hat{X} \;:=\; \{x \in \mathbb{R}^{3(2N+1)} \mid x \text{ satisfies (9.32a)(9.32b), (9.34), (9.32d)}\ \} \tag{9.35a}$$

and the convex relaxation of (9.33)

$$\min_x \; f(x) \qquad \text{s.t.} \qquad x \in \hat{X} \tag{9.35b}$$

We assume the problem parameters are such that the following condition is satisfied:

> C9.4: The feasible set X is nonempty and compact, the convex feasible set $\hat{X}$ is compact, and the real-valued cost function $f(x)$ is convex and continuous.

Then (9.33) and (9.35) are an example of (9.27) and (9.28) to which Theorem 9.2 applies.

To construct a Lyapunov-like function V and paths h_x for every infeasible point $x \in \hat{X} \setminus X$, we need additional assumptions:

> C9.5: The cost function $f(x) = f(p,q,v,\ell)$ is independent of line flows $S = (P,Q)$ and continuously differentiable in (p,q,ℓ) with nonnegative $\nabla_p f(x) \geq 0$ and $\nabla_q f(x) \geq 0$ for all $x \in \hat{X}$. Moreover there exists $c > 0$ such that $\frac{\partial f}{\partial \ell_l}(x) \geq c$ for all $l \in E$ and all $x \in \hat{X}$.
> C9.6: For each $j \in \overline{N}$, the injection limit $s_j^{\min} = -\infty - \mathbf{i}\infty$.
> C9.7: For each $j \to k \in E$, $z_{jk} =: (r_{jk}, x_{jk}) > 0$ and the line limit satisfies $|z_{jk}|^2 \ell^{\max} \leq v_j^{\min}$.

C9.5 implies that C is strictly increasing in each component of ℓ_j. Moreover, given any $x := (p,q,v,\ell,S) \in \hat{X}$, any nonnegative $(\delta p, \delta q, 0, \delta\ell) \geq 0$, and any scalar $t \geq 0$ we have (Exercise 9.12)

$$f((p,q,v,\ell) + t(\delta p, \delta q, 0, \delta\ell)) - f(p,q,v,\ell) \;\geq\; ct \sum_{(j,k)\in E} \delta\ell_{jk} \;=\; ct\|\delta\ell\|_1 \tag{9.36}$$

where $\|y\|_1 := \sum_j |y_j|$ is the l_1 norm. This property will be used in the proof below. C9.6 means that demands are large enough not to pose a constraint. C9.7 is realistic because typically $V_j = (1+\epsilon_j)e^{\mathbf{i}\theta_j}$ pu where $\epsilon_j \in [-0.1, 0.1]$ and the angle differences $\theta_{jk} := \theta_j - \theta_k$ are typically small in magnitude. Then the maximum value of $|V_j - V_k|^2 = \left|(1+\epsilon_j)e^{\mathbf{i}\theta_{jk}} - (1+\epsilon_k)\right|$, which is $|z_{jk}|^2\ell^{\max}$, should be much smaller than $v_j^{\min} \approx 1$ pu.

Theorem 11.3 of Chapter 11.2 shows that C9.5 and C9.6 imply that the SOCP relaxation (9.35) is exact with respect to (9.33). We now show that conditions C9.4–C9.7 guarantee that every local optimum of (9.33) is a global optimum.

Theorem 9.7 (Global optimality of (9.33))**.** Suppose C9.4–C9.7 holds for OPF (9.33) on radial networks. Then every local optimum of (9.33) is a global optimum.

Proof We will construct a Lyapunov-like function V and a path h_x in $\hat{X}$ for each infeasible point $\hat{x} \in \hat{X} \setminus X$ that, for OPF (9.33), satisfy C9.1–C9.3. The theorem then follows from Theorem 9.2.

Let

$$V(x) \;:=\; \sum_{j\to k\in E} \left(v_j \ell_{jk} - |S_{jk}|^2\right) \tag{9.37}$$

Clearly $V(x) \geq 0$ for all $x \in \hat{X}$ with equality if and only if $x \in X$, and hence $V(x)$ is a Lyapunov-like function.

Fix an $x \in \hat{X} \setminus X$. To construct a path h_x in $\hat{X}$, let $M := \{(j,k) \in E : v_j \ell_{jk} > |S_{jk}|^2\}$ be the set of lines where the quadratic equality is violated. For each $(j,k) \in M$, let ϕ_{jk} be the quadratic function:

$$\phi_{jk}(a) \;:=\; \frac{|z_{jk}|^2}{4}\, a^2 + \left(v_j - \mathrm{Re}\left(\bar{z}_{jk} S_{jk}\right)\right) a + \left(|S_{jk}|^2 - v_j \ell_{jk}\right) \tag{9.38}$$

Since $\phi_{jk}(0) < 0$, ϕ_{jk} has a unique positive root. Define Δ_{jk} to be this positive root if $(j,k) \in M$ and 0 otherwise. Furthermore for $(j,k) \in M$,

$$v_j - \mathrm{Re}(\bar{z}_{jk} S_{jk}) \;\geq\; v_j - |z_{jk}||S_{jk}| \;>\; v_j - |z_{jk}|\sqrt{v_j \ell_{jk}} \;\geq\; v_j - \sqrt{v_j \cdot |z_{jk}|^2 \ell_{jk}^{\max}} \geq v_j - \sqrt{v_j \cdot v_j}$$

where the second inequality follows from $(j,k) \in M$, and the last inequality follows from C9.7. This implies that the quadratic function $\phi_{jk}(a)$ is negative and strictly increasing over $[0, \Delta_{jk}]$. Consider the path $h_x(t) := \left(\tilde{s}(t), \tilde{v}(t), \tilde{\ell}(t), \tilde{S}(t)\right)$ for $t \in [0,1]$ where

$$\tilde{s}_j(t) \;:=\; s_j - \frac{t}{2} \sum_{i:i\to j} z_{ij} \Delta_{ij} - \frac{t}{2} \sum_{k:j\to k} z_{jk} \Delta_{jk}, \qquad j \in \overline{N} \tag{9.39a}$$

$$\tilde{v}_j(t) \;:=\; v_j, \qquad j \in \overline{N} \tag{9.39b}$$

$$\tilde{\ell}_{jk}(t) \;:=\; \ell_{jk} - t\Delta_{jk}, \qquad j \to k \in E \tag{9.39c}$$

$$\tilde{S}_{jk}(t) \;:=\; S_{jk} - \frac{t}{2} z_{jk} \Delta_{jk}, \qquad j \to k \in E \tag{9.39d}$$

Therefore $h_x(t) := x - tA\Delta(x)$ where the vector $\Delta(x) := (\Delta_{jk}, (j,k) \in E)$ depends on x through the quadratic function $\phi_{jk}(a)$, A is the following $3(2N+1) \times N$ matrix

$$A \;:=\; \begin{bmatrix} \frac{1}{2}|C|R \\ \frac{1}{2}|C|X \\ 0 \\ \mathbb{I}_N \\ \frac{1}{2}R \\ \frac{1}{2}X \end{bmatrix} \qquad \text{with} \qquad R := \mathrm{diag}(r_{jk}, j \to k \in E), \quad X := \mathrm{diag}(x_{jk}, j \to k \in E) \tag{9.40}$$

and $z_{jk} =: (r_{jk}, x_{jk})$. Here $|C|$ is obtained from the node-by-line incidence matrix C by replacing -1 by 1, and 0 and $\mathbb{I}_N$ denote the zero and identity matrices of appropriate sizes. Since $z_{jk} > 0$ (C9.7) and $\Delta_{jk} \geq 0$ by construction, each entry of the vector $A\Delta(x)$ is nonnegative and hence, for OPF (9.33), we have

$$\mathbf{1}^{\mathsf{T}} A\Delta(x) \;=\; \sum_k \left[A\Delta(x)\right]_k \;=\; \|A\Delta(x)\|_1, \qquad x \in \hat{X} \tag{9.41}$$

where $\|x\|_1 := \sum_k |x_k|$ is the l_1 norm. This is a property needed to establish C9.3 below.

We now show that V in (9.37) and $\{h_x : x \in \hat{X} \setminus X\}$ in (9.39) satisfy C9.1–C9.3.

1 Clearly $h_x(0) = x$ in $\hat{X} \setminus X$. It can be shown that $h_x(t) \in \hat{X}$ for all $t \in [0,1]$ and $h_x(1) \in X$ (Exercise 9.12). It suffices to show that both $f(h_x(t))$ and $V(h_x(t))$ are strictly decreasing in t on $[0,1]$ for C9.1 to be satisfied. Since f is strictly increasing in ℓ (C9.5) and $\Delta_{jk} > 0$ for $(j,k) \in M$, $\tilde{\ell} - \ell$ is nonnegative and nonzero from (9.39c) for $t > 0$. Hence $f(h_x(t)) = f(x - tA\Delta(x))$ is strictly decreasing in t on $[0,1]$. For $V(h_x(t))$ we have from (9.37) and (9.39)

$$V(h_x(t)) \;:=\; \sum_{(j,k)\in E} \Big(v_j(\ell_{jk} - t\Delta_{jk}) - |S_{jk} - (t/2) z_{jk}\Delta_{jk}|^2 \Big) \;=\; -\sum_{(j,k)\in M} \phi_{jk}(t\Delta_{jk})$$

because $v_j \ell_{jk} = |S_{jk}|^2$ for $(j,k) \notin M$. Since $\phi_{jk}(a)$ is strictly increasing in a over $[0, \Delta_{jk}]$, $V(h_x(t))$ is strictly decreasing in t over $[0,1]$. This proves C9.1.

2 C9.2 follows because $\hat{X}$ is a compact set and $h_x(t) = x - tA\Delta(x)$ is linear in t.

3 For C9.3 we will use (9.36) and (9.41) to show that there exists $\alpha > 0$ such that for all infeasible points $x \in \hat{X} \setminus X$ and all $0 \le \tau < t \le 1$, we have $f(h_x(\tau)) - f(h_x(t)) \ge \alpha \|h_x(\tau) - h_x(t)\|_1$. Fix $0 \le \tau < t \le 1$. Since $h_x(t) = x - tA\Delta(x)$, C9.5 and (9.40) imply

$$f(h_x(\tau)) - f(h_x(t)) \;=\; f((p,q,v,\ell) - \tau(\delta p, \delta q, \delta v, \delta \ell)) - f((p,q,v,\ell) - t(\delta p, \delta q, \delta v, \delta \ell))$$

where

$$\begin{bmatrix} \delta p \\ \delta q \end{bmatrix} \;=\; \begin{bmatrix} \frac{1}{2}|C|R \\ \frac{1}{2}|C|X \end{bmatrix} \Delta(x), \qquad \delta v \;=\; 0, \qquad \delta \ell \;=\; \Delta(x)$$

Hence (9.36) implies

$$f(h_x(\tau)) - f(h_x(t)) \;\ge\; c(t-\tau)\|\delta\ell\|_1 \;=\; c(t-\tau)\|\Delta(x)\|_1 \tag{9.42}$$

We will compare the right-hand side with $\|h_x(\tau) - h_x(t)\|_1 = (t-\tau)\|A\Delta(x)\|_1$. We have

$$\|h_x(\tau) - h_x(t)\|_1 \;=\; (t-\tau)\mathbf{1}^{\mathsf{T}} A\Delta(x) \;\le\; (t-\tau)\,\tilde{c} \sum_{(j,k)\in E} \Delta_{jk}(x) \;=\; (t-\tau)\,\tilde{c}\,\|\Delta(x)\|_1$$

where the first equality follows from (9.41), the last equality follows because every entry of $\Delta(x)$ is positive, and $\tilde{c} := \max_k [\mathbf{1}^{\mathsf{T}} A]_k > 0$ (recall that every entry of A is nonnegative). Substituting into (9.42) yields

$$f(h_x(\tau)) - f(h_x(t)) \;\ge\; \frac{c}{\tilde{c}} \|h_x(\tau) - h_x(t)\|_1$$

which is C9.3.

□

Remark 9.8 (Strong increase in Condition C9.5)**.** 1 C9.5 assumes f is strongly increasing in ℓ in the sense that $\frac{\partial f}{\partial \ell_j}(x) \ge c > 0$. Instead of ℓ, we can assume that f is strongly increasing in p or in q and Theorem 9.7 continues to hold. Specifically C9.5 can be modified to: there exists $c > 0$ such that for all $x \in \hat{X}$, $\frac{\partial f}{\partial \ell_l}(x) \ge c$ for all $l \in E$, or $\frac{\partial f}{\partial p_j}(x) \ge c$ for all $j \in \overline{N}$, or $\frac{\partial f}{\partial q_j}(x) \ge c$ for all $j \in \overline{N}$. Moreover Theorem 11.3 of Chapter 11.2 on exact relaxation continues to hold (see condition C11.1).

2 Continuous differentiability in C9.5 is not necessary because, since f is convex (C9.4), it is always subdifferentiable and we can replace $\frac{\partial f}{\partial \ell_j}(x) \geq c > 0$ by $\xi_j \geq c > 0$ for all subgradient ξ_j of f with respect to ℓ_j, for all j and all $x \in \hat{X}$. □

9.5 Techniques for scalability: case study

Practical OPF problems can be difficult to solve. This can be due to the sheer number of variables and constraints relative to available solution time. It can also arise from the nonsmoothness or the nonconvexity of the objective or constraint functions that often lead to numerical issues. The nonsmoothness or nonconvexity can take different forms, e.g., nonlinear power flow equations, discrete variables, nondifferentiability of the objective or constraint functions, complementarity or disjunctive constraints. All of these features are embodied in security constrained OPF (SCOPF). Practical solutions for a large optimization problem require not only the understanding of basic optimization theory, but also the development of many heuristics tailored to the structure of the specific problem.

In this section we illustrate these computational challenges and some solution techniques through an SCOPF problem proposed by the US Advanced Research Projects Agency - Energy (ARPA-E) in a multi-year Grid Optimization (GO) Competition. The GO Competition aims to accelerate the development of algorithms and software for solving large OPF problems. It was staged as a series of challenges. Challenge 1, which was conducted over the course of 2019, focused on real-time SCOPF [81]. In Chapter 9.5.1 we formulate the SCOPF problem and discuss computational challenges embodied in this problem. These challenges are also commonly found in other energy applications. In Chapters 9.5.2 and 9.5.3 we describe some of the techniques used by the top three winners of the GO Challenge 1 in addressing the nonsmoothness and scalability of SCOPF respectively [82, 83, 84]. The effective treatment of complementarity constraints, efficient contingency screening, and robust parallelization of computation have proved to be essential in devising a practical solution.

9.5.1 SCOPF formulation

The detailed SCOPF formulation is described in the official specification [81]. We present a highly simplified version to illustrate the main algorithmic ideas in [82, 83, 84] to overcome some of the computational challenges.

Constraints.

We start by formulating the constraints of the GO Challenge 1 problem. It can sometimes be difficult to exactly satisfy equality and inequality constraints in a realistic

problem. This can be due to modeling or numerical errors, not just the lack of computational resources. Energy management systems in practice however must recommend a decision even when it is impossible to satisfy all constraints of the model. One way to deal with this is to allow some constraint violations in order to practically eliminate infeasibility, but penalize them in the objective.

Let $k = 0$ denote the base case and $k = 1, \ldots, K$ denote contingencies, though we will often refer to the base case also as contingency $k = 0$. Let $\left(p^u_{ki}, q^u_{ki}\right)$ denote uncontrollable loads (or generations) and (p_{ki}, q_{ki}) denote controllable generation levels at buses $i \in \overline{N}$ in contingencies $k \geq 0$. For notational simplicity we assume without loss of generality that there is exactly one uncontrollable injection and one controllable generator at each bus i. We impose the standard voltage and generation limits:

$$\underline{v}_{ki} \leq |V_{ki}| \leq \overline{v}_{ki}, \quad \underline{p}_i \leq p_{ki} \leq \overline{p}_i, \quad \underline{q}_i \leq q_{ki} \leq \overline{q}_i, \qquad k \geq 0,\ i \in \overline{N} \tag{9.43}$$

where $\underline{v}_{ki} \leq \overline{v}_{ki}$, $\underline{p}_i \leq \overline{p}_i$, and $\underline{q}_i \leq \overline{q}_i$ are given constants.

For each line $(i, j) \in E$, let $\left(P_{k,ij}, Q_{k,ij}\right)$ denote the sending-end real and reactive power from buses i to j and $\left(P_{k,ji}, Q_{k,ji}\right)$ denote the sending-end line power in the opposite direction in contingencies $k \geq 0$. Instead of exact real and reactive power balance at bus i, we impose

$$p_{ki} \ - \ p^u_{ki} = \sum_{j:j\sim i} P_{k,ij} \ + \ \sigma^{p+}_{ki} \ - \ \sigma^{p-}_{ki}, \quad \left(\sigma^{p+}_{ki}, \sigma^{p-}_{ki}\right) \ \geq \ 0, \quad k \geq 0,\ i \in \overline{N} \tag{9.44a}$$

$$q_{ki} \ - \ q^u_{ki} = \sum_{j:j\sim i} Q_{ki} \ + \ \sigma^{q+}_{ki} \ - \ \sigma^{q-}_{ki}, \quad \left(\sigma^{q+}_{ki}, \sigma^{q-}_{ki}\right) \ \geq \ 0, \quad k \geq 0,\ i \in \overline{N} \tag{9.44b}$$

where the nonnegative variables $\left(\sigma^{p+}_{ki}, \sigma^{p-}_{ki}\right)$ are slack variables for real power violations and $\left(\sigma^{q+}_{ki}, \sigma^{q-}_{ki}\right)$ are slack variables for reactive power violations. These slack variables will be penalized in the objective as we will see below.

With a slight abuse of notation we use $(P_{k,ij}(\theta_k, |V_k|), Q_{k,ij}(\theta_k, |V_k|))$ to denote the line power as functions of voltage magnitudes and angles in contingencies $k \geq 0$ defined by:

$$P_{k,ij}(\theta_k, |V_k|) = \left(g^s_{ij} + g^m_{ij}\right)|V_{ki}|^2 \ - \ |V_{ki}||V_{kj}|\left(g^s_{ij}\cos(\theta_{ki} - \theta_{kj}) \ + \ b^s_{ij}\sin(\theta_{ki} - \theta_{kj})\right) \tag{9.45a}$$

$$Q_{k,ij}(\theta_k, |V_k|) = -\left(b^s_{ij} + b^m_{ij}\right)|V_{ki}|^2 \ + \ |V_{ki}||V_{kj}|\left(b^s_{ij}\cos(\theta_{ki} - \theta_{kj}) \ - \ g^s_{ij}\sin(\theta_{ki} - \theta_{kj})\right) \tag{9.45b}$$

where $\left(g^s_{ij}, b^s_{ij}\right)$ and $\left(g^m_{ij}, b^m_{ij}\right)$ are series and shunt admittances of line (i, j). Similarly for $\left(P_{k,ji}(\theta_k, |V_k|, \ Q_{k,ji}(\theta_k, |V_k|)\right)$ in the opposite direction on line (i, j). Then we impose the constraints

$$(P_{k,ij}, Q_{k,ij}) = (P_{k,ij}(\theta_k, |V_k|), Q_{k,ij}(\theta_k, |V_k|)), \qquad k \geq 0,\ (i, j) \in E \tag{9.45c}$$

$$(P_{k,ji}, Q_{k,ji}) = (P_{k,ji}(\theta_k, |V_k|), Q_{k,ji}(\theta_k, |V_k|)), \qquad k \geq 0,\ (i, j) \in E \tag{9.45d}$$

Line limits are expressed in terms of apparent power and the sending-end voltage magnitudes, on both ends of the lines $(i,j) \in E$:

$$\sqrt{P_{k,ij}^2 + Q_{k,ij}^2} \le P_{k,ij}^{\max} |V_{ki}| + \sigma_{k,ij}^e, \qquad k \ge 0,\ (i,j) \in E \tag{9.46a}$$

$$\sqrt{P_{k,ji}^2 + Q_{k,ji}^2} \le P_{k,ij}^{\max} |V_{kj}| + \sigma_{k,ij}^e, \qquad k \ge 0,\ (i,j) \in E \tag{9.46b}$$

$$\sigma_{k,ij}^e \ge 0, \qquad k \ge 0,\ (i,j) \in E \tag{9.46c}$$

where $P_{k,ij}^{\max}$ are given parameters and $\sigma_{k,ij}^e$ are slack variables that measure line limit violations.

When contingency $k \ge 1$ occurs the generators will adjust their real and reactive power to rebalance. This may be necessary even if the contingency is a transmission outage, i.e, the disconnection of a line or a transformer, instead of a generator outage, because the redistribution of line flows may result in different amounts of losses that need to be compensated for by these generators. Moreover the outage may also lead to deviation of tie-line flows from their scheduled values and hence nonzero area control error that must be corrected. The rebalancing is carried out at a fast timescale by frequency control mechanisms (see Chapter 6.3). The effect of the frequency control actions is modeled as follows. The real power at the generators is adjusted proportionally within their generation capacities:

$$p_{ki} = [p_{0i} + \alpha_i \Delta_k]_{\underline{p}_i}^{\overline{p}_i}, \qquad k \ge 1,\ i \in \overline{N} \tag{9.47a}$$

where p_{0i} are the output levels of generators i in the base case $k = 0$, $\left(\underline{p}_i, \overline{p}_i\right)$ are their lower and upper capacity limits, Δ_k are the total real power contingency response, and $\alpha_i \ge 0$ are called the participation factors of generators i with $\sum_i \alpha_i = 1$. (If generator i does not participate in contingency response then $\alpha_i = 0$.) Here, for real scalars x, $a \le b$, we define $[x]_a^b := \max(a, \min(x,b))$. The reactive power of generators i is adjusted within their capacity limits in an attempt to restore the voltage magnitudes $|V_{ki}|$ to their pre-contingency values, as expressed in:

$$\left\{\underline{q}_i \le q_{ki} \le \overline{q}_i,\ |V_{ki}| = |V_{0i}|\right\} \cup \left\{q_{ki} = \underline{q}_i,\ |V_{ki}| \ge |V_{0i}|\right\} \cup \left\{q_{ki} = \overline{q}_i,\ |V_{ki}| \le |V_{0i}|\right\}, \quad k \ge 1,\ i \in$$

(9.4

Variables.

To simplify notation define the following nodal vector variables for each contingency:

$$(p_k, q_k, |V_k|, \theta_k) := \left(p_{ki}, q_{ki}, |V_{ki}|, \theta_{ki},\ i \in \overline{N}\right), \quad \sigma_k^{p+} := \left(\sigma_{ki}^{p+}, i \in \overline{N}\right), \quad k \ge$$

(9.48a)

and similarly for $\left(\sigma_k^{p-}, \sigma_k^{q+}, \sigma_k^{q-}\right)$. Define the following branch variables for each contingency:

$$(P_k, Q_k) := \left(P_{k,ij}, Q_{k,ij}, P_{k,ji}, Q_{k,ji},\ (i,j) \in E\right), \quad \sigma_k^e := \left(\sigma_{k,ij}^e,\ (i,j) \in E\right), \tag{9.48b}$$

Let

$$\sigma_k := \left(\sigma_k^{p+}, \sigma_k^{p-}, \sigma_k^{q+}, \sigma_k^{q-}, \sigma_k^e\right), \qquad k \geq 0 \tag{9.48c}$$

$$x_k := (p_k, q_k, |V_k|, \theta_k,\ P_k, Q_k,\ \sigma_k), \qquad k \geq 0 \tag{9.48d}$$

$$y_k := (x_k, \Delta_k), \qquad k \geq 1 \tag{9.48e}$$

The vector x_0 collects base-case decisions and y_k collect responses to contingencies $k \geq 1$.

SCOPF.

The SCOPF problem in the GO Challenge 1 takes the form:

$$\min \sum_i c_i^g(p_{0i}) \ + \ \delta c_0(\sigma_0) \ + \ (1-\delta)\frac{1}{|K|}\sum_{k\geq 1} c_k(\sigma_k) \tag{9.49a}$$

$$\text{over } x_0, (y_k,\ k \geq 1) \tag{9.49b}$$

$$\text{s.t. } (9.43)(9.44)(9.45)(9.46)(9.47) \tag{9.49c}$$

where $c_i^g(p_{0i})$ are the generation costs at buses i in the base case, $c_0(\sigma_0)$ and $c_k(\sigma_k)$ are the penalty functions for constraint violations in the base case $k = 0$ and contingencies $k \geq 1$ respectively, defined as:

$$c_k(\sigma_k) := \sum_{i \in N}\left(c_{ki}^p\left(\sigma_{ki}^{p+} + \sigma_{ki}^{p-}\right) + c_{ki}^q\left(\sigma_{ki}^{q+} + \sigma_{ki}^{q-}\right)\right) + \sum_{(i,j)\in E} c_{k,ij}^e\left(\sigma_{k,ij}^e\right), \quad k \geq 0 \tag{9.49d}$$

and $\delta \in [0,1]$ is the weight to trade off the penalty in the base case against the average contingency penalty. The functions c_{ki}^p, c_{ki}^q, $c_{k,ij}^e$, $k \geq 0$, are convex piecewise linear, each with three segments of increasing slopes.

Two-stage formulation.

The problem (9.49) can also be treated as a two-stage optimization where the first-stage optimization is over the base-case decision x_0 and the second-stage optimization is over the contingency response y_k in each contingency $k \geq 1$. It can be rewritten as

$$\min_{x_0} \sum_i c_i^g(p_{0i}) \ + \ \delta c_0(\sigma_0) \ + \ (1-\delta)\frac{1}{|K|}\sum_{k\geq 1} r_k(x_0) \tag{9.50a}$$

$$\text{s.t. } (9.43)(9.44)(9.45)(9.46) \quad \text{with} \quad k := 0 \tag{9.50b}$$

where the recourse functions from the second-stage optimization are: for $j \geq 1$,

$$r_j(x_0) \quad := \quad \min_{y_j} \ c_j(\sigma_j) \tag{9.51a}$$

$$\text{s.t. } (9.43)(9.44)(9.45)(9.46)(9.47) \ \text{ with } \ k := j \tag{9.51b}$$

where the penalty functions $c_k(\sigma_k)$ are defined in (9.49d). The second-stage problem is used for contingency evaluation. (Two-stage optimization with recourse is studied in Chapter 13.)

Remark 9.9 (Key structures of SCOPF)**.** 1 The constraints (9.43) and (9.44) are linear. The constraint (9.45) is smooth but nonconvex. The constraints (9.46) (9.47) are nonsmooth and computationally difficult especially for interior-point methods (e.g., Ipopt [85]) used by all three teams [82, 83, 84]. All three teams devise methods to effectively handle these nonsmooth constraints, as discussed in Chapters 9.5.2 and 9.5.3.

2 The constraints (9.43) (9.44) (9.45) (9.46) apply to both the base case $k = 0$ and contingencies $k \geq 1$, but (9.47) where complementarity constraints must be dealt with applies only to contingencies $k \geq 1$ and hence only appears in the second-stage problem (9.51). As noted above (9.47) models the steady-state effect of frequency control actions after a contengency.

3 All constraints except (9.47) are separable in k. The constraint (9.47) couples the base case variables x_0 and contingency response y_k for each k. The SCOPF problem is therefore highly parallelizable and this is exploited by all three teams. □

Computational challenges

The GO Challenge 1 includes a SCOPF test where a base case decision x_0 must be computed within 10 or 45 minutes depending on the category of competition. It includes another test that computes contingency responses given the base-case decision x_0 with a time limit corresponding to 2 seconds per contingency. The problem (9.49) does not include unit commitment decisions or switched devices such as transformer taps, capacitor banks and switchable transmission lines. They are included in Challenge 2 of the GO Competition that was conducted in 2021 and introduce discrete variables that add to the computational difficulty.

There are three main computational challenges with (9.49):

1 *Nonsmoothness*. Interior-point solvers, which all three winning teams use, by default require the problem to be smooth but constraints (9.46)(9.47) are both nonsmooth. The line limit (9.46) specifies a second-order cone (studied in Chapter 8.2.1) of the form:

$$\sqrt{\sum_{i=1}^{n-1} x_i^2} \leq x_n + a_n, \qquad x \in \mathbb{R}^n,\ a_n \in \mathbb{R} \tag{9.52a}$$

This constraint is convex but nondifferentiable at the origin. The real power generation limit (9.47a) in each contingency is of the form

$$y = [x]_a^b := \max(a, \min(x, b)), \qquad x, y, a, b \in \mathbb{R} \tag{9.52b}$$

and also nondifferentiable at $x = a$ or $x = b$. The reactive power generation limit in each contingency (9.47b) is a logical constraint of the form

$$\{a \le x \le b,\ y = z\} \cup \{x = a,\ y \ge z\} \cup \{x = b,\ y \le z\}, \qquad x, y, z, a, b \in \mathbb{R} \tag{9.52c}$$

Logical constraints are generally difficult to compute.

2 *Large problem size*. For a network with G generators and M transmission lines or transformers[2], if we are to evaluate security against the outage of every single generator or line/transformer, it can increase the number of constraints by a factor of $G + M$ under $N - 1$ security. If the dispatch has to be secure against $N - k$ security then the number of constraints will be increased by a factor of $(G + M)!/(k!(G+M-k)!)$. For example the largest network used in the GO Challenge 1 has 30,000 buses, 3,526 generators, 32,020 transmission lines, 3,373 transformers [84, Table EC.1], yielding $G + M = 3,526 + 32,020 + 3,373 = 38,919$. This would have increased the number of constraints by 4 orders of magnitude under $N - 1$ security, or almost 9 orders of magnitude under $N - 2$ security ($(G+M)!/(k!(G+M-k)!) = 757,324,821$). The GO Competition adopts $N - 1$ security and specifies about 16,000 contingency scenarios which is still an increase of 4 orders of magnitude. For real-time SCOPF any practical solution must include methods to efficiently rank contingencies and solve an approximate problem with only a few highly ranked contingencies.

3 *Nonconvexity*. The power flow constraint (9.45) is nonconvex. As we have seen in Chapter 9.3, OPF is NP-hard which means that it is hard to scale in the worst case.

Methods to deal with nonconvexity through convex relaxations are studied in Chapters 10 and 11. It is however difficult to scale these methods to large problems. All three teams use a solver (Ipopt [85]) that applies a local interior-point algorithm to the nonconvex problem. Though local algorithms are generally not guaranteed to produce a global optimum, they often perform well in practice, as we have discussed in Chapter 9.4.

We therefore focus in the rest of this section on techniques use by the GO Competition teams to handle nonsmoothness and large problem size.

9.5.2 Handling nonsmoothness

The types of nonsmoothness in (9.52) are common in OPF problems. A basic approach is to approximate nondifferentiable functions by smooth functions and convert logical constraints into equivalent complementarity constraints or mixed integer constraints. For small problems the resulting complementarity problems or mixed integer problems can be solved directly. For large problems the complementarity constraints or mixed

[2] The official GO Challenge 1 formulation models transformers with slightly different capacity limits than (9.46).

integer constraints are approximated by smooth constraints that can be solved using standard solvers. We next describe three techniques from [82, 83, 84].

9.5.2.1 Smooth approximation

A common technique to handle the nondifferentiable second-order cone constraint (9.52a), $\sqrt{\sum_{i=1}^{n-1} x_i^2} \le x_n + a_n$, is to consider instead

$$\sum_{i=1}^{n-1} x_i^2 \;\le\; (x_n + a_n)^2, \qquad x_n + a_n \;\ge\; 0$$

The first constraint $\sum_{i=1}^{n-1} x_i^2 \le (x_n + a_n)^2$ is differentiable but nonconvex. Even though they are different representations of the same set, the resulting optimization problem can have different duality and computational properties; see Chapter 8.3.7. Instead of including the nonconvex constraint $\sum_{i=1}^{n-1} x_i^2 \le (x_n + a_n)^2$, [82] replaces it by a log-barrier function in the cost function for each contingency:

$$\log\left((x_n + a_n)^2 \;-\; \sum_{i=1}^{n-1} x_i^2\right)$$

which is convex.

The constraint (9.52b), $y = [x]_a^b := \max(a, \min(x, b))$, is nondifferentiable at $x = a$ or $x = b$. It is approximated by a smooth constraint in [84], as follows. The function $f(x) := \max(0, x)$, $x \in \mathbb{R}$, can be over approximated by

$$f^\epsilon(x) := \epsilon \ln\left(1 + e^{x/\epsilon}\right), \qquad \epsilon > 0 \tag{9.53a}$$

and the function $g(x) := \min(0, x)$, $x \in \mathbb{R}$, can be under approximated by

$$g^\epsilon(x) := -\epsilon \ln\left(1 + e^{-x/\epsilon}\right), \qquad \epsilon > 0 \tag{9.53b}$$

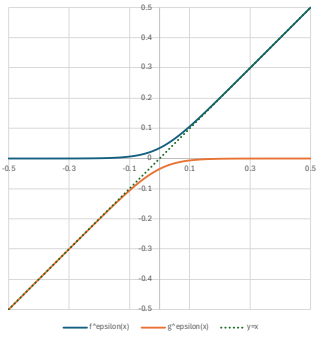
See Figure 9.4(a). The approximation errors are respectively

$$\begin{aligned} f^\epsilon(x) - \epsilon \ln 2 \;&\le f(x) \;<\; f^\epsilon(x), \qquad \epsilon > 0,\ x \in \mathbb{R} \\ g^\epsilon(x) + \epsilon \ln 2 \;&\ge g(x) \;>\; g^\epsilon(x), \qquad \epsilon > 0,\ x \in \mathbb{R} \end{aligned}$$

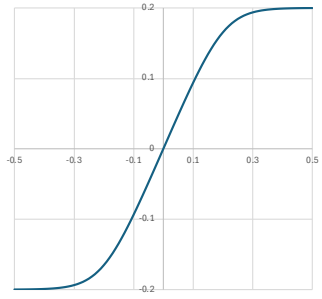
Hence the approximation becomes tight as $\epsilon \to 0$, but a small ϵ can cause numerical issues since the second derivatives $\frac{d^2}{dx^2} f^\epsilon(0)$ and $\frac{d^2}{dx^2} g^\epsilon(0)$ evaluated at $x = 0$ diverges as $\epsilon \to 0$. Hence a good heuristic must strike a balance between accuracy and numerical stability. This method leads to a smooth approximation of $h(x) := \max(a, \min(x, b))$ given by

$$h^\epsilon(x) := a \;+\; \epsilon \ln\left(1 \;+\; \frac{e^{(b-a)/\epsilon}}{1 \;+\; e^{(b-x)/\epsilon}}\right) \tag{9.53c}$$

See Exercise 9.13 for approximations of $\max(a, x)$, $\min(x, b)$ and $\max(a, \min(x, b))$. Then the constraint $y = [x]_a^b$ can be replaced by its smooth approximation $y = h^\epsilon(x)$.



(a) $f^{\epsilon}(x)$: blue, $g^{\epsilon}(x)$: orange, $y = x$: dashed



(b) $h^{\epsilon}(x)$

Figure 9.4 (a) The nonsmooth functions $f(x) := \max(0,x)$ and $g(x) := \min(0,x)$ and their smooth approximations $f^{\epsilon}(x)$ and $g^{\epsilon}(x)$ respectively ($\epsilon = 0.05$). (a) The nonsmooth functions $h(x) := [x]_a^b$ and and its smooth approximation $h^{\epsilon}(x)$ ($\epsilon = 0.05$, $a = -0.2$, $b = 0.2$).

9.5.2.2 Reformulation as mixed integer constraints

Both the nondifferentiable constraint (9.52b) and the logical constraint (9.52c) can be reformulated as equivalent mixed integer constraints using the big-M method. Specifically $y = [x]_a^b := \max(0, \min(x,b))$ if and only if there exist binary variables $\underline{z}, \overline{z} \in \{0,1\}$ such that (Exercise 9.14):

$$a \le y \le b, \quad y - a \le M\underline{z}, \quad y - x \le M(1-\underline{z}), \quad b - y \le M\overline{z}, \quad x - y \le M(1-\overline{z}), \quad x, y \in \mathbb{R}$$

where $M \in \mathbb{R}_+$ is a sufficiently large constant. Similarly the logical constraint (9.52c) can also be reformulated as an equivalent mixed integer constraint (Exercise 9.15): $(x,y,z) \in \mathbb{R}^3$ satisfies

$$\{a \le x \le b,\ y = z\} \cup \{x = a,\ y \ge z\} \cup \{x = b,\ y \le z\}$$

if and only if there exist binary variables $(\underline{z}, \overline{z})$ such that

$$\begin{aligned} &a \le x \le b, && \underline{z}, \overline{z} \in \{0,1\} && \\ &x - a \le M\underline{z}, && z - y \le M\underline{z}, && y - z \le M(1-\underline{z}) \\ &b - x \le M\overline{z}, && y - z \le M\overline{z}, && z - y \le M(1-\overline{z}) \end{aligned}$$

After all nondifferentiable constraints and logical constraints have been replaced by equivalent mixed integer constraints, the resulting mixed integer problem can be solved exactly by standard solvers if the problem is small. Otherwise one can relax the integrality constraints, e.g., relax $\underline{z}, \overline{z} \in \{0,1\}$ to $\underline{z}, \overline{z} \in [0,1]$, and solve the relaxation instead.

9.5.2.3 Reformulation as complementarity constraints

Alternatively the nondifferentiable constraint (9.52b) can be reformulated as an equivalent complementarity constraint: $y = [x]_a^b := \max(0, \min(x, b))$ if and only if there exist slack variables $\rho^-, \rho^+ \in \mathbb{R}$ such that (Exercise 9.14)

$$y + \rho^+ - \rho^- = x, \quad 0 \le \rho^- \perp y - a \ge 0, \quad 0 \le \rho^+ \perp b - y \ge 0, \quad x, y \in \mathbb{R} \tag{9.54a}$$

Similarly the logical constraint (9.52c) can also be reformulated as an equivalent complementarity constraint (Exercise 9.15): $(x, y, z) \in \mathbb{R}^3$ satisfies

$$\{a \le x \le b,\ y = z\} \cup \{x = a,\ y \ge z\} \cup \{x = b,\ y \le z\}$$

if and only if there exist slack variables $\rho^-, \rho^+ \in \mathbb{R}$ such that

$$y + \rho^+ - \rho^- = z, \quad 0 \le \rho^- \perp x - a \ge 0, \quad 0 \le \rho^+ \perp b - x \ge 0 \tag{9.54b}$$

The ability to convert between these nonsmooth constraints allows algorithm designers to choose different representations and derive different strategies to handle them, as the GO Competition teams do.

Solving complementarity constraints such as those in (9.54), e.g., given x, finding (y, ρ^-, ρ^+) that satisfies (9.54a) is called a linear complementarity problem. More generally, given a matrix $M \in \mathbb{R}^{m \times n}$ and vector $c \in \mathbb{R}^m$ the *linear complementarity problem LCP*(M, c) is to find vectors $(z, x) \in \mathbb{R}^{m+n}$ such that

$$z \ge 0, \quad Mx + c \ge 0, \quad z^\mathsf{T}(Mx + c) = 0 \tag{9.55a}$$

The shorthand for (9.55a) is

$$0 \le z \perp Mx + c \ge 0$$

Note that the set $\{(z, x) \in \mathbb{R}^2 : 0 \le z \perp x \ge 0\}$ is a nonconvex set and hence LCP can be difficult to solve exactly. We often encounter the special case where M is square and $x := z$ are imposed in (9.55a), i.e., find $z \in \mathbb{R}^n$ such that

$$z \ge 0, \quad Mz + c \ge 0, \quad z^\mathsf{T}(Mz + c) = 0 \tag{9.55b}$$

In this case a sufficient condition for the existence and uniqueness of a solution z is that M satisfies $x^\mathsf{T} M x \ge 0$ for all $x \in \mathbb{R}^m$ whether or not M is symmetric.[3] In particular M being positive definite or symmetric is not necessary (Exercise 9.18). A *nonlinear complementarity problem NCP*(h) for a function $h : \mathbb{R}^n \to \mathbb{R}^m$ is to find vectors $(z, x) \in \mathbb{R}^{n+m}$ such that

$$z \ge 0, \quad h(x) \ge 0, \quad z^\mathsf{T} h(x) = 0 \tag{9.56}$$

It reduces to LCP(M, a) when $h(x) := Mx + c$. Complementarity problems originally arise as solving KKT conditions of optimization problems; in particular solving the KKT condition of a quadratic program is a linear complementarity problem (Exercise

[3] For a matrix M over the field $\mathbb{R}$, we define M to be positive definite only for symmetric M; see Definition A.2 and Remark A.1 in Chapter A.5.

9.16). To see that, given x, finding (y, ρ^-, ρ^+) that satisfies (9.54a) is a LCP, substitute $y = \rho^- - \rho^+ + x$ into the complementarity constraints to get:

$$0 \le \rho^- \perp (\rho^- - \rho^+) + (x - a) \ge 0$$
$$0 \le \rho^+ \perp (-\rho^- + \rho^+) + (b - x) \ge 0$$

or finding a solution $(\rho^-, \rho^+) \in \mathbb{R}^2$ to the following LCP(M, c):

$$0 \le \begin{bmatrix} \rho^- \\ \rho^+ \end{bmatrix} \perp \underbrace{\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}}_{M} \begin{bmatrix} \rho^- \\ \rho^+ \end{bmatrix} + \underbrace{\begin{bmatrix} x - a \\ b - x \end{bmatrix}}_{c} \ge 0$$

After all nondifferentiable constraints and logical constraints have been replaced by equivalent complementarity constraints, the resulting problem can be solved exactly by LCP solvers if the problem is small. Otherwise the complementarity constraints can be approximated by simpler smooth constraints that can be solved for larger problems, as we discuss next.

Consider the complementarity constraint of the form

$$x \ge 0, \qquad y \ge 0, \qquad xy = 0, \qquad x, y \in \mathbb{R} \tag{9.57}$$

The bilinear constraint $xy = 0$ is nonconvex. The function $\phi(x, y) := x + y - \sqrt{x^2 + y^2}$ is called the Fischer-Burmeister function and well studied for nonlinear complementarity problems. It is easy to check that (9.57) holds if and only it $\phi(x, y) = 0$. A common way to handle the complementarity constraint (9.57) is to replace it with the Fischer-Burmeister function $\phi(x, y)$ as a penalty term in the objective. The function ϕ is convex and Lipschitz continuous. (It is however not differentiable at $(0, 0)$ and [82] finds this approach numerically unstable for the SCOPF problem.)

In many applications some bounds on x, y are known, e.g., the capacity limit of the largest generator poses a bound on all generators' output levels a priori. Suppose $\underline{x} \le x \le \overline{x}$ and $\underline{y} \le y \le \overline{y}$ where $(\underline{x}, \overline{x})$ and $(\underline{y}, \overline{y})$ are known. Then the bilinear function $f(x, y) := xy$, $x, y \in \mathbb{R}$ can be approximated by a *McCormick envelop*. Generally a McCormick envelop is a convex relaxation of a nonconvex function $f(x, y)$, $x, y \in \mathbb{R}$. For the bilinear constraint:

$$w = xy, \qquad \underline{x} \le x \le \overline{x}, \qquad \underline{y} \le y \le \overline{y}, \qquad w, x, y \in \mathbb{R}$$

the relaxation is a set of linear inequalities in $(w, x, y) \in \mathbb{R}^3$ (Exercise 9.19):

$$\text{lower bounds on } w: \qquad w \ge \underline{y}x + \underline{x}y - \underline{x}\,\underline{y}, \qquad w \ge \overline{y}x + \overline{x}y - \overline{x}\,\overline{y} \tag{9.58a}$$
$$\text{upper bounds on } w: \qquad w \le \overline{y}x + \underline{x}y - \underline{x}\,\overline{y}, \qquad w \le \underline{y}x + \overline{x}y - \overline{x}\,\underline{y} \tag{9.58b}$$
$$\underline{x} \le x \le \overline{x}, \qquad \underline{y} \le y \le \overline{y} \tag{9.58c}$$

The quality of the approximation depends on how tight the lower and upper bounds on x, y are.

Example 9.3 (McCormick envelops)**.** Consider the QCQP:

$$\min_{x\in\mathbb{R}^n} \sum_{i,j} c_{ij}x_ix_j \qquad \text{s.t.} \qquad \sum_{i,j} c^l_{ij}x_ix_j \le b_l, \qquad l=1,\dots,L$$
$$\underline{x} \le x \le \overline{x}$$

Derive a convex relaxation based on the McCormick envelops.

Solution. Let $w_{ij} := x_ix_j$. Applying (9.58) leads to the convex relaxation:

$$\min_{x\in\mathbb{R}^n} \sum_{i,j} c_{ij}w_{ij} \qquad \text{s.t.} \qquad \sum_{i,j} c^l_{ij}w_{ij} \le b_l, \qquad l=1,..$$
$$w_{ij} \ge \underline{x}_jx_i+\underline{x}_ix_j-\underline{x}_i\underline{x}_j, \quad w_{ij} \ge \overline{x}_jx_i+\overline{x}_ix_j-\overline{x}_i\overline{x}_j, \quad i,j=1,..$$
$$w_{ij} \le \underline{x}_jx_i+\overline{x}_ix_j-\overline{x}_i\underline{x}_j, \quad w_{ij} \le \overline{x}_jx_i+\underline{x}_ix_j-\underline{x}_i\overline{x}_j, \quad i,j=1,..$$
$$\underline{x} \le x \le \overline{x}$$

□

Suppose there are known upper bounds on $x, y \in \mathbb{R}$ in the complementarity constraint (9.57):

$$\overline{x} \ge x \ge 0, \qquad \overline{y} \ge y \ge 0, \qquad xy = 0, \qquad x,y\in\mathbb{R} \tag{9.59}$$

(Note that this may not be the case for KKT conditions as one of the variables x, y will be a dual variable which may not have an upper bound.) In this case, substituting $w = xy$ into (9.59) leads to the following linear relaxation of the nonconvex constraint $xy = 0$:

$$0 \le x \le \overline{x}, \qquad 0 \le y \le \overline{y}, \qquad w = 0, \qquad x,y\in\mathbb{R}$$
$$\max\{0,\ \overline{y}x+\overline{x}y-\overline{xy}\} \le 0 \le \min\{\overline{x}y,\ \overline{y}x\}$$

or equivalently

$$0 \le x \le \overline{x}, \qquad 0 \le y \le \overline{y}, \qquad 0 \le \min\{\overline{x}y,\ \overline{y}x\}, \qquad x,y\in\mathbb{R}$$

9.5.3 Scaling computation

To scale the computation of (9.49), or its two-stage formulation (9.50)(9.51), efficient software implementation is critical, especially how to effectively use multi-core platforms for parallel computation, how to detect and reduce numerical instability, and how to handle software failures such as solver divergence or convergence to an infeasible point even when the problem is provably feasible. For example, the number of nonlinear subproblems that needs to be solved in [82] can be as high as 100,000, each with 2,000,000 variables and constraints. Software implementation issues in such a large-scale computational regime are highly nontrivial.

In the rest of this subsection however we will focus only on algorithmic techniques

for scalability. In particular we summarize four techniques used in [82, 83, 84] to illustrate some of the ideas in solving industrial-scale OPF problems.

Approximate or relax nonsmooth functions.

To avoid infeasibility, some hard constraints $g(x) = 0$, $h(x) \le 0$ such as power balanced have been replaced by soft constraints $g(x) = \sigma_1$, $h(x) \le \sigma_2$ respectively and a violation cost $c(\sigma_1, \sigma_2)$ is added to the cost function to penalize constraint violation. Nonsmooth cost functions $f(x)$, e.g., piecewise linear (convex) constraint violation costs $c^p_{ki}(s), c^q_{ki}(s), c^e_{k,ij}(s)$ in (9.49d), are approximated by quadratic functions of the form $\hat{f}(x) := ax^2 + bx$ with parameters (a, b) determined by linear regression. Nondifferentiable or combinatorial constraints, e.g., (9.46)(9.47), are approximated or relaxed by smooth constraints, as discussed in Chapter 9.5.2. Smooth problems are generally easier to solve and what most standard solvers can handle.

Approximate optimal recourse function $r_k(x_0)$.

The approach of [82] uses the two-stage formulation (9.50)(9.51) of the SCOPF problem. A two-stage problem is computationally difficult because an explicit form of the second-stage recourse function $r_k(x_0)$ is generally not available. Moreover the recourse function is in general nonsmooth; we will study nonsmooth convex optimization in Chapter 12 and two-stage stochastic optimization in Chapter 13.4. The key idea of [82] is to approximate $r_k(x_0)$ by an explicit polynomial function $\hat{r}_k(x_0; \pi_k)$ of the form

$$\hat{r}_k(x_0; \pi_k) \;:=\; \pi_k \hat{f}_k(x_0) \tag{9.60a}$$

where $\hat{f}_k(x_0)$ is a low-degree polynomial that depends on the device (a generator or a line) that is disconnected in contingency k under $N-1$ security and π_k is a scaling factor in the approximation to be determined. This reduces the first-stage to a much simpler approximate problem of the form

$$\min_{x_0} \quad \hat{f}_0(x_0) \;+\; \frac{1}{|\hat{K}|} \sum_{k \in \hat{K}} \hat{r}_k\,(x_0; \pi_k) \tag{9.60b}$$

where the cost functions $\hat{f}_0$ and $\hat{r}_k$ are either quadratic or low-degree polynomials and $\hat{K}$ is a reduced set of credible contingencies (see discussions below). Given an optimal solution x_0 of the approximate first-stage problem (9.60), an approximate version of the second-stage problem (9.51) is solved to determine the scaling factor π_k. Since the second-stage problem is separable in k, given x_0, the approximate (9.51) is solved in parallel across contingencies k, to obtain an (approximate) optimal $r_k(x_0)$. Using (9.60a), $\pi_k(x_0)$ is then set to be

$$\pi_k(x_0) := \frac{r_k(x_0)}{\hat{f}_k(x_0)} \tag{9.61}$$

This leads to an algorithm that solves approximate first-stage problem and approximate second-stage problem iteratively: for $t = 0, 1, \ldots$, repeat until a stopping criterion is satisfied ($\pi_k(0) = 0$, i.e., start with the base case):

1 Given $\hat{r}_k(x_0; \pi_k(t))$, solve the approximate first-stage problem (9.60) to obtain an optimal solution $x_0(t+1)$.
2 Given $x_0(t+1)$, solve an approximate version of the second-stage problems (9.51) in parallel to obtain optimal solutions $r_k(x_0(t+1))$. Construct $\pi_k(t+1) := \pi_k(x_0(t+1))$ according to (9.61) and $\hat{r}_k(x_0; \pi_k(t+1))$ according to (9.60a).

The two subproblems in this algorithm are made much simpler by techniques that handle nonsmoothness (discussed in Chapter 9.5.2) and techniques that screen contingencies quickly to identify and include only contingencies that are likely to have large recourse costs $r_k(x_0)$, which we discuss next.

Fast contingency selection.

The approach of [83] focuses on continuously and iteratively evaluate contingencies and include only the top three contingencies in the solution of SCOPF (9.49) in each iteration. Three main contingency selection techniques are used to identify top three contingencies:

1 *Initial ranking using machine learning.* Initial contingency ranking uses supervised learning to predict the importance of a contingency on overall cost based on various features, such as different expressions of generation levels and line power, generator ratings, degrees of buses, etc. It finds that the apparent line power
$$\max\left\{\sqrt{P^2_{0,i(k)j(k)} + Q^2_{0,i(k)j(k)}}, \sqrt{P^2_{0,j(k)i(k)} + Q^2_{0,j(k)i(k)}}\right\}$$
has the best predictive power. This is consistent with the intuition used to approximate the recourse function $r_k(x_0)$ in [82] (it is used in $\hat{f}_k(x_0)$ in (9.60a)).
2 *Contingency evaluation.* Each contingency k identified by the initial ranking as credible is then evaluated more carefully by solving the second-stage problem (9.51), in two steps. First, given a first-stage decision x_0, an upper bound on the second-stage cost $r_k(x_0)$ is computed by solving a reduced problem with only the power flow equations and linear constraints associated with complementarity constraints predicted by an active set method to handle complementarity constraints. In particular this reduced problem does not include any operational constraints. Only if this upper bound exceeds a certain threshold will a full evaluation of the contingency be carried out by solving the second-stage problem using the active set method.
3 *Dominated contingencies.* Inclusion of the constraints due to contingency j may cause the constraints due to other contingencies k to be automatically (possibly approximately) satisfied. To identify these constraints, let $\sigma_k^{\max}$ be the largest entry of the vector σ_k defined in (9.48c), i.e., $\sigma_k^{\max}$ is the largest slack variable measuring

the violation of power balance or a line limit in contingency k. We say that *contingency k is dominated by contingency j* if $\sigma_j^{\max} > \sigma_k^{\max}$. Only contingencies that are not dominated by another contingency are included in the solution of the master problem (9.49).

The screening of contingencies and solving of SCOPF (9.49) with top three contingencies in each iteration both require techniques to handle complementarity constraints, evaluate contingencies quickly, remove dominated contingencies, and effective parallelization of computation.

Exploit distributed problem structure by ADMM algorithm.

The approach of [84] uses smooth approximation of constraint (9.47) and develops an ADMM-based algorithm to exploit the problem's distributed structure. The base case $k = 0$ and the contingencies $k \geq 1$ are coupled only through the first-stage decision x_0 in the constraint (9.47) that appears in the set of second-stage problems (9.51), one for each contingency $k \geq 1$. By introducing a local copy x_k^0 of x_0 for each contingency subproblem these second-stage problems are decoupled and can therefore be computed in parallel, with a consensus constraint that all local copies x_k^0 equal to x_0 at optimality. Hence the SOCP problem (9.49) can be equivalently reformulated into the form

$$\min \; f_0(x_0) \; + \; \sum_{k\geq 1} f_k\left(x_k^0, y_k\right) \tag{9.62a}$$

$$\text{over } x_0, \left(x_k^0, y_k, \; k \geq 1\right) \tag{9.62b}$$

$$\text{s.t. } x_0 \in X_0, \; \left(x_k^0, y_k\right) \in X_k, \; k \geq 1 \tag{9.62c}$$

$$x_k^0 = x_0, \; k \geq 1 \tag{9.62d}$$

where the constraint $x_0 \in X_0$ means that x_0 satisfies (9.43)–(9.46), and the constraint $\left(x_k^0, y_k\right) \in X_k$ means that y_k satisfies (9.43)–(9.46) and $\left(x_k^0, y_k\right)$ satisfies the smooth approximations of (9.47). These constraints (9.62c) are decoupled across k. The coupling of the $K+1$ variables x_0 and (x_k^0, y_k), $k \geq 1$, is only through K linear (consensus) constraint (9.62d). This is a form that is suitable for distributed solution using the alternating direction method of multipliers (ADMM) studied in Chapter 8.5.5.

Define the augmented Lagrangian function that relaxes the coupling constraint (9.62d):

$$L_\rho\left(x_0, (x_k^0, y_k), k \geq 1; \lambda\right) := f_0(x_0) + \sum_{k\geq 1} f_k\left(x_k^0, y_k\right) \; + \; \lambda^\mathsf{T}\left(\mathbf{1}_K \otimes x_0 - x^0\right) \; + \; \frac{\rho}{2}\left\|\mathbf{1}_K \otimes x_0 - x^0\right\|_2^2$$

where $\mathbf{1}_K$ is the vector of all 1s of size K and $x^0 := (x_1^0, \cdots, x_K^0)$ is a column vector.

The ADMM algorithm is

$$x_0(t+1) \;:=\; \arg\min_{x_0 \in X_0} L_\rho\left(x_0, \left(x_k^0(t), y_k(t)\right), k \geq 1; \lambda(t)\right) \tag{9.63a}$$

$$\left(x_k^0(t+1), y_k(t+1)\right) \;:=\; \arg\min_{(x_k^0, y_k) \in X_k} L_\rho\left(x_0(t+1), \left(x_k^0, y_k\right), k \geq 1; \lambda(t)\right), \quad k \geq 1 \tag{9.63b}$$

$$\lambda(t+1) \;:=\; \lambda(t) \;+\; \rho\left(\mathbf{1}_K \otimes x_0(t+1) - x^0(t+1)\right) \tag{9.63c}$$

The expression (9.63b) is a shorthand for one-pass of a Gauss-Seidel method across the K contingencies: for $k = 1, \ldots, K$,

$$\left(x_k^0(t+1), y_k(t+1)\right) \;:=\; \arg\min_{(x_k^0, y_k) \in X_k} L_\rho\left(x_0(t+1), (x_1^0(t+1), y_1(t+1)), \ldots, \left(x_k^0, y_k\right), \ldots, \left(x_K^0(t), y_K(t)\right); \lambda(t\right.$$

Given the Lagrange multiplier $\lambda_k(t)$ associated with contingency k, the $K+1$ subproblems (9.63a)(9.63b) can be computed in parallel. The algorithm of [84] applies this idea to SCOPF (9.62) with two main refinements. First it relaxes the coupling constraint (9.62d) with a slack variable z_k for each contingency $k \geq 1$ which is penalized in the objective function with a term $\beta\|z_k\|_2^2$. As a result the solution returned by the ADMM algorithm may violate by a large amount the coupling constraint and is therefore infeasible for the original SCOPF. The second refinement is an outer loop where the weight β on the penalty is increased if the worst violation $\max_{k\geq 1}\|z_k\|$ across contingencies is too large and the approximate SCOPF is solved again using ADMM. The outer loop terminates when $\max_{k\geq 1}\|z_k\|$ is small enough (and the stationarity condition is sufficiently satisfied). Even though the problem is nonconvex it is proved in [84] that the two-level ADMM algorithm with both the inner and outer loops converges under the condition that each inner-loop iteration (9.63a)(9.63b) produces sufficient descent.

9.6 Bibliographical notes

As for most chapters, this section is now a placeholder with references collected in a somewhat random fashion during the writing of the text. Major rewrite later.

There has been a great deal of research on OPF since Carpentier's first formulation in 1962 [86]. An early solution appears in [87] and extensive surveys can be found in e.g. [88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 39, 101]. It is nonconvex and has been shown to be NP-hard in general [102, 69, 70].

Many references for 3-phase OPF: e.g. [103, 104, 105]

There are many excellent texts on optimization theory especially for convex prob-

lems, e.g., [62, 57, 54]. Optimization texts with power system applications include [106, 107]. In particular Chapter 8.5.3 mostly follows the presentation in [57, Chapter 11]. A popular interior-point solver for OPF problems is [108].

A classic text on computational complexity is [75]. OPF has been shown to be NP-hard in general [102, 69, 70, 72, 74]. [77] surveys combinatorial OPF and proves approximation results and conditions for exactness (when there are no discrete variables). It shows that OPF with discrete injections cannot be efficiently approximated. The hardness results complement those in [73, 68, 69, 70]; see [77, Chapter 5] and its Section 5.6 for comparison.

Chapter 9.4 on global optimality is taken form [80, 79]

[109] shows that, by dualizing clique tree conversion, a class of nonconvex problems, including OPF problems, the per-iteration cost of an interior-point method is linear $O(n)$ in time and in memory, so an ϵ-accurate and ϵ-feasible iterate is obtained after $O(\sqrt{n}\log(1/\epsilon))$ iterations in $O(n^{1.5}\log(1/\epsilon))$ time.

9.7 Problems

Chapter 9.1

Exercise 9.1 (OPF: power losses as quadratic form)**.** We revisit Exercise 4.12 to write power losses as quadratic forms. For each line $(j,k) \in E$, let its admittances be $y^s_{jk} = g^s_{jk} + \mathbf{i}b^s_{jk}$ and $y^m_{jk} = g^m_{jk} + \mathbf{i}b^m_{jk}$. Suppose $y^s_{jk} = y^s_{kj}$ and $g^s_{jk} \geq 0$, $g^m_{jk} \geq 0$ (these conditions are satisfied if (j,k) models a transmission line).

1 Define the total real power loss as:

$$C(V) \;:=\; \sum_j \mathrm{Re}\left(s_j(V)\right) \;=\; \sum_j \mathrm{Re}\left(\sum_{k:j\sim k} \bar{y}^s_{jk}\left(|V_j|^2 - V_j\bar{V}_k\right) \;+\; \bar{y}^m_{jj}|V_j|^2 \right)$$

Show that $C(V)$ is a quadratic form $C(V) = V^{\mathsf{H}}C_0V$ where the cost matrix $C_0 := \frac{1}{2}\left(Y^{\mathsf{H}} + Y\right) = \mathrm{Re}(Y)$ is the Hermitian component of the admittance matrix Y. Show that C_0 is a positive semidefinite matrix.

2 Suppose $y^m_{jk} = y^m_{kj} = 0$. Define the total thermal loss as:

$$C(V) \;:=\; \sum_{(j,k)\in E} r^s_{jk}\,|I_{jk}(V)|^2 = \sum_{(j,k)\in E} r^s_{jk}\left|y^s_{jk}(V_j - V_k)\right|^2$$

where $z^s_{jk} = r^s_{jk} + \mathbf{i}x^s_{jk} := 1/y^s_{jk}$. Show that $C(V)$ is a quadratic form $C(V) = V^{\mathsf{H}}C_0V$ where the cost matrix $C_0 = \mathrm{Re}(Y)$ when $y^m_{jk} = y^m_{kj} = 0$.

3 Therefore, when $y^m_{jk} = y^m_{kj} = 0$, the total real power loss in part 1 reduces to the

total thermal loss in part 2. As an alternative proof that sheds light on the physics behind this mathematical property, show that

$$\sum_j s_j(V) \;=\; \sum_{(j,k)\in E} z^s_{jk}\left|\frac{V_j - V_k}{z^s_{jk}}\right|^2 \;+\; \sum_{(j,k)\in E}\left(\bar{y}^m_{jk}|V_j|^2 \;+\; \bar{y}^m_{kj}|V_k|^2\right)$$

where $(V_j - V_k)/z^s_{jk}$ is the current through the series impedance of line (j,k).

Exercise 9.2 (OPF: quadratic line limit)**.** Consider the line limit

$$|S_{jk}(V)|^2 \;\le\; \overline{S}^2_{jk}, \qquad\qquad |S_{kj}(V)|^2 \;\le\; \overline{S}^2_{kj}, \qquad\qquad (j,k)\in E$$

where

$$S_{jk}(V) \;:=\; V_j\bar{I}_{jk}(V) \;=\; \bar{y}^s_{jk}\left(|V_j|^2 - V_j\bar{V}_k\right) \;+\; \bar{y}^m_{jk}|V_j|^2, \qquad (j,k)\in E$$

$$S_{kj}(V) \;:=\; V_k\bar{I}_{kj}(V) \;=\; \bar{y}^s_{kj}\left(|V_k|^2 - V_k\bar{V}_j\right) \;+\; \bar{y}^m_{kj}|V_k|^2, \qquad (j,k)\in E$$

Show that the line limit can be written as an inhomogeneous quadratic form.

Exercise 9.3 (Inner product and trace)**.** Let $A, B \in \mathbb{C}^{n\times n}$ be square complex matrices. The inner product of A, B is defined to be $A \cdot B := \mathrm{tr}(A^{\mathsf{H}}B)$. Show that:

1 $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.
2 $A \cdot B := \mathrm{tr}(A^{\mathsf{H}}B) = \mathrm{tr}(AB)$ if A is Hermitian. The converse does not necessarily hold.
3 If A and B are both Hermitian then $A \cdot B = B \cdot A$.

Exercise 9.4 (Skew-symmetric and Hermitian matrices)**.** Show that:

1 If $C \in \mathbb{R}^{n\times n}$ is a skew symmetric matrix (i.e., $C^{\mathsf{T}} = -C$) then $x^{\mathsf{T}}Cx = 0$ for any $x \in \mathbb{R}^n$.
2 If $C \in \mathbb{C}^{n\times n}$ is a Hermitian matrix (i.e., $C^{\mathsf{H}} = C$) then $x^{\mathsf{H}}Cx \in \mathbb{R}$ for any $x \in \mathbb{C}^n$.
3 If $C \in \mathbb{C}^{n\times n}$ is a Hermitian matrix, then $\mathrm{tr}(CX)$ is a real number for any rank-1 matrix $X \in \mathbb{C}^{n\times n}$ (psd or nsd).
4 Let $C := C_r + \mathbf{i}C_i$ where $C_r, C_i \in \mathbb{R}^{n\times n}$. If C is Hermitian then $C_r^{\mathsf{T}} = C_r$ and $C_i^{\mathsf{T}} = -C_i$.

Let $A \in \mathbb{C}^{n\times n}$ and $x \in \mathbb{C}^n$. Define the Hermitian and skewed Hermitian components of A:

$$B_r \;:=\; \frac{1}{2}\left(A + A^{\mathsf{H}}\right), \qquad B_i \;:=\; \frac{1}{2\mathbf{i}}\left(A - A^{\mathsf{H}}\right)$$

Show that

5. B_r and B_i are both Hermitian for arbitrary A, so that $x^{\mathsf{H}}B_r x$ and $x^{\mathsf{H}}B_i x$ are both real numbers.
6. Moreover $x^{\mathsf{H}}Ax = x^{\mathsf{H}}B_r x + \mathbf{i}x^{\mathsf{H}}B_i x$.

Exercise 9.5 (Real QCQP)**.** Show that the complex QCQP (9.10) is equivalent to the real QCQP (9.11) of twice the dimension. Show that D_l are symmetric matrices.

Exercise 9.6 (Homogenization)**.** Let $x, a, b \in \mathbb{C}^n$.

1 Let $e_j \in \{0,1\}^n$ be the unit vector with a single 1 at the jth position. Show that the set of inequalities $a_j \le x_j \le b_j$, $j = 1,\dots,n$, is equivalent to the following set of homogeneous quadratic inequalities in $(\hat{x}, t)$ with $x := \hat{x}\bar{t}$: for $j = 1,\dots,n$,

$$\mathrm{Re}(a_j) \le \begin{bmatrix}\hat{x}\\ t\end{bmatrix}^{\mathsf{H}} \begin{bmatrix}0 & \zeta_j\\ \zeta_j^{\mathsf{H}} & 0\end{bmatrix} \begin{bmatrix}\hat{x}\\ t\end{bmatrix} \le \mathrm{Re}(b_j), \quad \mathrm{Im}(a_j) \le \begin{bmatrix}\hat{x}\\ t\end{bmatrix}^{\mathsf{H}} \begin{bmatrix}0 & \mathbf{i}\zeta_j\\ -\mathbf{i}\zeta_j^{\mathsf{H}} & 0\end{bmatrix} \begin{bmatrix}\hat{x}\\ t\end{bmatrix} \le \mathrm{Im}(b_j \tag{9.64a}$$

$$1 \le \begin{bmatrix}\hat{x}\\ t\end{bmatrix}^{\mathsf{H}} \begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix} \begin{bmatrix}\hat{x}\\ t\end{bmatrix} \le 1 \tag{9.64b}$$

where $\zeta_j = e_j/2$.

2 Let $c_j \in \mathbb{C}^n$ for $j = 1,\dots,n$. Show that the set of inequalities $a_j \le c_j^{\mathsf{H}} x \le b_j$, $j = 1,\dots,n$, is equivalent to (9.64) with $\zeta_j = e_j/2$ replaced by $\zeta_j = c_j/2$.

Chapter 9.2.

Exercise 9.7.

Chapter 9.3.

Exercise 9.8 (Angle constraint)**.** Show that (9.23e) is equivalent to the constraint on apparent power $P_{jk}^2 + Q_{jk}^2 \le S(\bar{\theta})$ for some real number $S(\bar{\theta})$ that depends on $\bar{\theta}$, provided $|\theta_{jk}| \le \bar{\theta} \in (0, \pi/2]$.

Exercise 9.9 (NP-hardness [70])**.** Let $x := (p_j, j \in N_G;\ \theta_j, j \in N_G \cup N_L;\ P_{jk}, Q_{jk}, (j,k) \in E)$ be a solution of (9.23).

1 Consider the line flow $P_{a0} = a\hat{P}(\theta_{a0}) := a\,(g(1-\cos\theta_{a0}) - b\sin\theta_{a0})$ on line $(a,0) \in E$. Show that, if $\theta_{a0} < 0$, then $\hat{P}(\theta_{a0}) < 0$.

2 Show that $\hat{Q}(-\overline{\theta}) > 0$.
3 Show that $\hat{P}(-\overline{\theta})\hat{Q}(\theta_{0a}) \geq \hat{Q}(-\overline{\theta})\hat{P}(\theta_{0a})$ with equality if and only if $\theta_{a0} \in \{0, \overline{\theta}\}$.

(Hint: use $\tan(\phi/2) = (1-\cos\phi)/\sin\phi$, $|\theta_{a0}| \leq \overline{\theta} \leq \pi/2$ and $\hat{P}(-\overline{\theta}) < 0$.)

Exercise 9.10 (NP-hardness of nonconvex quadratic program [110])**.** Show that determining the global solution of smooth nonlinear program is NP-hard by reducing the NP-complete subset sum problem to nonconvex quadratic program. (Hint: Write a subset sum problem instance (A, σ) in terms of determining a binary vector x of size $|A|$ and reduce it to a smooth nonconvex quadratic program.)

Chapter 9.4

Exercise 9.11 (Feasible region of OPF [111])**.** By introducing slack variables, the constraints that define the feasible region of OPF (e.g., (9.16)) is of the form $f(x) = 0$ for some $f : \mathbb{R}^n \to \mathbb{R}^n$. Consider the (energy) function $E(x) := \frac{1}{2}\|f(x)\|_2^2$ and the problem $\min_{x\in\mathbb{R}^n} E(x)$.

1 What is the gradient flow (continuous time version of gradient descent algorithm) to minimize $E(x)$?
2 Show that if $\bar{x}$ is a feasible point (i.e., $f(\bar{x}) = 0$), then $\bar{x}$ is an stable equilibrium point of the gradient flow dynamic. The converse is not necessarily true.
3 Show that the converse is true if the Jacobian $\frac{\partial f}{\partial x}(x)$ is nonsingular on $\mathbb{R}^n$.

Exercise 9.12 (OPF global optimality)**.** This exercise fills in some details in the proof of Theorem 9.7 in Chapter 9.4.4.

1 Show that condition C9.5 implies (9.36).
2 For $\{h_x : x \in \hat{X} \setminus X\}$ defined in (9.39), show that $h_x(t) \in \hat{X}$ for all $t \in [0,1]$ and $h_x(1) \in X$.

Chapter 9.5

Exercise 9.13 (Smooth approximation)**.** This problem considers smooth approximations of $\max(a,x)$ and $\min(a,x)$.

1 Let $f(x) := \max(0,x)$ and its over approximation $f^\epsilon(x) := \epsilon \ln\left(1 + e^{x/\epsilon}\right)$ for $x \in \mathbb{R}$ and $\epsilon > 0$. For any $\epsilon > 0$ show that $f^\epsilon(x) - \epsilon\ln 2 \leq f(x) < f^\epsilon(x)$ for all $x \in \mathbb{R}$, with equality if and only if $x = 0$.

2 What is the corresponding approximation $\tilde{f}^{\epsilon}(x)$ for $\tilde{f}(x) := \max(a,x)$ for any $a \in \mathbb{R}$?
3 Let $g(x) := \min(0,x)$ and its under approximation $g^{\epsilon}(x) := -\epsilon \ln\left(1+e^{-x/\epsilon}\right)$ for $x \in \mathbb{R}$ and $\epsilon > 0$. For any $\epsilon > 0$ show that $g^{\epsilon}(x) < g(x) \leq g^{\epsilon}(x) + \epsilon \ln 2$ for all $x \in \mathbb{R}$, with equality if and only if $x = 0$.
4 What is the corresponding approximation $\tilde{g}^{\epsilon}(x)$ for $\tilde{g}(x) := \min(x,b)$ for any $b \in \mathbb{R}$?
5 What is the approximation for $h(x) := \max(a, \min(x,b))$ for $a < b$ if we apply the approximations for $\tilde{f}$ and $\tilde{g}$?

Exercise 9.14 (Complementarity and big-M constraints)**.** Consider the nondifferentiable constraint $y = [x]_a^b := \max(0, \min(x,b))$ where $x, y \in \mathbb{R}$ are variables and $a < b \in \mathbb{R}$ are given constants.

1 Show that it is equivalent to a complementarity constraint: $y = [x]_a^b := \max(0, \min(x,b))$ if and only if there exist slack variables $(\rho^-, \rho^+) \in \mathbb{R}^2$ such that

$$y + \rho^+ - \rho^- = x, \quad 0 \leq \rho^- \perp y - a \geq 0, \qquad 0 \leq \rho^+ \perp b - y \geq 0, \quad x, y \in \mathbb{R} \tag{9.65}$$

Given $x \in \mathbb{R}$, show that finding a solution $(y, \rho^-, \rho^+) \in \mathbb{R}^3$ to this complementarity constraint is a standard linear complementarity problem LCP(M,q) for a 2×2 matrix M.

2. Show that it is equivalent to a big-M mixed integer constraint: $y = [x]_a^b := \max(0, \min(x,b))$ if and only if there exist binary variables $\underline{z}, \overline{z}$ such that

$$\begin{aligned} a \leq y \leq b, & \qquad \underline{z}, \overline{z} \in \{0,1\} && \text{(9.66a)} \\ y - a \leq M\underline{z}, & \qquad y - x \leq M(1-\underline{z}) && \text{(9.66b)} \\ b - y \leq M\overline{z}, & \qquad x - y \leq M(1-\overline{z}) && \text{(9.66c)} \end{aligned}$$

where $M \in \mathbb{R}_+$ is a sufficiently large constant. What value of $(\underline{z}, \overline{z})$ will result in infeasibility?

3. Show that it is also equivalent to (in the same sense):

$$\begin{aligned} x - a \leq M\underline{z}, \; a - x \leq M(1-\underline{z}) && \text{(9.67a)} \\ b - x \leq M\overline{z}, \; x - b \leq M(1-\overline{z}) && \text{(9.67b)} \end{aligned}$$

together with (the nonlinear equality)

$$(y-a)(1-\underline{z}) + (y-b)(1-\overline{z}) + (y-x)\underline{z}\overline{z} = 0, \qquad \underline{z}, \overline{z} \in \{0,1\} \tag{9.67c}$$

What value of $(\underline{z}, \overline{z})$ will result in infeasibility?

Exercise 9.15 (Complementarity and big-M constraints)**.** Consider the logical constraint (9.52c) on the variables $(x, y, z) \in \mathbb{R}^3$, reproduced here

$$\{a \le x \le b,\ y = z\} \cup \{x = a,\ y \ge z\} \cup \{x = b,\ y \le z\}, \qquad (x, y, z) \in \mathbb{R}^3 \quad (9.68)$$

where $a < b$ are given scalars. Unlike the complementarity problem in Exercise 9.14, the equality constraint $y = z$ here involves another variable x.

1 Show that it is equivalent to the following complementarity constraint: $(x, y, z) \in \mathbb{R}^3$ satisfies (9.68) if and only if there exist slack variables $\rho^-, \rho^+ \in \mathbb{R}$ such that

$$y + \rho^+ - \rho^- = z,\ \ 0 \le \rho^- \perp x - a \ge 0,\ \ \ 0 \le \rho^+ \perp b - x \ge 0 \quad (9.69)$$

Given z, show that finding (x, y, ρ^-, ρ^+) that solves (9.69) is an LCP.

2 Show that it is equivalent to the following mixed integer constraint: $(x, y, z) \in \mathbb{R}^3$ satisfies (9.68) if and only if there exist binary variables $(\underline{z}, \overline{z})$ such that

$$\begin{aligned} a \le x \le b, \qquad & \underline{z}, \overline{z} \in \{0, 1\} & & \\ x - a \le M\underline{z}, \qquad & z - y \le M\underline{z}, & \qquad & y - z \le M(1 - \underline{z}) \\ b - x \le M\overline{z}, \qquad & y - z \le M\overline{z}, & \qquad & z - y \le M(1 - \overline{z}) \end{aligned}$$

where M is a sufficiently large constant.

Exercise 9.16 (LCP for quadratic program)**.** 1 Consider the quadratic optimization:

$$\min_{x \in \mathbb{R}^n} \ \frac{1}{2} x^\mathsf{T} Q x + c^\mathsf{T} x \ \text{ s.t. } \ Ax \le b,\ x \ge 0 \qquad (9.70a)$$

Show that solving the associated KKT condition is a LCP(M, q) with

$$M := \begin{bmatrix} Q^\mathsf{T} & A^\mathsf{T} \\ -A & 0 \end{bmatrix}, \qquad q := \begin{bmatrix} c \\ b \end{bmatrix} \qquad (9.70b)$$

2 Consider the quadratic optimization without the nonnegativity constraint on x:

$$\min_{x \in \mathbb{R}^n} \ \frac{1}{2} x^\mathsf{T} Q x + c^\mathsf{T} x \ \text{ s.t. } \ Ax \le b \qquad (9.70c)$$

If Q is positive definite show that solving the associated KKT condition is equivalent to the following LCP:

$$\lambda \ge 0,\ \ M\lambda + q \ge 0,\ \ \lambda^\mathsf{T}(M\lambda + q) = 0$$

Determine M and q.

Exercise 9.17 (LCP)**.** Suppose $A, B \in \mathbb{R}^{n \times n}$ are square matrices and $a, b \in \mathbb{R}^n$. Consider the problem of finding z such that

$$0 \le Az + a \perp Bz + b \ge 0$$

Show that this is equivalent to a LCP if A is nonsingular.

Exercise 9.18 (Linear complementarity problem)**.** Let

$$M := \begin{bmatrix} 1 & 1 \\ -1 & 0 \end{bmatrix}, \quad q := \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

Solve the LCP(M,q): find $x := [x_1\ x_2]^\mathsf{T}$ such that

$$x \geq 0, \quad Mx+q \geq 0, \quad x^\mathsf{T}(Mx+q) = 0$$

Note that there exists a unique solution even though M is neither positive definite nor symmetric.

Exercise 9.19 (McCormick envelop of $w = xy$)**.** For the bilinear constraint

$$w = xy, \qquad \underline{x} \leq x \leq \overline{x}, \qquad \underline{y} \leq y \leq \overline{y}, \qquad w,x,y \in \mathbb{R}$$

derive its McCormick envelops. (Hint: For (9.58a), if we let $a := x - \underline{x}$ and $b := y - \underline{y}$ or $a := \overline{x} - x$ and $b := \overline{y} - y$ then $ab \geq 0$. Similarly for the lower bounds.)

10 Semidefinite relaxations: BIM

Chapter 9 formulates OPF as a nonconvex quadratically constrained quadratic program (QCQP) and shows that it is NP-hard in general. There are three common approaches to deal with nonconvexity. First, one can solve a linear approximation of the original nonconvex problem. For instance DC OPF is a linear program approximation that is widely used for electricity market operations (see Chapter 6.4). Second, one can apply local algorithms such as Newton-Raphson or interior-point methods to compute a local solution. Some of these algorithms are studied in Chapter 8.5, but because the problem is nonconvex, the optimality conditions of Chapter 8.3 for convex problems are generally not applicable. Theorem 9.2 of Chapter 9.4 provides a Lyapunov-like condition that guarantees that if an algorithm does produce a local optimum, it will be a global optimum. The condition also ensures that convex relaxations of OPF will be exact and therefore a third approach is to solve a convex relaxation for a global solution, to which the optimality conditions of Chapter 8.3 do apply. In this and the next chapters we study a particular type convex relaxation, called semidefinite relaxation, of OPF.

There is a rich theory and extensive empirical experiences in applying semidefinite relaxation to many engineering problems. Besides being a method for seeking a global solution, a semidefinite relaxation allows us to check if a feasible solution produced by a local algorithm is globally optimal. If it is not, the solution of a relaxation provides a lower bound on the minimum cost and hence a bound on how far any feasible solution is from optimality. Unlike approximations, if a relaxed problem is infeasible, it is a certificate that the original OPF is infeasible.

In Chapter 10.1 we define semidefinite relaxation of QCQP in general and explain how to use the concept of partial matrices and their psd rank-1 completion to reduce the computational complexity of semidefinite relaxation for large sparse networks. In Chapter 10.2 we apply these results to write the single-phase OPF in terms of partial matrices to reveal structures that enable exact relaxations (we will use the QCQP formulation of OPF studied in Chapter 9.1.3 for the bus injection model). In Chapters 10.3 and 10.4 we study two sufficient conditions for exact relaxations of OPF on single-phase radial networks. We study semidefinite relaxations in the branch flow model in Chapter 11. The sufficient conditions in this and the next chapter complement the exactness condition of Chapter 9.4 (see Lemma 9.4).

10.1 Semidefinite relaxations of QCQP

OPF is formulated in (9.16) as a standard homogeneous QCQP. The computational difficulty arises from the nonconvex feasible set of OPF. Informally one can regard a relaxation of OPF as minimizing the same cost function over a convex superset (though in a lifted space). Different choices of convex supersets lead to different relaxations, but they all provide lower bounds to OPF. If an optimal solution of a relaxation happens to lie in the feasible set of the original OPF problem, then it is optimal for the original OPF. In this case we say the relaxation is exact. In this section we describe three types of semidefinite relaxation of OPF and explain equivalence relations among them.

10.1.1 SDP relaxation

Since these methods are not restricted to OPF, we will discuss them using the general QCQP formulation (9.10), reproduced here:

$$C^{\text{opt}} := \min_{x\in\mathbb{C}^n} \quad x^{\mathsf{H}} C_0 x \tag{10.1a}$$

$$\text{s.t.} \quad x^{\mathsf{H}} C_l x \le b_l, \quad l = 1,\dots,L \tag{10.1b}$$

Using $x^{\mathsf{H}} C_l x = \operatorname{tr}\left(C_l x x^{\mathsf{H}}\right)$ we can rewrite (10.1) as

$$\begin{aligned} \min_{X\in\mathbb{S}^n, x\in\mathbb{C}^n} \quad & \operatorname{tr}(C_0 X) \\ \text{s.t.} \quad & \operatorname{tr}(C_l X) \le b_l, \quad l = 1,\dots,L \\ & X = x x^{\mathsf{H}} \end{aligned}$$

Any positive semidefinite (psd) rank-1 matrix $X \in \mathbb{S}_+^{n\times n}$ has a spectral decomposition $X = xx^{\mathsf{H}}$ for some $x \in \mathbb{C}^n$; see Chapter A.6. The factor x is unique *up to a rotation*, i.e., x satisfies $X = xx^{\mathsf{H}}$ if and only if $x e^{j\theta}$ does for any $\theta \in \mathbb{R}$. Hence (10.1) is equivalent to the following problem where the optimization is over the set $\mathbb{S}^n$ of Hermitian matrices X:

$$\min_{X\in\mathbb{S}^n} \quad \operatorname{tr}(C_0 X) \tag{10.2a}$$

$$\text{s.t.} \quad \operatorname{tr}(C_l X) \le b_l, \quad l = 1,\dots,L \tag{10.2b}$$

$$X \succeq 0, \quad \operatorname{rank}(X) = 1 \tag{10.2c}$$

Recall that $\operatorname{tr}(C_l X) = \sum_{j,k} [C_l]_{jk} X_{kj} = \sum_{j,k} [C_l]_{jk} X_{jk}^{\mathsf{H}}$ where the second equality follows when X is Hermitian. While the objective function and the constraints in (10.1) are quadratic in x, they are linear in X in (10.2a)(10.2b). The constraint $X \succeq 0$ in (10.2c) is convex ($\mathbb{S}_+^n$ is a convex cone; see Chapter 8.2.2). The rank constraint in (10.2c) is the only nonconvex constraint. These two problems are equivalent in the sense that, given a feasible (or optimal) solution x to QCQP (10.1), there is an $X := xx^{\mathsf{H}}$ that is feasible (or optimal) to the semidefinite program (10.2). Conversely, given an X that is feasible (or optimal) to (10.2), a solution x to (10.1) can be recovered through rank-1

factorization $X = xx^{\mathsf{H}}$. It is in this sense that we also say that the feasible sets of (10.1) and (10.2) are equivalent. This is referred to as *lifting* the original QCQP problem from n dimensional space $\mathbb{C}^n$ to the higher-dimensional space of $n \times n$ Hermitian matrices.

Removing the rank constraint (10.2c) results in a semidefinite program (SDP):

$$\begin{aligned} \min_{X \in \mathbb{S}^n} \quad & \operatorname{tr}(C_0 X) && \text{(10.3a)} \\ \text{s.t.} \quad & \operatorname{tr}(C_l X) \;\le\; b_l, \quad l = 1, \ldots, L && \text{(10.3b)} \\ & X \;\succeq\; 0 && \text{(10.3c)} \end{aligned}$$

which is a convex problem. (Strong duality and KKT condition of semidefinite program is studied in Chapter 8.4.5.) We call (10.3) a *semidefinite relaxation* or an *SDP relaxation* of QCQP (10.1) because the feasible set of the equivalent problem (10.2) is a subset of the feasible set of SDP (10.3). A strategy for solving QCQP (10.1) is to solve SDP (10.3) for an optimal matrix X^{opt} and check its rank. If $\operatorname{rank}\left(X^{\text{opt}}\right) = 1$ then X^{opt} is feasible and hence optimal for (10.2) as well and an optimal solution x^{opt} of QCQP (10.1) can be recovered from X^{opt} through spectral decomposition $X^{\text{opt}} = x^{\text{opt}}(x^{\text{opt}})^{\mathsf{H}}$. If $\operatorname{rank} X^{\text{opt}} > 1$ then, in general, no feasible solution of QCQP can be directly obtained from X^{opt} but the optimal objective value of SDP provides a lower bound on that of QCQP.

10.1.2 Partial matrices and rank-1 completion

Even though the relaxation (10.3) is a convex problem computing its solution can still be challenging if the problem size n is large. If the underlying network is sparse, much more efficient relaxations can be used. To develop these ideas precisely, the key is to study the feasible sets of QCQP and its relaxations.

We start with the concept of partial matrices and their completions. An instance of QCQP (10.1) is specified by a set of matrices and scalars $(C_0, C_l, b_l, l = 1, \ldots, L)$. We assume the matrices $C_l, l = 0, 1, \ldots, L$, are Hermitian so that $x^{\mathsf{H}} C_l x$ are real. They define an underlying undirected graph $F := (N, E)$ with n nodes and m edges where distinct nodes j and k are adjacent (i.e., $(j,k) \in E$) if and only if there exists an $l \in \{0, 1, \ldots, L\}$ such that $[C_l]_{jk} = [C_l]_{kj}^{\mathsf{H}} \neq 0$. Assume without loss of generality that the graph F is connected (otherwise restrict ourselves to each connected component). For any $x \in \mathbb{C}^n$ note that the quadratic form $x^{\mathsf{H}} C_l x$ depends on $|x_j|^2$ and on $x_j^{\mathsf{H}} x_k$ if and only if $(j,k) \in E$ is a link in F, i.e., if and only if there exists an l such that the coefficient of $x_j^{\mathsf{H}} x_k$ is nonzero. Indeed

$$x^{\mathsf{H}} C_l x = \sum_{j,k} [C_l]_{jk}\, x_j^{\mathsf{H}} x_k \;=\; \sum_j [C_l]_{jj}\, |x_j|^2 \;+\; 2 \sum_{\substack{j<k \\ (j,k)\in E}} \operatorname{Re}\left([C_l]_{jk}\, x_j^{\mathsf{H}} x_k \right)$$

where the last equality follows from $[C_l]_{kj} x_k^{\mathsf{H}} x_j = [C_l]_{jk}^{\mathsf{H}} x_k^{\mathsf{H}} x_j = \left([C_l]_{jk}\, x_j^{\mathsf{H}} x_k \right)^{\mathsf{H}}$ since

C_l is Hermitian. Hence $x_j^{\mathsf{H}} x_k$ is not constrained by $x^{\mathsf{H}} C_l x \le b_l$ if $(j,k) \notin E$ for any l, in which case X_{jk} of the lifted variable X is not constrained by $\mathrm{tr}(C_l X) \le b_l$ for any l. This can be used to relax the psd and rank-1 constraints on the entire matrix X using the concept of partial matrices, greatly simplifying computation when the underlying graph F of the QCQP is sparse.

Given a graph $F := (N,E)$, a *partial matrix* X_F defined on F is a set of $2m+n$ complex numbers:

$$X_F := \left\{ [X_F]_{jj}, [X_F]_{jk}, [X_F]_{kj} \ : \ \text{nodes } j \in N \text{ and links } (j,k) \in E \right\}$$

X_F can be interpreted as a matrix with entries partially specified by these complex numbers. The (j,k)th entry of X_F that does not correspond to an edge in F is not specified. If F is a complete graph (in which there is an edge between every pair of vertices) then X_F is a fully specified $n \times n$ matrix. A *completion* X of X_F is any fully specified $n \times n$ matrix that agrees with X_F on graph F, i.e.,

$$[X]_{jj} \ = \ [X_F]_{jj}, \quad [X]_{jk} \ = \ [X_F]_{jk}, \quad [X]_{kj} \ = \ [X_F]_{kj}, \quad j \in N, \ (j,k) \in E$$

Given an $n \times n$ matrix X we use X_F to denote the *submatrix of X on F,* i.e., the partial matrix consisting of the entries of X defined on graph F. If q is a clique (a fully connected subgraph) of F then let $X_F(q)$ denote the fully-specified principal submatrix of X_F defined on q, i.e., if the clique q has k nodes then $X_F(q)$ is a $k \times k$ matrix and, for every node j and link (j,k) in q,

$$[X_F(q)]_{jj} \ := \ [X_F]_{jj}, \qquad [X_F(q)]_{jk} \ := \ [X_F]_{jk}, \qquad [X_F(q)]_{kj} \ := \ [X_F]_{kj}$$

We extend the definitions of Hermitian, psd, rank-1, and the trace operation for matrices to partial matrices.

Definition 10.1 (Partial matrix X_F)**.** Let X_F be a partial matrix on a graph $F := (N,E)$.

1 X_F is *Hermitian*, denoted by $X_F = X_F^{\mathsf{H}}$, if $[X_F]_{kj} = [X_F]_{jk}^{\mathsf{H}}$ for all $(j,k) \in E$.
2 X_F is *positive semidefinite* (psd), denoted by $X_F \succeq 0$, if X_F is Hermitian and the principal submatrices $X_F(q)$ are psd for all cliques q of F.
3 X_F is *rank-1*, denoted by $\mathrm{rank}(X_F) = 1$, if the principal submatrices $X_F(q)$ are rank-1 for all cliques q of F.
4 X_F is 2×2 *psd* if, for all edges $(j,k) \in F$, the 2×2 principal submatrices

$$X_F(j,k) := \begin{bmatrix} [X_F]_{jj} & [X_F]_{jk} \\ [X_F]_{kj} & [X_F]_{kk} \end{bmatrix}$$

are psd (and necessarily Hermitian).
5 X_F is 2×2 *rank-1* if, for all edges $(j,k) \in F$, the 2×2 principal submatrices $X_F(j,k)$ are rank-1.
6 The trace operation on X_F is defined as

$$\mathrm{tr}\,(C_l X_F) \ := \ \sum_{j \in N} [C_l]_{jj} \, [X_F]_{jj} \ + \ \sum_{\substack{j<k \\ (j,k) \in E}} \left([C_l]_{jk} \, [X_F]_{kj} + [C_l]_{kj} \, [X_F]_{jk} \right)$$

□

The condition $X_F(j,k) \succeq 0$ is equivalent to: the matrix $X_F(j,k)$ is Hermitian, i.e., $X_F(j,k) = X_F(j,k)^{\mathsf{H}}$, and

$$[X_F]_{jj} \geq 0, \qquad [X_F]_{kk} \geq 0, \qquad [X_X]_{jj}[X_X]_{kk} \geq \left|[X_F]_{jk}\right|^2$$

This is a rotated second-order cone studied in Chapter 8.2.1 (see (8.17)). The condition $\operatorname{rank}(X_F(j,k)) = 1$ is equivalent to:

$$[X_F]_{jj}\,[X_F]_{kk} = \left|[X_F]_{jk}\right|^2 \;>\; 0$$

If both C_l and X_F are Hermitian then $[C_l]_{kj}[X_F]_{jk} = \left([C_l]_{jk}[X_F]_{kj}\right)^{\mathsf{H}}$ and hence

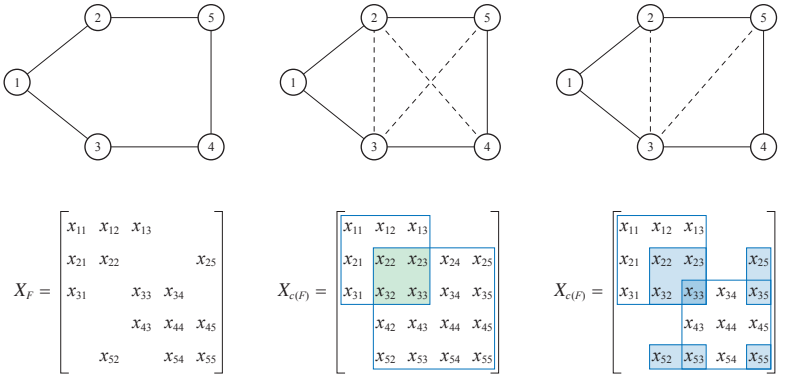
$$\operatorname{tr}(C_l X_F) = \sum_{j\in N} [C_l]_{jj}\,[X_F]_{jj} + 2 \sum_{\substack{j<k \\ (j,k)\in E}} \operatorname{Re}\left([C_l]_{jk}\,[X_F]_{kj}\right)$$

is a real scalar.

We call F a *chordal graph* if either F has no cycle or all its minimal cycles (ones without chords) are of length three. A *chordal extension* $c(F)$ of F is a chordal graph that contains F, i.e., $c(F)$ has the same vertex set as F but an edge set that is a superset of F's edge set. In that case we call the partial matrix $X_{c(F)}$ a *chordal extension* of the partial matrix X_F. Every graph F has a chordal extension, generally nonunique. In particular a complete supergraph of F is a trivial chordal extension of F. Chordal graphs are important for us because of the result [112, Theorem 7] that every psd partial matrix has a psd completion if and only if the underlying graph is chordal. When a positive definite completion exists, there is a *unique* positive definite completion, in the class of all positive definite completions, whose determinant is maximal. We extend this result to rank-1 partial matrices after presenting an example.

Example 10.1 (Partial matrices and chordal extensions)**.** Consider the graph F and the partial matrix X_F in Figure 10.1(a). X_F is Hermitian if $x_{jk} = x_{kj}^{\mathsf{H}}$. The only cliques in F consist of two nodes that are adjacent, and hence X_F is psd if it is 2×2 psd and X_F is rank-1 if it is 2×2 rank-1. X_F is not chordal as it contains a cycle of length greater than 3.

Figure 10.1(b) and (c) depict two chordal extensions $c(F)$ of F and the partial matrices $X_{c(F)}$ defined on these chordal extensions. The chordal extension in Figure 10.1(b) has 2 maximal cliques, $q_1 := (1,2,3)$ and $q_2 := (2,3,4,5)$. These cliques share two nodes, 2 and 3. The (fully specified) submatrices $X_{c(F)}(q_1)$ and $X_{c(F)}(q_2)$ defined on the cliques q_1 and q_2 respectively are outlined in the figure with overlapping entries shaded in green. The chordal extension in Figure 10.1(c) has 3 maximal cliques whose (fully specified) submatrices are outlined. The clique $q_2 := (2,3,5)$ overlaps with the other two cliques and the overlapping entries in $X_{c(F)}(q_2)$ are shaded in blue. (The shared nodes between maximal cliques introduce complications in formulating semidefinite relaxation based on chordal extensions; see Chapter 10.1.6.) □



(a) Partial matrix X_F (b) Chordal extension $X_{c(F)}$ (c) Chordal extension $X_{c(F)}$

Figure 10.1 Example 10.1: The overlapping maximal cliques of chordal extensions $X_{c(F)}$ are in shades.

Consider the following conditions on a $n \times n$ matrix X and partial matrices $X_{c(F)}$ and X_F associated with a given graph F:

$$X \succeq 0, \operatorname{rank}(X) = 1 \tag{10.4a}$$

$$X_{c(F)} \succeq 0, \operatorname{rank}(X_{c(F)}) = 1 \tag{10.4b}$$

$$X_F(j,k) \succeq 0, \operatorname{rank}(X_F(j,k)) = 1, \quad (j,k) \in E \tag{10.4c}$$

We say that a partial matrix X_F satisfies the *cycle condition* if for every cycle c in F

$$\sum_{(j,k)\in c} \angle [X_F]_{jk} = 0 \mod 2\pi \tag{10.5}$$

where $x = \phi \mod 2\pi$ means $x = \phi + 2k\pi$ for some integer k. For instance if $\angle [X_F]_{jk}$ represent the voltage phase differences across lines (j,k) then the cycle condition imposes that they sum to zero (mod 2π) around any cycle c. The next theorem, proved in [113, Theorem 3] and [33], implies that X_F has a psd rank-1 completion X if and only if X_F has a chordal extension $X_{c(F)}$ that is psd rank-1, if and only if X_F is 2×2 psd rank-1 on F and satisfies the cycle condition (10.5). [1] All proofs in this section are deferred to Chapter 10.1.8

Theorem 10.1 (Rank-1 characterization)**.** Fix a connected graph $F := (N, E)$ with $n := |N|$ nodes. Consider any chordal extension $c(F)$ of F. Suppose $X_{jj} > 0$, $\left[X_{c(F)}\right]_{jj} > 0$ and $[X_F]_{jj} > 0$, $j \in N$, for the matrix X and submatrices X_F and $X_{c(F)}$ below. Then

(1) Given a $n \times n$ matrix X that satisfies (10.4a), its submatrix $X_{c(F)}$ satisfies (10.4b).
(2) Given a partial matrix $X_{c(F)}$ that satisfies (10.4b), its submatrix X_F satisfies (10.4c) and the cycle condition (10.5).

[1] The theorem also holds with psd replaced by negative semidefinite.

(3) Given a partial matrix X_F that satisfies (10.4c) and the cycle condition (10.5), there is a completion X of X_F that satisfies (10.4a). Moreover the completion X is unique.

Informally Theorem 10.1 says that (10.4a) is equivalent to (10.4b) which is equivalent to (10.4c)(10.5). It implies in particular that, for a chordal graph, X is psd rank-1 if and only if the principal submatrix $X(q)$ of X is psd rank-1 for every maximal clique q of the graph. It characterizes a property of the full matrix X (that X is psd and rank-1) in terms of its submatrices $X_{c(F)}$ and X_F. This is important because the submatrices are typically much smaller than X for large sparse networks and much easier to compute. We discuss how to construct a chordal extension $c(F)$ of F and formulate $X_{c(F)}$ in Chapter 10.1.6. Theorem 10.1 thus allows us to solve smaller problems in terms of partial matrices as we now explain.

10.1.3 Feasible sets

To develop semidefinite relaxations of QCQP we start by studying their feasible sets. Fix C_l, $l = 0, 1, \dots L$, and its underlying graph F. Define the feasible set of the QCQP (10.1) as:

$$\mathbb{V} := \{x \in \mathbb{C}^n \,|\, x^{\mathsf{H}} C_l x \;\; \leq \;\; b_l,\; l = 1, \dots, L\} \tag{10.6}$$

Given an $x \in \mathbb{V}$, it defines a unique (up to a rotation) psd rank-1 matrix $X := xx^{\mathsf{H}}$ and therefore a unique psd rank-1 partial matrix X_F that satisfies $\mathrm{tr}(C_l X_F) \leq b_l$. The converse is not always true: given a partial matrix X_F that is psd rank-1 and satisfies $\mathrm{tr}(C_l X_F) \leq b_l$, it is not always possible to recover an x in $\mathbb{V}$. This is possible if and only if X_F has a psd rank-1 completion X that satisfies $\mathrm{tr}(C_l X) \leq b_l$. We now characterize the set of partial matrices from which $x \in \mathbb{V}$ can be recovered.

Define the set of Hermitian matrices:

$$\mathbb{X} := \{\; X \in \mathbb{S}^n \;|\; X \text{ satisfies } \mathrm{tr}(C_l X) \leq b_l,\; l = 1, \dots, L, \;\; (10.4a) \;\} \tag{10.7a}$$

i.e., $X \in \mathbb{X}$ satisfies $\mathrm{tr}(C_l X) \leq b_l$ for all l and (10.4a). Fix a connected graph F. Fix any chordal extension $c(F)$ of F and define the set of Hermitian partial matrices $X_{c(F)}$:

$$\mathbb{X}_{c(F)} := \{\; X_{c(F)} \;|\; X_{c(F)} \text{ satisfies } \mathrm{tr}\left(C_l X_{c(F)}\right) \leq b_l,\; l = 1, \dots, L, \;\; (10.4b) \;\} \tag{10.7b}$$

Finally define the set of Hermitian partial matrices X_F:

$$\mathbb{X}_F := \{\; X_F \;|X_F \text{ satisfies } \mathrm{tr}(C_l X_F) \leq b_l,\; l = 1, \dots, L, \;\; (10.4c)(10.5) \;\} \tag{10.7c}$$

Note that the definition of psd for partial matrices implies that $X_{c(F)}$ and X_F are Hermitian partial matrices (see Definition 10.1).

Theorem 10.1 implies that given a partial matrix $X_{c(F)} \in \mathbb{X}_{c(F)}$ or a partial matrix $X_F \in \mathbb{X}_F$ there is a psd rank-1 completion $X \in \mathbb{X}$. Moreover the completion X is unique.

Corollary 10.2 (Uniqueness of rank-1 completion)**.** Fix a connected graph F. Given a partial matrix $X_{c(F)} \in \mathbb{X}_{c(F)}$ or $X_F \in \mathbb{X}_F$ there is a unique psd rank-1 completion $X \in \mathbb{X}$.

The corollary implies that, given any Hermitian partial matrix $X_F \in \mathbb{X}_F$, the set of *all* completions of X_F consists of a single psd rank-1 matrix and infinitely many indefinite or non-rank-1 matrices.

We say two sets A and B are *equivalent*, denoted $A \equiv B$, if there is a bijection between them. Even though $\mathbb{X}, \mathbb{X}_{c(F)}, \mathbb{X}_F$ are different kinds of spaces, Theorem 10.1 and Corollary 10.2 imply that they are all equivalent to the feasible set of QCQP (10.1) once an arbitrary reference angle is fixed, e.g., $\angle x_1 := 0$.

Theorem 10.3 (Equivalence)**.** $\mathbb{V} \equiv \mathbb{X} \equiv \mathbb{X}_{c(F)} \equiv \mathbb{X}_F$.

Since the cost function $x^{\mathsf{H}} C_0 x$ of (10.1) depends on X only through the partial matrix X_F, Theorem 10.3 suggests three problems that are equivalent to QCQP (10.1): for $\hat{\mathbb{X}} \in \{\mathbb{X}, \mathbb{X}_{c(F)}, \mathbb{X}_F\}$,

$$\min_X C(X_F) \;\text{ subject to } X \in \hat{\mathbb{X}} \tag{10.8}$$

Specifically, given an optimal solution X^{opt} in $\mathbb{X}$, it can be decomposed into $X^{\mathrm{opt}} = x^{\mathrm{opt}}(x^{\mathrm{opt}})^{\mathsf{H}}$ where x^{opt} is unique up to an arbitrary reference angle. Then x^{opt} is in $\mathbb{V}$ and an optimal solution of QCQP (10.1). Alternatively given an optimal solution $X_F^{\mathrm{opt}} \in \mathbb{X}_F$ or $X_{c(F)}^{\mathrm{opt}} \in \mathbb{X}_{c(F)}$, Corollary 10.2 guarantees that it has a unique psd rank-1 completion X^{opt} in $\mathbb{X}$ from which an optimal $x^{\mathrm{opt}} \in \mathbb{V}$ can be recovered. This suggests solving the QCQP (10.1) by computing X_F^{opt} or $X_{c(F)}^{\mathrm{opt}}$ instead of X^{opt} because both of them are typically much smaller in size than X^{opt} for a large sparse network. Indeed the number of complex variables in a Hermitian X is $n(n+1)/2$ while the number of complex variables in X_F is only $n+|E|$, which is much smaller if F is large but sparse. Given a partial matrix $X_F \in \mathbb{X}_F$ (or $X_{c(F)} \in \mathbb{X}_{c(F)}$), however, there is a more direct construction of a feasible solution $x \in \mathbb{V}$ of QCQP than through its completion (see Chapter 10.1.4).

Remark 10.1 (Graph $\hat{F}$ underlying QCQP)**.** Note that the feasible sets $\mathbb{V}, \mathbb{X}, \mathbb{X}_{c(F)}, \mathbb{X}_F$ defined in (10.6) (10.7) depend only on the constraint matrices C_l, $l = 1, \ldots, L$, but not on the cost matrix C_0. Equivalence among these sets will therefore hold if we replace F in Theorem 10.1, Corollary 10.2 and Theorem 10.3 with a subgraph $\hat{F}$ that is induced by C_l only for $l \geq 1$, i.e., two nodes j and k in $\hat{F}$ are adjacent if and only if $[C_l]_{jk} \neq 0$ for some $l \in \{1, \ldots, L\}$.

The matrix F is needed for the proper definition of cost function. For the optimization problems in (10.8) to be equivalent, we need to compute the partial matrices X_F and $X_{c(F)}$. The partial matrices $X_{\hat{F}}$ will have missing terms $[X_{\hat{F}}]_{jk}$ in the cost function if (j,k) is in F but not in $\hat{F}$, i.e., if $[C_0]_{jk} \neq 0$ but $[C_l]_{jk} = 0$ for all $l \geq 1$. Similarly for $X_{c(\hat{F})}$. □

10.1.4 Semidefinite relaxations and solution recovery

Hence solving QCQP (10.1) is equivalent to solving (10.8) over any of $\mathbb{X}, \mathbb{X}_{c(F)}, \mathbb{X}_F$ for an appropriate matrix variable. The difficulty with solving (10.8) is that the feasible sets $\mathbb{X}$, $\mathbb{X}_{c(F)}$, and $\mathbb{X}_F$ are still nonconvex due to the rank-1 constraint and the cycle condition (10.5). Their removal leads to three types of semidefinite relaxations of QCQP (10.1).

Semidefinite relaxations.

Relax $\mathbb{X}$, $\mathbb{X}_{c(F)}$ and $\mathbb{X}_F$ to the following convex supersets:

$$\begin{aligned}
\mathbb{X}^+ &:= \{X \in \mathbb{S}^n \mid X_F \text{ satisfies } \operatorname{tr}(C_l X) \le b_l,\ l = 1,\dots,L,\ \ X \succeq 0\} \\
\mathbb{X}^+_{c(F)} &:= \{X_{c(F)} \mid X_F \text{ satisfies } \operatorname{tr}\left(C_l X_{c(F)}\right) \le b_l,\ l = 1,\dots,L,\ \ X_{c(F)} \succeq 0\} \\
\mathbb{X}^+_F &:= \{X_F \mid X_F \text{ satisfies } \operatorname{tr}\left(C_l X_F\right) \le b_l,\ l = 1,\dots,L,\ \ X_F(j,k) \ge 0,\ (j,k) \in E\}
\end{aligned}$$

These feasible sets are defined for different (partial) matrices and differ in the definition of psd. Remark 10.1 applies to these relaxed feasible sets regarding the underlying graph and the corresponding partial matrices. The following problems are semidefinite relaxations of QCQP (10.1) with different sizes and tightness:

QCQP-sdp:

$$C^{\text{sdp}} := \min_X C(X_F) \text{ subject to } X \in \mathbb{X}^+ \tag{10.9a}$$

QCQP-ch:

$$C^{\text{ch}} := \min_{X_{c(F)}} C(X_F) \text{ subject to } X_{c(F)} \in \mathbb{X}^+_{c(F)} \tag{10.9b}$$

QCQP-socp:

$$C^{\text{socp}} := \min_{X_F} C(X_F) \text{ subject to } X_F \in \mathbb{X}^+_F \tag{10.9c}$$

We call (10.9a) a *SDP relaxation*, (10.9b) a *chordal relaxation*, and (10.9c) a *SOCP relaxation*. In Chapter 10.1.6 we describe how to construct the set of constraints $X_{c(F)} \succeq 0$ in $\mathbb{X}^+_{c(F)}$ and show that chordal relaxation is equivalent to a semidefinite program (and similarly for SOCP relaxation).

Solution recovery.

When the semidefinite relaxations OPF-sdp, OPF-ch, OPF-socp are exact, i.e., if their optimal solutions $X^{\text{sdp}}, X^{\text{ch}}_{c(F)}, X^{\text{socp}}_F$ happen to lie in $\mathbb{X}, \mathbb{X}_{c(F)}, \mathbb{X}_F$ respectively, then an optimal solution $x^{\text{opt}} \in \mathbb{V}$ of the original QCQP can be recovered from these solutions. Indeed the recovery method works not just for an optimal solution, but any feasible solution that lies in $\mathbb{X}$, $\mathbb{X}_{c(F)}$ or $\mathbb{X}_F$. Moreover, given an $X \in \mathbb{X}$ or an $X_{c(F)} \in \mathbb{X}_{c(F)}$, the construction of x depends on X or $X_{c(F)}$ only through their submatrix X_F. We hence describe a method for recovering an $x \in \mathbb{V}$ from an X_F, which may be a partial

matrix in $\mathbb{X}_F$ or the submatrix of a (partial) matrix in $\mathbb{X}$ or $\mathbb{X}_{c(F)}$. The solution x is unique if F is connected and, say, $\angle x_1$ is fixed.

Take an *arbitrary* spanning tree of F rooted at bus 1 with orientation where lines pointing away from bus 1. Let P_j denote the unique path from bus 1 to bus j in the spanning tree. Set $|x_1| := \sqrt{[X_F]_{11}}$ and $\angle x_1$ to an arbitrary value. For $j = 2, \ldots, n$,

$$|x_j| := \sqrt{[X_F]_{jj}}, \qquad \angle x_j := \angle V_1 - \sum_{(i,k)\in \mathsf{P}_j} \angle [X_F]_{ik} \tag{10.10}$$

Then, on link (j,k), $\angle x_j - \angle x_k = \angle [X_F]_{jk}$ and $[X_F]_{jk} = x_j x_k^{\mathsf{H}}$ since X_F is 2×2 psd rank-1. It can be checked that x is in the feasible set $\mathbb{V}$ of QCQP, i.e., $x^{\mathsf{H}} C_l x \le b_l$, $l = 1, \ldots, L$ (Exercise 10.1). The cycle condition (10.5) ensures that the angle calculation (10.10) gives the same result for any spanning tree.

This method for recovering x from X_F is generally more efficient than computing the psd rank-1 completion X of X_F and factorizing X, as suggested in Theorem 10.3, and is used in the proof of Theorem 10.1 (see Chapter 10.1.8). It is equivalent to the method (5.12c) of Chapter 5.1.2 for recovering voltage angles in the branch flow model for radial networks, with $\beta_{jk} = [X_F]_{jk}$.

10.1.5 Tightness of relaxations

Recall that $\mathbb{V} \equiv \mathbb{X} \equiv X_{c(F)} \equiv X_F$ (Theorem 10.3). Since $\mathbb{X} \subseteq \mathbb{X}^+$, $\mathbb{X}_{c(F)} \subseteq \mathbb{X}^+_{c(F)}$, $\mathbb{X}_F \subseteq \mathbb{X}^+_F$, the relaxations OPF-sdp, OPF-ch, OPF-socp all provide lower bounds on OPF (9.9). OPF-socp is the simplest computationally. OPF-ch usually requires heavier computation than OPF-socp but much lighter than OPF-sdp for large sparse networks (even though OPF-ch can be as complex as OPF-sdp in the worse case [114, 115]). The relative tightness of the relaxations depends on the network topology. For a general network that may contain cycles, OPF-ch is as tight a relaxation as OPF-sdp and they are strictly tighter than OPF-socp. For a tree (radial) network the hierarchy collapses and all three are equally tight. We now make this precise.

Consider the relaxed feasible sets $\mathbb{X}^+$, $\mathbb{X}^+_{c(F)}$ and $\mathbb{X}^+_F$. Consider two sets A and B and the corresponding cost functions $C_A : A \to \mathbb{R}$ and $C_B : B \to \mathbb{R}$. For instance $A := \mathbb{C}^n$, $B := \mathbb{S}^n$, $C_A(x) := x^{\mathsf{H}} C x$ and $C_B(X) := \mathrm{tr}\,(CX)$ for a given Hermitian matrix C. We say that A is an *effective subset* of B with respect to the cost functions C_A, C_B, denoted by $A \sqsubseteq B$, if, given any $a \in A$, there is a $b \in B$ that has the same cost $C_A(a) = C_B(b)$. We say A is *similar to* B with respect to the cost functions C_A, C_B, denoted by $A \simeq B$, if $A \sqsubseteq B$ and $B \sqsubseteq A$. Note that $A \equiv B$ implies $A \simeq B$ but the converse may not hold. Even though effective subset and similarity are defined with respect to some cost functions C_A, C_B, we often omit the cost functions when their existence is understood and unimportant for the discussion, and simply say A is an effective subset of B or A is similar to B.

The feasible set of QCQP (10.1) is an effective subset of the feasible sets of its relaxations; moreover these relaxations have similar feasible sets when the network is radial.

Theorem 10.4 (Tightness of relaxations)**.** 1 $\mathbb{V} \sqsubseteq \mathbb{X}^+ \simeq \mathbb{X}^+_{c(F)} \sqsubseteq \mathbb{X}^+_F$.
2 If F is a tree then $\mathbb{V} \sqsubseteq \mathbb{X}^+ \simeq \mathbb{X}^+_{c(F)} \simeq \mathbb{X}^+_F$.

The reason $\mathbb{X}^+_{c(F)}$ is similar, but not equivalent, to $\mathbb{X}^+$ is that psd completions of a psd submatrix $X \in \mathbb{X}^+_{c(F)}$ are generally nonunique. In contrast, the psd rank-1 completion of a psd rank-1 submatrix $X \in \mathbb{X}_{c(F)}$ is unique according to Corollary 10.2.

Let $C^{\text{opt}}, C^{\text{sdp}}, C^{\text{ch}}, C^{\text{socp}}$ be the optimal values of QCQP (10.1), QCQP-sdp (10.9a), QCQP-ch (10.9b), QCQP-socp (10.9c) respectively. Theorem 10.3 and Theorem 10.4 directly imply

Corollary 10.5. 1 $C^{\text{opt}} \geq C^{\text{sdp}} = C^{\text{ch}} \geq C^{\text{socp}}$.
2 If F is a tree then $C^{\text{opt}} \geq C^{\text{sdp}} = C^{\text{ch}} = C^{\text{socp}}$.

Remark 10.2 (Tightness)**.** Theorem 10.4 and Corollary 10.5 imply that for radial networks one should always solve QCQP-socp, not QCQP-sdp or QCQP-ch, since it is the tightest and the simplest relaxation of the three. For networks that contain cycles there is a tradeoff between QCQP-socp and QCQP-ch/QCQP-sdp: the latter is tighter but requires heavier computation. Between QCQP-ch and QCQP-sdp, QCQP-ch is preferable as they are equally tight but QCQP-ch is usually much faster to solve for large sparse networks. □

10.1.6 Chordal relaxation

Theorem 10.1 through Corollary 10.5 apply to *any* chordal extension $c(F)$ of F. The choice of $c(F)$ does not affect the optimal value of the chordal relaxation but determines its complexity. We now explain how to construct the set of constraints $X_{c(F)} \succeq 0$ in the definition of $\mathbb{X}^+_{c(F)}$ and show that chordal relaxation (10.9b) is equivalent to a semidefinite program. The method is applicable to SOCP relaxation (10.9c) as well (see Example 10.2).

The constraint $X_{c(F)} \succeq 0$ consists of multiple constraints that the (fully specified) principal submatrices $X_{c(F)}(q) \succeq 0$, one for each maximal clique q of $c(F)$. We will discuss the tradeoffs in choosing a chordal extension $c(F)$ of F later. Once a $c(F)$ is chosen the construction of $X_{c(F)} \succeq 0$ involves two steps:

1 List all the maximal cliques q_k of $c(F)$, $k = 1, \ldots, K$.
2 Use as relaxation variables appropriate Hermitian matrices X_k corresponding to q_k. Then $X_{c(F)} \succeq 0$ is a shorthand for: $X_k \succeq 0$ for $k = 1, \ldots, K$.

We elaborate on both steps. Computing all maximal cliques of a general graph is NP-hard. It can however be done efficiently for a chordal graph because a graph is chordal if and only if it has a perfect elimination ordering [116] and computing this ordering takes linear time in the number of nodes and edges [117]. Given a perfect elimination ordering all maximal cliques q_k can be enumerated and $X_F(q_k)$ constructed efficiently [114]. For most OPF applications the computation depends only on the topology of the power network, not on operational data, and therefore can be done offline.

Suppose the set of maximal cliques $\{q_k, k = 1, \ldots, K\}$ has been identified in which clique q_k consists of n_k nodes. It is tempting to simply use K matrix variables X_k each of size $n_k \times n_k$, require $X_k \succeq 0$ in the chordal relaxation (10.9b), and integrate the K optimal (fully specified) matrix solutions X_k^{opt} of (10.9b) into a single optimal partial matrix $X_{c(F)}^{\text{opt}}$. Unfortunately this approach fails if some of the maximal cliques q_k share nodes. In that case their X_k share entries and cannot be integrated as principal submatrices of an $n \times n$ matrix, as explained in Example 10.1. Therefore when maximal cliques of $c(F)$ share nodes, their corresponding matrices must be decoupled by introducing auxiliary variables and equality constraints on the auxiliarty variables. We now sketch this procedure using Example 10.1 (see [114, 115] for more details). It also illustrates the difficulty in choosing a good chordal extension $c(F)$.

Suppose we have chosen the chordal extension $c(F)$ in Figure 10.1(b) with two cliques $q_1 := (1,2,3)$ and $q_2 := (2,3,4,5)$ that share nodes 2 and 3. The (fully specified) matrices X_1 and X_2 defined on the cliques q_1 and q_2 respectively are outlined in Figure 10.1(b). They overlap in 4 entries and require 4 decoupling variables u_{jk}. To decouple these matrices, replace X_1 by the 3×3 matrix

$$X_1' := \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & u_{22} & u_{23} \\ x_{31} & u_{32} & u_{33} \end{bmatrix}$$

where the decoupling variables u_{jk} are constrained to be:

$$u_{jk} = x_{jk} \quad \text{for } j,k = 2,3 \tag{10.11a}$$

Then the psd constraints $X_{c(F)} \succeq 0$ in chordal relaxation (10.9b) is not $X_1 \succeq 0$ and $X_2 \succeq 0$, but

$$X_1' \succeq 0, \qquad X_2 \succeq 0 \tag{10.11b}$$

We can write the chordal relaxation as a SDP in standard form (10.3) by defining the 7×7 block-diagonal matrix

$$X' := \begin{bmatrix} X_1' & 0 \\ 0 & X_2 \end{bmatrix}$$

Then chordal relaxation (10.9b) is equivalent to:

$$\begin{aligned}
\min_{X' \in \mathbb{S}^7} \quad & \operatorname{tr}(C_0' X') && && (10.12a)\\
\text{s.t.} \quad & \operatorname{tr}(C_l' X') \le b_l, && l = 1,\ldots,L && (10.12b)\\
& \operatorname{tr}(C_r' X') = 0, && r = 1,2,3,4 && (10.12c)\\
& X' \succeq 0 && && (10.12d)
\end{aligned}$$

for appropriate C_l', $l = 0,\ldots,L$. The constraint $X' \succeq 0$ in (10.12d) is equivalent to the psd constraints (10.11b) on X_1' and X_2. The matrices C_r' in (10.12c) are chosen to enforce the linear decoupling constraints (10.11a). See Example 10.2 for an explicit construction of these matrices.

As the example illustrates, the choice of chordal extension $c(F)$ determines the number and sizes of matrices X_k associated with the maximal cliques as well as the number of decoupling variables and constraints. In our example, the full SDP computes a 5×5 matrix X for 25 variables (counting x_{jk} and $x_{kj} = \bar{x}_{jk}$ as two variables). Chordal relaxation defined by (10.11) computes a 3×3 matrix X_1' and a 4×4 matrix X_2 for 25 variables, plus 4 decoupling variables and (linear) constraints. If we have chosen the chordal extension $c(F)$ in Figure 10.1(c) with three cliques $q_1 := (1,2,3)$, $q_2 := (3,4,5)$, and $q_3 := (2,3,5)$, then chordal relaxation will involve three 3×3 matrices with 27 variables, plus 8 decoupling variables and constraints. (Despite these examples, chordal relaxation is typically much less computationally intensive than a full SDP for large sparse network.)

The optimal choice of chordal extension $c(F)$ that minimizes the complexity of QCQP-ch is NP-hard to compute. This difficulty is due to two conflicting factors in choosing a $c(F)$. On the one hand if $c(F)$ contains few cliques q then the submatrices $X_{c(F)}(q)$ tend to be large and expensive to compute (e.g. if $c(F)$ is the complete graph then there is a single clique, but $X_{c(F)} = X$ and QCQP-ch is identical to QCQP-sdp). On the other hand if $c(F)$ contains many small cliques q then there tends to be more overlap and chordal relaxation tends to require more decoupling variables and constraints. Hence choosing a good chordal extension $c(F)$ of F is important but nontrivial.

Example 10.2 (SOCP relaxation)**.** We apply the same method to construct SOCP relaxation (10.9c) on the graph in Figure 10.1(a). It has 5 links $(1,2)$, $(1,3)$, $(3,4)$, $(4,5)$, $(2,5)$. (In this example each link is a maximal clique but this fact is not important for SOCP relaxation, i.e., for a general network F we can choose an arbitrary spanning tree T_F and construct SOCP relaxation on T_F.) Every link (j,k) shares node j with a link (i,j) and node k with another link (k,l). We introduce 5 decoupling variables to decouple the five 2×2 variables:

$$X_{12} := \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}, \quad X_{13}' := \begin{bmatrix} u_{11} & x_{13} \\ x_{31} & u_{33} \end{bmatrix}, \quad X_{34}' := \begin{bmatrix} u_{33} & x_{34} \\ x_{43} & x_{44} \end{bmatrix}, \quad X_{45}' := \begin{bmatrix} u_{44} & x_{45} \\ x_{54} & x_{55} \end{bmatrix}, \quad X_{25}' := \begin{bmatrix} u_{22} & \\ x_{52} & \end{bmatrix}$$

with 5 decoupling constraints:

$$u_{11} = x_{11}, \quad u_{33} = x_{33}, \quad u_{44} = x_{44}, \quad u_{22} = x_{22}, \quad u_{55} = x_{55} \tag{10.13a}$$

Then the set of 2×2 psd constraints in $\mathbb{X}_F^+$ are:

$$X_{12} \succeq 0, \quad X'_{13} \succeq 0, \quad X'_{34} \succeq 0, \quad X'_{45} \succeq 0, \quad X'_{25} \succeq 0 \tag{10.13b}$$

We can convert this into a semidefinite program in standard form, i.e., we will construct the matrices C'_l in (10.12).

Define the 10×10 matrix

$$X' := \operatorname{diag}\left(X_{12}, X'_{13}, X'_{34}, X'_{45}, X'_{25}\right)$$

Then (10.13b) is equivalent to $X' \succeq 0$. To convert an original constraint $\operatorname{tr}(C_l X_F) \le b_l$ into $\operatorname{tr}(C'_l X') \le b_l$ we have (each c_{jk} may be zero or nonzero, but all blank entries are zero):

$$\operatorname{tr}(C_l X_F) \le b_l \quad \Leftrightarrow \quad \operatorname{tr}\begin{bmatrix} c_{11} & c_{12} & c_{13} & & \\ c_{21} & c_{22} & & & c_{25} \\ c_{31} & & c_{33} & c_{34} & \\ & & c_{43} & c_{44} & c_{45} \\ & c_{52} & & c_{54} & c_{55} \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} & x_{13} & & \\ x_{21} & x_{22} & & & x_{25} \\ x_{31} & & x_{33} & x_{34} & \\ & & x_{43} & x_{44} & x_{45} \\ & x_{52} & & x_{54} & x_{55} \end{bmatrix} \le b_l$$

To construct C'_l, define

$$C_{12} := \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}, \quad C'_{13} := \begin{bmatrix} 0 & c_{13} \\ c_{31} & c_{33} \end{bmatrix}, \quad C'_{34} := \begin{bmatrix} 0 & c_{34} \\ c_{43} & c_{44} \end{bmatrix}, \quad C'_{45} := \begin{bmatrix} 0 & c_{45} \\ c_{54} & c_{55} \end{bmatrix}, \quad C'_{25} := \begin{bmatrix} 0 \\ c_{52} \end{bmatrix}$$

i.e., C'_{jk} has the same pattern as X'_{jk} with entries corresponding to decoupling variables u_{jj} set to zero. Then

$$C'_l := \operatorname{diag}\left(C_{12}, C'_{13}, C'_{34}, C'_{45}, C'_{25}\right)$$

and

$$\operatorname{tr}(C_l X_F) \le b_l \quad \Longleftrightarrow \quad \operatorname{tr}(C'_l X') \le b_l$$

Finally to enforce the decoupling constraints (10.13a) define (e_j is the unit vector of size 10 with 1 in the jth place and 0 elsewhere)

$$\begin{aligned} C'_{11} &:= e_1 e_1^{\mathsf{T}} - e_3 e_3^{\mathsf{T}}, & C'_{33} &:= e_4 e_4^{\mathsf{T}} - e_5 e_5^{\mathsf{T}}, & C'_{44} &:= e_6 e_6^{\mathsf{T}} - e_7 e_7^{\mathsf{T}} \\ C'_{22} &:= e_2 e_2^{\mathsf{T}} - e_9 e_9^{\mathsf{T}}, & C'_{55} &:= e_8 e_8^{\mathsf{T}} - e_{10} e_{10}^{\mathsf{T}} \end{aligned}$$

Then (10.13a) is equivalent to

$$\operatorname{tr}(C'_r X') = 0, \qquad r = 1,2,3,4,5$$

□

10.1.7 Strong SOCP relaxations: mesh network

1 Strong SOCP relaxations are proposed and their relation with SOCP and SDP relaxations are studied in [118].
2 SDP, SOCP and strong SOCP relaxations are applied to a two-stage robust AC OPF problem, and column-and-constraint generation method of [119, 120] are used to solve these relaxations.
3 Check out Lingling Fan's recent paper: A sparse Convex AC OPF Solver and Convex Iteration Implementation Based on 3-Node Cycles Minyue Ma, Lingling Fan, Zhixin Miao, Bo Zeng, Hossein Ghassempour.

10.1.8 Proofs

Proof of Theorem 10.1: Rank-1 characterization.

We will prove (1) $\Rightarrow$ (2) $\Rightarrow$ (3) $\Rightarrow$ (1). If X is psd rank-1 then all its principal submatrices are psd and of rank 1 (the submatrix cannot be of rank 0 because, by assumption, $X_{jj} > 0$ for all $j \in N$). This implies that its submatrix $X_{c(F)}$ is psd and rank-1. Hence (1) $\Rightarrow$ (2).

Fix a partial matrix $X_{c(F)}$ that is psd and rank-1 and consider its submatrix X_F. Since each link $(j,k) \in E$ is a clique of $c(F)$ the 2×2 principal submatrix $X_F(j,k)$ is psd and rank-1. Therefore to prove that (2) $\Rightarrow$ (3), it suffices to show that X_F satisfies the cycle condition (10.5). We now prove the following statement by induction on k: for all cycles $c := (j_1,\ldots,j_k)$ of length $3 \le k \le n$ in $c(F)$, such that the lines $(j_i, j_{i+1}) \in c$ with $j_{k+1} := j_1$, we have

$$\sum_{i=1}^{k} \angle [X_F]_{j_i j_{i+1}} = 0 \mod 2\pi \tag{10.14}$$

For $k = 3$, a cycle $c := (j_1, j_2, j_3)$ is a clique of $c(F)$ and therefore the following principal submatrix of $X_{c(F)}$:

$$X_{c(F)}(j_1,j_2,j_3) := \begin{bmatrix} [X_{c(F)}]_{j_1 j_1} & [X_{c(F)}]_{j_1 j_2} & [X_{c(F)}]_{j_1 j_3} \\ [X_{c(F)}]_{j_2 j_1} & [X_{c(F)}]_{j_2 j_2} & [X_{c(F)}]_{j_2 j_3} \\ [X_{c(F)}]_{j_3 j_1} & [X_{c(F)}]_{j_3 j_2} & [X_{c(F)}]_{j_3 j_3} \end{bmatrix}$$

defined on the cycle is psd rank-1. Hence $X_{c(F)}(j_1,j_2,j_3) = xx^{\mathsf{H}}$ for some $x := (x_1,x_2,x_3) \in \mathbb{C}^3$. Then

$$\sum_{i=1}^{3} \angle [X_F]_{j_i j_{i+1}} = \angle \left(x_1 x_2^{\mathsf{H}}\right) + \angle \left(x_2 x_3^{\mathsf{H}}\right) + \angle \left(x_3 x_1^{\mathsf{H}}\right) \;=\; 0 \mod 2\pi$$

Suppose (10.14) holds for all cycles in $c(F)$ of length up to $k > 3$. Consider now a cycle $(j_1,\ldots,j_{k+1})$ of length $k+1$ in $c(F)$. Since $c(F)$ is chordal there is a chord, say,

$(j_1, j_m) \in E$ for some $1 < m < k+1$. Since both cycles $(j_1, \ldots, j_m)$ and $(j_1, j_m, \ldots, j_{k+1})$ satisfy (10.14) we have

$$\sum_{i=1}^{m-1} \angle [X_F]_{j_i j_{i+1}} + \angle [X_F]_{j_m j_1} = 0 \mod 2\pi$$

$$\angle [X_F]_{j_1 j_m} + \sum_{i=m}^{k+1} \angle [X_F]_{j_i j_{i+1}} = 0 \mod 2\pi$$

where $j_{k+2} := j_1$. Since X_F is Hermitian, $\angle [X_F]_{j_m j_1} = -\angle [X_F]_{j_1 j_m}$ and hence adding the above equations yields

$$\sum_{i=1}^{k+1} \angle [X_F]_{j_i j_{i+1}} = 0 \mod 2\pi$$

proving (10.14) for $k+1$. This completes the proof of (2) $\Rightarrow$ (3).

For (3) $\Rightarrow$ (1), fix any partial matrix X_F that is 2×2 psd rank-1 and satisfies the cycle condition (10.5). We can construct a psd rank-1 completion X of X_F, by constructing a vector $x \in \mathbb{C}^n$ such that $X = xx^{\mathsf{H}}$, using the method (10.10) of Chapter 10.1.4 for solution discovery, applied to each connected component of F if F is not connected, with an arbitrary spanning tree for each connected component. This defines x_j for all $j \in \{1, \ldots, n\}$. Clearly $X = xx^{\mathsf{H}}$ is a psd rank-1 completion of X_F. For uniqueness of X see the proof of Corollary 10.2. This completes the proof. □

Proof of Corollary 10.2: Uniqueness of rank-1 completion.

The proof of Theorem 10.1 shows that given a partial matrix $X_{c(F)} \in \mathbb{X}_{c(F)}$, the (unique) submatrix X_F of $X_{c(F)}$ has a psd rank-1 completion $X \in \mathbb{X}$. Therefore to prove the corollary it suffices to prove that any partial matrix $X_F \in \mathbb{X}_F$ has a unique psd rank-1 completion $X \in \mathbb{X}$. To this end fix an $X_F \in \mathbb{X}_F$ and suppose there are two psd rank-1 completions $X := xx^{\mathsf{H}}$ and $\hat{X} := \hat{x}\hat{x}^{\mathsf{H}}$ in $\mathbb{X}$. Since $X_F = \hat{X}_F$ we have

$$|x_j| = \sqrt{[X_F]_{jj}} = |\hat{x}_j|, \quad j \in \overline{N}$$

and

$$\theta_j - \theta_k = \angle [X_F]_{jk} = \hat{\theta}_j - \hat{\theta}_k, \quad (j,k) \in E$$

i.e., $C^{\mathsf{T}}\theta = C^{\mathsf{T}}\hat{\theta}$ where C is the $|N| \times |E|$ incidence matrix of the graph $G := (N, E)$:

$$C_{jl} := \begin{cases} 1 & \text{if } l = j \to k \text{ for some bus } k \\ -1 & \text{if } l = i \to j \text{ for some bus } i \\ 0 & \text{otherwise} \end{cases}, \qquad j \in N,\ l \in E$$

This means that $C^{\mathsf{T}}\left(\hat{\theta} - \theta\right) = 0$. The cycle condition (10.5) in $\mathbb{X}_F$ guarantees that there is a solution for $\hat{\theta} - \theta$ when the graph F is not a tree. Since the graph F is connected,

the null space of C^{T} is span($\mathbf{1}$), and therefore, $\hat{\theta} = \theta + \gamma\mathbf{1}$ for any $\gamma \in \mathbb{R}$. Hence $\hat{x} = xe^{\mathbf{i}\gamma}$. This implies that

$$\hat{X} \;=\; \hat{x}\hat{x}^{\mathsf{H}} \;=\; \left(xe^{\mathbf{i}\gamma}\right)\left(xe^{\mathbf{i}\gamma}\right)^{\mathsf{H}} \;=\; X$$

i.e., the psd rank-1 completion is unique. □

Proof of Theorem 10.4: Tightness of relaxations.

First $\mathbb{V} \sqsubseteq \mathbb{X}^+ \sqsubseteq \mathbb{X}^+_{c(F)} \sqsubseteq \mathbb{X}^+_F$ follows from Theorem 10.3 and the definitions of $\mathbb{X}^+$, $\mathbb{X}^+_{c(F)}$, $\mathbb{X}^+_F$ (recall that by assumption the cost function C depends on $V, X, X_{c(F)}$ only through the submatrix X_F). Since $c(F)$ is chordal, [112, Theorem 7] implies that every $X_{c(F)}$ in $\mathbb{X}^+_{c(F)}$ has a psd completion X in $\mathbb{X}^+$, i.e., $\mathbb{X}^+_{c(F)} \sqsubseteq \mathbb{X}^+$. Hence $\mathbb{X}^+ \simeq \mathbb{X}^+_{c(F)}$.

Suppose F is a tree and consider any chordal extension $c(F)$. We need to show that $\mathbb{X}^+_F \sqsubseteq \mathbb{X}^+_{c(F)}$, i.e., given any $X_F \in \mathbb{X}^+_F$ there is a $X_{c(F)} \in \mathbb{X}^+_{c(F)}$ with the same cost. Since F is itself chordal, [112, Theorem 7] implies that X_F has a psd completion X in $\mathbb{X}^+$. The submatrix $X_{c(F)}$ of X defined on $c(F)$ is the desired partial matrix in $\mathbb{X}^+_{c(F)}$ with the same cost. This proves $\mathbb{X}^+_F \sqsubseteq \mathbb{X}^+_{c(F)}$ and hence $\mathbb{X}^+_F \simeq \mathbb{X}^+_{c(F)}$ for radial networks. □

10.2 Application to OPF

In this section we apply the results of Chapter 10.1 to single-phase OPF problems in the bus injection model. In Chapter 10.2.1 we write OPF (9.16) as a standard QCQP but expressed in terms of the partial matrix defined on the network graph G. Its semidefinite relaxations then follow from (10.9). In Chapter 10.2.2 we define exact relaxation of OPF. Sufficient conditions for exact relaxations of OPF for radial networks will be studied in Chapters 10.3 and 10.4.

10.2.1 Semidefinite relaxations

Constraints.

Recall the undirected connected graph $G = (\overline{N}, E)$ that models a power network with $N+1$ buses and M lines. Given a voltage vector $V \in \mathbb{V}$ define the partial matrix $W_G := W_G(V)$:

$$[W_G]_{jj} := |V_j|^2, \quad j \in \overline{N}; \quad [W_G]_{jk} := V_j V_k^{\mathsf{H}} =: [W_G]_{kj}^{\mathsf{H}}, \quad (j,k) \in E$$

Then the constraints in OPF (9.16) as a QCQP can be written in terms of the partial matrix $W_G := W_G(V)$ as:

$$p_j^{\min} \le \operatorname{tr}(\Phi_j W_G) \le p_j^{\max}, \qquad j \in \overline{N} \tag{10.16a}$$
$$q_j^{\min} \le \operatorname{tr}(\Psi_j W_G) \le q_j^{\max}, \qquad j \in \overline{N} \tag{10.16b}$$
$$v_j^{\min} \le \operatorname{tr}(E_j W_G) \le v_j^{\max}, \qquad j \in \overline{N} \tag{10.16c}$$
$$\operatorname{tr}(\hat{Y}_{jk} W_G) \le \ell_{jk}^{\max}, \qquad (j,k) \in E \tag{10.16d}$$
$$\operatorname{tr}(\hat{Y}_{kj} W_G) \le \ell_{kj}^{\max}, \qquad (j,k) \in E \tag{10.16e}$$

Cost function.

Common cost functions can also be expressed in terms of the partial matrix W_G. For example if the cost is a weighted sum of real generation power then

$$C(W_G) = \sum_{j:\text{gens}} c_j \operatorname{Re}(s_j) = \sum_{j:\text{gens}} c_j \operatorname{tr}(\Phi_j W_G)$$

In particular the real line loss in the network is:

$$C(W_G) = \sum_j \operatorname{Re}(s_j) = \sum_j \operatorname{tr}(\Phi_j W_G)$$

We present a less obvious example.

Example 10.3 (Cost function)**.** Consider the problem of minimizing the total deviation of squared voltage magnitudes from their squared nominal values $a_j \in \mathbb{R}$

$$\min_{V \in \mathbb{C}^{N+1}} \sum_j \left(|V_j|^2 - a_j \right)^2 \qquad \text{s.t.} \qquad V \in \mathbb{V} \tag{10.17}$$

where the feasible set $\mathbb{V}$ is defined by quadratic constraints in terms of the partial matrix W_G: $V \in \mathbb{V}$ if and only if

$$V^{\mathsf{H}} C_l V = \operatorname{tr}(C_l W_G) \le b_l, \qquad l = 1, \ldots, L$$

with some matrices C_l and real numbers b_l such that $[C_l]_{jk} = 0$ if $(j,k) \notin E$. Even though the cost function is not a quadratic form in terms of W_G, show that the problem can be equivalently expressed as a QCQP in terms of W_G with additional variables and constraints.

Solution. The cost function is $\sum_j \left(|V_j|^4 - 2a_j |V_j|^2 + a_j^2 \right)$. We can omit the constants a_j^2 in the cost and hence (10.17) is equivalent to the following problem:

$$\min_{V \in \mathbb{C}^{N+1}} \sum_j \left(|U_j|^2 - 2a_j U_j \right) \quad \text{s. t.} \quad V \in \mathbb{V}, \quad U_j = |V_j|^2, \quad j \in \overline{N} \tag{10.18a}$$

Let $V := \left(V_j, j \in \overline{N} \right) \in \mathbb{C}^{N+1}$, $U := \left(U_j, j \in \overline{N} \right) \in \mathbb{C}^{N+1}$, $a := \left(a_j, j \in \overline{N} \right)$, and $e_j \in$

$\{0,1\}^{N+1}$ with a single 1 at the jth entry and 0 elsewhere. In terms of the variable $x := (V, U) \in \mathbb{C}^{2(N+1)}$, we rewrite (10.18a) as an inhomogeneous QCQP of the form:

$$\min_{x \in \mathbb{C}^{2(N+1)}} \; x^{\mathsf{H}} C_0 x + \left(c_0^{\mathsf{H}} x + x^{\mathsf{H}} c_0\right) \quad \text{s. t.} \quad V \in \mathbb{V}, \quad x^{\mathsf{H}} C_j x + \left(c_j^{\mathsf{H}} x + x^{\mathsf{H}} c_j\right) = 0, \quad j \in \overline{N} \tag{10.18b}$$

Indeed

$$\begin{aligned} \sum_j \left(|U_j|^2 - 2a_j U_j\right) &= U^H U - \left(a^H U + U^H a\right) \\ |V_j|^2 - U_j &= V^{\mathsf{H}} \left(e_j e_j^{\mathsf{H}}\right) V - \frac{1}{2}\left(e_j^{\mathsf{H}} U_j + U_j^{\mathsf{H}} e_j\right), \qquad j \in \overline{N} \end{aligned}$$

since a_j and $U_j = |V_j|^2$ are real numbers. Therefore (10.18a) is an inhomogeneous QCQP of the form (10.18b) with

$$\begin{aligned} C_0 &:= \begin{bmatrix} 0 & 0 \\ 0 & \mathbb{I}_{N+1} \end{bmatrix}, & c_0 &:= \begin{bmatrix} 0 \\ -a \end{bmatrix} \\ C_j &:= \begin{bmatrix} e_j e_j^{\mathsf{H}} & 0 \\ 0 & 0 \end{bmatrix}, & c_j &:= \begin{bmatrix} 0 \\ -\frac{1}{2} e_j \end{bmatrix}, & j &\in \overline{N} \end{aligned}$$

where $\mathbb{I}_{N+1}$ is the identity matrix of size $N+1$. Since the cost function and the new constraints depends on V only through $|V_j|^2$, in particular, it does not depend on $V_j V_k^{\mathsf{H}}$, $j \neq k$, the problem (10.18b) depends only on W_G. Indeed W_G appears only in the term $V^{\mathsf{H}} \left(e_j e_j^{\mathsf{H}}\right) V = \mathrm{tr}\left(\left(e_j e_j^{\mathsf{H}}\right) V V^{\mathsf{H}}\right) = \mathrm{tr}\left(\left(e_j e_j^{\mathsf{H}}\right) W_G\right)$.

As explained in Chapter 9.1.3, the inhomogeneous QCQP (10.18b) is equivalent to the following homogeneous QCQP with an auxiliary scalar variable $t \in \mathbb{C}$:

$$\begin{aligned} \min_{x \in \mathbb{C}^{2(N+1)}, t \in \mathbb{C}} \; & \begin{bmatrix} x^{\mathsf{H}} & t^{\mathsf{H}} \end{bmatrix} \begin{bmatrix} C_0 & c_0 \\ c_0^{\mathsf{H}} & 0 \end{bmatrix} \begin{bmatrix} x \\ t \end{bmatrix} \\ \text{s. t.} \;\; & V \in \mathbb{V} \\ & \begin{bmatrix} x^{\mathsf{H}} & t^{\mathsf{H}} \end{bmatrix} \begin{bmatrix} C_j & c_j \\ c_j^{\mathsf{H}} & 0 \end{bmatrix} \begin{bmatrix} x \\ t \end{bmatrix} = 0, \qquad j \in \overline{N} \\ & \begin{bmatrix} x^{\mathsf{H}} & t^{\mathsf{H}} \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ t \end{bmatrix} = 1 \end{aligned}$$

in the sense that, if $(x^{\mathrm{opt}}, t^{\mathrm{opt}}) \in \mathbb{C}^{2N+3}$ is optimal for the homogeneous QCQP, then their product $x^{\mathrm{opt}} t^{\mathrm{opt}} = x^{\mathrm{opt}} e^{\mathbf{i}\theta^{\mathrm{opt}}}$ is optimal for the inhomogeneous problem (10.18b). □

Henceforth we will abuse notation and use C to denote the cost function both as a function $C(V)$ of the voltage vector $V \in \mathbb{C}^{N+1}$ and as a function $C(W_G)$ of a partial matrix W_G.

OPF and relaxations.

Recall the OPF problem (9.16) as a QCQP, reproduced here

$$\min_{V} \; C(V) \quad \text{s.t.} \quad V \in \mathbb{V} := \left\{ V \in \mathbb{C}^{N+1} \mid V^{\mathsf{H}} C_l V \le b_l,\ l = 1, \ldots, L \right\} \tag{10.19}$$

where the constraint matrices C_l are given in (10.16). To avoid triviality we will assume unless otherwise specified that OPF (10.19) is feasible. Define the set of Hermitian matrices:

$$\mathbb{W} := \{\, W \in \mathbb{S}^{N+1} \mid W \text{ satisfies (10.16) with } W_G \text{ replaced by } W,\ \text{(10.4a)} \,\}$$

Fix any chordal extension $c(G)$ of G and define the set of Hermitian partial matrices $W_{c(G)}$:

$$\mathbb{W}_{c(G)} := \{\, W_{c(G)} \mid W_{c(G)} \text{ satisfies (10.16) with } W_G \text{ replaced by } W_{c(G)},\ \text{(10.4b)} \,\}$$

Finally define the set of Hermitian partial matrices W_G:

$$\mathbb{W}_G := \{\, W_G \mid W_G \text{ satisfies (10.16)(10.4c)(10.5)} \,\}$$

Then Theorem 10.3 implies that OPF (10.19) is equivalent to

$$\min_{W} C(W_G) \quad \text{s.t. } W \in \hat{\mathbb{W}}$$

where $\hat{\mathbb{W}}$ is any one of the equivalent feasible sets $\mathbb{W}, \mathbb{W}_{c(G)}, \mathbb{W}_G$. Its semidefinite relaxation relaxes $\hat{\mathbb{W}}$ to semidefinite cones:

$$\begin{aligned}
\mathbb{W}^{+} &:= \{\, W \in \mathbb{S}^{N+1} \mid W_G \text{ satisfies (10.16)}, \ W \succeq 0 \,\} \\
\mathbb{W}^{+}_{c(G)} &:= \{\, W_{c(G)} \mid W_G \text{ satisfies (10.16)}, \ W_{c(G)} \succeq 0 \,\} \\
\mathbb{W}^{+}_{G} &:= \{\, W_G \mid W_G \text{ satisfies (10.16)}, \ W_G(j,k) \ge 0, \ (j,k) \in E \,\}
\end{aligned}$$

i.e., the semidefinite relaxations of OPF (10.19) is:

$$\min_{W} C(W_G) \quad \text{s.t. } W \in \hat{\mathbb{W}}^{+}$$

where $\hat{\mathbb{W}}^{+}$ is any one of the feasible sets $\mathbb{W}^{+}, \mathbb{W}^{+}_{c(G)}, \mathbb{W}^{+}_{G}$. Explicitly, these relaxations are (c.f. (10.9)):

OPF-sdp:

$$\min_{W \in \mathbb{S}^{N+1}} \; C(W_G) \quad \text{s.t.} \quad \operatorname{tr}(C_l W) \le b_l, \quad l = 1, \ldots, L, \quad W \succeq 0 \tag{10.20a}$$

OPF-ch:

$$\min_{W_{c(G)}} \; C(W_G) \quad \text{s.t.} \quad \operatorname{tr}\left(C_l W_{c(G)}\right) \le b_l, \quad l = 1, \ldots, L, \quad W_{c(G)} \succeq 0 \tag{10.20b}$$

OPF-socp:

$$\min_{W_G} \; C(W_G) \quad \text{s.t.} \quad \operatorname{tr}(C_l W_G) \le b_l, \quad l = 1, \ldots, L, \quad W_G(j,k) \ge 0, \quad (j,k) \in E \tag{10.20c}$$

where C_l are given in (10.16). Since OPF (9.16) as a QCQP does not require assumption C4.1 that $y^s_{jk} = y^s_{kj}$, neither does its semidefinite relaxations (10.20). They can therefore accommodate single-phase transformers that have complex turns ratios.

As discussed in Remark 10.2, if the network graph G is a tree, then we should solve OPF-socp to compute the partial matrix W_G because it will be as tight as OPF-sdp that computes the entire matrix W, but much simpler computationally. Otherwise we can solve OPF-ch to compute $W_{c(G)}$ corresponding to a chordal extension $c(G)$ of G which is usually much simpler than OPF-sdp for large sparse network but as tight.

Example 10.4 (Two-bus network)**.** For the two-bus network in Figure 10.2, suppose the line is a series admittance $y = g + \mathbf{i}b$ and the load (p_2, q_2) is given. Write OPF and its relaxation as QCCPs assuming C is the cost matrix and line limits are neglected.

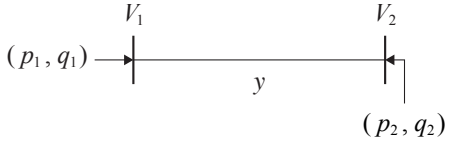


Figure 10.2 Example 10.4.

Solution. The complex form power flow solution is (from Chapter 4.3.1):

$$s_1 = \bar{y}\left(|V_1|^2 - V_1\bar{V}_2\right), \qquad s_2 = \bar{y}\left(|V_2|^2 - V_2\bar{V}_1\right)$$

Therefore the admittance matrix and the associated Y_1, Y_2 are:

$$Y := \begin{bmatrix} y & -y \\ -y & y \end{bmatrix}, \quad Y_1 := e_1 e_1^{\mathsf{T}} Y = \begin{bmatrix} y & -y \\ 0 & 0 \end{bmatrix}, \quad Y_2 := e_2 e_2^{\mathsf{T}} Y = \begin{bmatrix} 0 & 0 \\ -y & y \end{bmatrix}$$

The matrices in (10.16) are:

$$\Phi_1 := \frac{1}{2}\left(Y_1^{\mathsf{H}} + Y_1\right) = \begin{bmatrix} g & -y/2 \\ -\bar{y}/2 & 0 \end{bmatrix}, \quad \Psi_1 := \frac{1}{2\mathbf{i}}\left(Y_1^{\mathsf{H}} - Y_1\right) = \begin{bmatrix} -b & y/(2\mathbf{i}) \\ -\bar{y}/(2\mathbf{i}) & 0 \end{bmatrix}$$

$$\Phi_2 := \frac{1}{2}\left(Y_2^{\mathsf{H}} + Y_2\right) = \begin{bmatrix} 0 & -\bar{y}/2 \\ -y/2 & g \end{bmatrix}, \quad \Psi_2 := \frac{1}{2\mathbf{i}}\left(Y_2^{\mathsf{H}} - Y_2\right) = \begin{bmatrix} 0 & -\bar{y}/(2\mathbf{i}) \\ y/(2\mathbf{i}) & -b \end{bmatrix}$$

$J_1 = e_1 e_1^{\mathsf{T}}$ and $J_2 = e_2 e_2^{\mathsf{T}}$. Then OPF is:

$$\begin{aligned} \min_{V \in \mathbb{C}^2} \; V^{\mathsf{H}} C V \quad \text{s.t.} \quad & p_1^{\min} \le p_1 = V^{\mathsf{H}}\Phi_1 V \le p_1^{\max}, \quad q_1^{\min} \le q_1 = V^{\mathsf{H}}\Psi_1 V \le q_1^{\max} \\ & v_1^{\min} \le |V_1|^2 = V^{\mathsf{H}} J_1 V \le v_1^{\max}, \quad v_2^{\min} \le |V_2|^2 = V^{\mathsf{H}} J_2 V \le v_2^{\max} \\ & V^{\mathsf{H}}\Phi_2 V = p_2, \quad V^{\mathsf{H}}\Psi_2 V = q_2 \end{aligned}$$

Its SDP relaxation is:

$$\min_{W \in \mathbb{S}^2} \ \operatorname{tr}(CW) \qquad \text{s.t.} \qquad \begin{aligned} & p_1^{\min} \le \operatorname{tr}(\Phi_1 W) \le p_1^{\max}, && q_1^{\min} \le \operatorname{tr}(\Psi_1 W) \le q_1^{\max} \\ & v_1^{\min} \le \operatorname{tr}(J_1 W) \le v_1^{\max}, && v_2^{\min} \le \operatorname{tr}(J_2 W) \le v_2^{\max} \\ & \operatorname{tr}(\Phi_2 W) = p_2, \quad \operatorname{tr}(\Psi_2 W) = q_2, && W \succeq 0 \end{aligned}$$

□

10.2.2 Exact relaxation: definition

Consider the single-phase OPF (10.19) as a standard QCQP and its semidefinite relaxations (10.20).

Definition 10.2 (Strong exactness)**.** We say that

1. OPF-sdp (10.20a) is *exact* if every optimal solution W^{sdp} of OPF-sdp is psd rank-1;
2. OPF-ch (10.20b) is *exact* if every optimal solution $W^{\text{ch}}_{c(G)}$ of OPF-ch is psd rank-1, i.e., the principal submatrices $W^{\text{ch}}_{c(G)}(q)$ of $W^{\text{ch}}_{c(G)}$ are psd rank-1 for all maximal cliques q of the chordal extension $c(G)$ of graph G;
3. OPF-socp (10.20c) is *exact* if every optimal solution W^{socp}_G of OPF-socp
   - is 2×2 psd rank-1, i.e., the 2×2 principal submatrices $W^{\text{socp}}_G(j,k)$ are psd rank-1 for all $(j,k) \in E$; and
   - satisfies the cycle condition (10.5).

Exactness does not guarantee the existence of an optimal solution. If a relaxation is infeasible then the original OPF is also infeasible. To recover an optimal solution V^{opt} of OPF (10.19) from an optimal solution W^{sdp} or $W^{\text{ch}}_{c(G)}$ or W^{socp}_G of its relaxations, see Chapter 10.1.4. The strong exactness notion in Definition 10.2 is convenient because it ensures that any algorithm that solves an exact relaxation always produces a globally optimal solution to the OPF problem. For a weaker notion of exactness that requires at least one (not necessarily all) optimal solution of the relaxation, if exists, be feasible and therefore optimal for the original nonconvex OPF problem, an algorithm may not be guaranteed to produce an optimal solution of OPF by solving its relaxation. This strong notion of exactness is however more stringent than necessary under the sufficient exactness conditions of Chapters 10.3 and 10.4 for radial networks. See Remark 10.3 after Theorem 10.6 and Remark 10.4 after Theorem 10.9 (and Remarks 11.1 and 11.3 for BFM). These conditions guarantee that an optimal solution to OPF can always be recovered from *any* optimal solution of OPF-socp for radial networks, even when the OPF-socp is not exact under Definition 10.2.

In the rest of this chapter we present sufficient conditions for exact semidefinite relaxations when the network is radial, i.e., the network graph is a tree. We restrict our discussion to single-phase networks though exactness conditions exist in the literature for three-phase radial networks.

10.3 Exactness condition: linear separability

Theorem 10.4 implies that, for a single-phase radial network whose graph G is a tree, if SOCP relaxation is exact then SDP and chordal relaxations are also exact. We hence focus on the exactness of OPF-socp (10.20c). Since the cycle condition (10.5) is vacuous for radial networks, OPF-socp (10.20c) is exact if all of its optimal solutions are 2×2 rank-1. To avoid triviality we assume OPF (10.19) is feasible.

We will first present a general result on the exactness of the SOCP relaxation of general QCQP on a tree graph G and then apply it to OPF-socp (10.20c) for single-phase radial networks.

10.3.1 Sufficient condition for QCQP

Fix an undirected graph $G = (N, E)$ where $|N| = n$ and $E \subseteq N \times N$. Fix Hermitian matrices $C_l \in \mathbb{S}^n$, $l = 0, \ldots, L$, defined on G, i.e., $[C_l]_{jk} = 0$ if $(j,k) \notin E$. Consider QCQP:

$$C^{\text{opt}} := \min_{x\in\mathbb{C}^n} \; x^H C_0 x \qquad \text{s.t.} \qquad x^H C_l x \le b_l, \;\; l = 1, \ldots, L \tag{10.21}$$

where $b_l \in \mathbb{R}$, $l = 1, \ldots, L$, and its SOCP relaxation where the optimization variable ranges over Hermitian partial matrices X_G:

$$C^{\text{socp}} := \min_{X_G} \; \text{tr}\,(C_0 X_G) \qquad \text{s.t.} \qquad \text{tr}\,(C_l X_G) \le b_l, \;\; l = 1, \ldots, L \tag{10.22a}$$

$$X_G(j,k) \succeq 0, \;\; (j,k) \in E \tag{10.22b}$$

The following result can be regarded as an extension of [121] on the SOCP relaxation of QCQP from the real domain to the complex domain. Consider: [2]

- C10.1: For each link $(j,k) \in E$ there exists an α_{jk} such that $\angle\,[C_l]_{jk} \in [\alpha_{ij}, \alpha_{ij} + \pi]$ for all $l = 0, \ldots, L$.
- C10.2: The cost matrix C_0 is positive definite.

Condition C10.1 is illustrated in Figure 10.3. Let C^{opt} and C^{socp} denote the optimal values of QCQP (10.21) and SOCP (10.22) respectively.

Theorem 10.6 (Linear separability)**.** Suppose G is a tree and C10.1 holds. Then $C^{\text{opt}} = C^{\text{socp}}$ and an optimal solution $x^{\text{opt}} \in \mathbb{C}^n$ of QCQP (10.21) can be recovered from every optimal solution X_G^{socp} of SOCP (10.22).

Remark 10.3 (Strong exactness)**.** The SOCP relaxation may not be exact in the strong sense of Definition 10.2, i.e., some optimal solutions of (10.22) may be 2×2 psd

[2] All angles should be interpreted as "mod 2π", i.e., projected onto $(-\pi, \pi]$.

but not 2×2 rank-1, but Theorem 10.6 says that C10.1 guarantees that an optimal solution of QCQP (10.21) can always be recovered from *any* optimal solution x^{socp} of its SOCP relaxation (10.22) whether or not x^{socp} is 2×2 rank-1. The proof of the theorem prescribes a simple procedure to do that; see Chapter 10.3.3. □

If the objective function is strictly convex however then the optimal solution is unique and SOCP (10.22) is indeed exact in the sense of Definition 10.2.

Corollary 10.7. Suppose G is a tree and C10.1, C10.2 hold. Then SOCP (10.22) is exact.

10.3.2 Application to OPF

We now apply Theorem 10.6 to our OPF problem (10.19) where the constraint matrices C_l are given in (10.16). Since the formulation does not require assumption $y^s_{jk} = y^s_{kj}$ (assumption C4.1) and allows nonzero shunt admittances (y^m_{jk}, y^m_{kj}), and can therefore accommodate single-phase transformers that have complex turns ratios.

To simplify illustration we ignore the branch constraints (10.16d)(10.16e), which reduces (10.19) to:

$$\min_{x\in\mathbb{C}^n} \; V^{\mathsf{H}} C_0 V \qquad \text{s.t.} \quad V^{\mathsf{H}}\Phi_j V \le p_j^{\max}, \quad V^{\mathsf{H}}(-\Phi_j)V \le -p_j^{\min}, \quad j\in\overline{N} \tag{10.23a}$$

$$V^{\mathsf{H}}\Psi_j V \le q_j^{\max}, \quad V^{\mathsf{H}}(-\Psi_j)V \le -q_j^{\min}, \quad j\in\overline{N} \tag{10.23b}$$

$$V^{\mathsf{H}}E_j V \le v_j^{\max}, \quad V^{\mathsf{H}}(-E_j)V \le -v_j^{\min}, \quad j\in\overline{N} \tag{10.23c}$$

for some Hermitian matrices C_0, Φ_j, Ψ_j, E_j where $j\in\overline{N}$. Condition C10.1 depends only on the off-diagonal entries of C_0, Φ_j, Ψ_j (E_j are diagonal matrices). It implies a simple pattern on the power injection constraints (10.23a)(10.23b). Write the series admittances in terms of its real and imaginary parts $y^s_{jk} =: g^s_{jk} + \mathbf{i}b^s_{jk}$ with $g^s_{jk} > 0, b^s_{jk} < 0$. (Note that C10.1 does not depend on the shunt admittances $\left(y^m_{jk}, y^m_{kj}\right)$.) Then we have

$$[\Phi_k]_{ij} = \begin{cases} \frac{1}{2}Y_{ij} = -\frac{1}{2}(g^s_{ij} + \mathbf{i}b^s_{ij}) & \text{if } k = i \\ \frac{1}{2}Y^{\mathsf{H}}_{ij} = -\frac{1}{2}(g^s_{ij} - \mathbf{i}b^s_{ij}) & \text{if } k = j \\ 0 & \text{if } k \notin \{i,j\} \end{cases}$$

$$[\Psi_k]_{ij} = \begin{cases} \frac{-1}{2\mathbf{i}}Y_{ij} = \frac{1}{2}(b^s_{ij} - \mathbf{i}g^s_{ij}) & \text{if } k = i \\ \frac{1}{2\mathbf{i}}Y^{\mathsf{H}}_{ij} = \frac{1}{2}(b^s_{ij} + \mathbf{i}g^s_{ij}) & \text{if } k = j \\ 0 & \text{if } k \notin \{i,j\} \end{cases}$$

Hence for each line $(j,k) \in E$ the relevant angles for C10.1 are those of $[C_0]_{jk}$ and

$$\begin{aligned} [\Phi_j]_{jk} &= -\frac{1}{2}\left(g^s_{jk} + \mathbf{i} b^s_{jk}\right), & [\Phi_k]_{jk} &= -\frac{1}{2}\left(g^s_{jk} - \mathbf{i} b^s_{jk}\right) \\ [\Psi_j]_{jk} &= \frac{1}{2}\left(b^s_{jk} - \mathbf{i} g^s_{jk}\right), & [\Psi_k]_{jk} &= \frac{1}{2}\left(b^s_{jk} + \mathbf{i} g^s_{jk}\right) \end{aligned}$$

as well as the angles of $-[\Phi_j]_{jk}, -[\Phi_k]_{jk}$ and $-[\Psi_j]_{jk}, -[\Psi_k]_{jk}$. These quantities are shown in Figure 10.3 with their magnitudes normalized to a common value and explained in the caption of the figure.

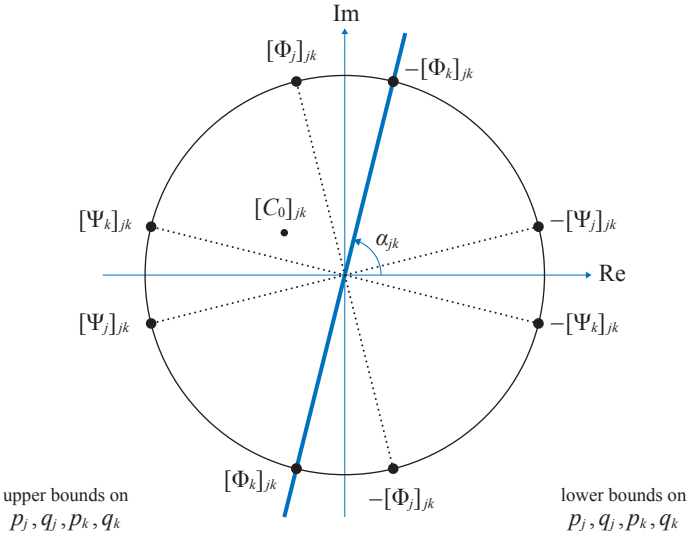


Figure 10.3 Condition C10.1' for OPF on a line $(j,k) \in E$. The quantities $([\Phi_j]_{jk}, [\Phi_k]_{jk}, [\Psi_j]_{jk}, [\Psi_k]_{jk})$ on the left-half plane correspond to finite upper bounds on (p_j, p_k, q_j, q_k) in (10.23a)(10.23b); $(-[\Phi_j]_{jk}, -[\Phi_k]_{jk}, -[\Psi_j]_{jk}, -[\Psi_k]_{jk})$ on the right-half plane correspond to finite lower bounds on (p_j, p_k, q_j, q_k).

Condition C10.1 applied to OPF (10.23) takes the following form (see Figure 10.3):

C10.1': For each link $(j,k) \in E$ there is a line in the complex plane through the origin such that $[C_0]_{jk}$ as well as those $\pm[\Phi_i]_{jk}$ and $\pm[\Psi_i]_{jk}$ corresponding to *finite* lower or upper bounds on (p_i, q_i), for $i = j, k$, are all on one side of the line, possibly on the line itself.

Let C^{opt} and C^{socp} denote the optimal values of OPF and OPF-socp respectively.

Corollary 10.8. Suppose G is a tree and C10.1' holds.

1 $C^{\text{opt}} = C^{\text{socp}}$. Moreover an optimal solution V^{opt} of OPF (10.23) can be recovered from every optimal solution X_G^{socp} of OPF-socp.

2 If, in addition, C10.2 holds then OPF-socp is exact.

It is clear from Figure 10.3 that condition C10.1' cannot be satisfied if there is a line where both the real and reactive power injections at both ends are both lower and upper bounded (8 combinations as shown in the figure). C10.1' requires that some of them be unconstrained. When the cost function is convex, this is the same as requiring that the constraints be inactive at optimality (see Exercise 10.3). The result proved in [122] also includes constraints on real branch power flows and line losses. Corollary 10.8 includes several sufficient conditions in the literature for exact relaxation as special cases. Referring to Figure 10.3, the load over-satisfaction condition in [123, 124] corresponds to the red line in the figure being the Im-axis that excludes all quantities on the right-half plane. The sufficient condition in [125, Theorem 2] corresponds to the red line in the figure that allows a finite lower bound on the real power at one end of the line, i.e., p_j or p_k but not both, and no finite lower bounds on reactive powers q_j and q_k.

10.3.3 Proofs

We now prove Theorem 10.6 and Corollary 10.7, following [126]. It is equivalent to the argument of [127] and simpler than the original duality proof in [122].

Proof of Theorem 10.6.

Fix any partial matrix X_G that is feasible for SOCP (10.22). We will construct an $x \in \mathbb{C}^n$ that satisfies

$$x^{\mathsf{H}} C_l x \le \operatorname{tr} C_l X_G, \qquad l = 0, 1, \ldots, L$$

i.e., x is feasible for QCQP (10.21) and has an equal or lower cost than X_G. Since the minimum cost of QCQP is lower bounded by that of its SOCP relaxation this means that an optimal solution $x \in \mathbb{C}^n$ of QCQP (10.21) can be obtained from every optimal solution X_G of SOCP (10.22), whether or not (10.22) is exact in the sense of Definition 10.2.

Now $X_G(j,k) \ge 0$ for every $(j,k) \in E$ implies that $[X_G]_{jj} \ge 0$ for all $j \in N$ and

$$[X_G]_{jj}\,[X_G]_{kk} \ge \left|[X_G]_{jk}\right|^2, \quad (j,k) \in E$$

Case 1: X_G is 2×2 psd rank-1. Suppose $[X_G]_{jj}[X_G]_{kk} = |[X_G]_{jk}|^2$ for all $(j,k) \in E$. We will construct an $x \in \mathbb{C}^n$ that is feasible for QCQP and has an equal cost. To construct such an x let $|x_j| := \sqrt{[X_G]_{jj}}$, $j \in N$. Recall that G is a (connected) tree with node 1 as its root. Let $\angle x_1 := 0$. Traversing the tree starting from the root the angles can be successively assigned: given $\angle x_j$ at one end of a link (j,k), let $\angle x_k := \angle x_j - \angle [X_G]_{jk}$ at the other end. Given any X_G which is 2×2 psd rank-1, angles $\angle x_j$ can always be

consistently assigned if and only if G is a tree. (If G contains cycles then X_G must also satisfy the cycle condition according to Theorem 10.1).

With this x constructed from X_G we have, for $l = 0, 1, \ldots, L$,

$$x^{\mathsf{H}} C_l x = \sum_{j,k} [C_l]_{jk} x_j^{\mathsf{H}} x_k \;=\; \sum_{j,k} [C_l]_{jk} |x_j| |x_k| e^{\mathbf{i}(\angle x_k - \angle x_j)} \;=\; \sum_{j,k} [C_l]_{jk} \left|[X_G]_{jk}\right| e^{-\mathbf{i}\angle [X_G]_{jk}} =$$

where the last equality follows from $\mathrm{tr}(C_l X_G) = \sum_{j,k} [C_l]_{jk} [X_G]_{jk}^{\mathsf{H}}$ and that X_G is a Hermitian partial matrix. Hence x is feasible for QCQP (10.21) and has the same cost as X_G.

Case 2: X_G is 2×2 psd but not 2×2 rank-1. Suppose $[X_G]_{jj}[X_G]_{kk} > |[X_G]_{jk}|^2$ for some (j,k). We will

1 Construct an $\hat{X}_G$ that is 2×2 psd rank-1.
2 Show that C10.1 implies

$$\mathrm{tr}\, C_l \hat{X}_G \le \mathrm{tr}\, C_l X_G, \qquad l = 0, 1, \ldots, L \tag{10.24}$$

Then an $x \in \mathbb{C}^n$ can be constructed from $\hat{X}_G$ as in Case 1 and step 2 ensures that for $l = 0, 1, \ldots, L$

$$x^{\mathsf{H}} C_l x \;=\; \mathrm{tr}\, C_l \hat{X}_G \;\le\; \mathrm{tr}\, C_l X_G$$

i.e., x is feasible for QCQP (10.21) and has an equal or lower cost than X_G.

To construct such an $\hat{X}_G$ let $[\hat{X}_G]_{jj} = [X_G]_{jj}$, $j \in \overline{N}$. For each line $(j,k) \in E$ let

$$[\hat{X}_G]_{jk} - [X_G]_{jk} =: r_{jk} e^{-\mathbf{i}(\frac{\pi}{2} - \alpha_{jk})}$$

for some $r_{jk} > 0$ to be determined and α_{jk} in condition C10.1. For $\hat{X}_G$ to be 2×2 psd rank-1 we need to choose $r_{jk} > 0$ such that $[\hat{X}_G]_{jj}[\hat{X}_G]_{kk} = \left|[\hat{X}_G]_{jk}\right|^2$ for all $(j,k) \in E$, i.e.,

$$[X_G]_{jj}\,[X_G]_{kk} = \left|[X_G]_{jk} + r_{jk} e^{-\mathbf{i}(\frac{\pi}{2} - \alpha_{jk})}\right|^2$$

or

$$r_{jk}^2 + 2b\, r_{jk} - c = 0$$

where

$$b \;:=\; \mathrm{Re}\left([X_G]_{jk}\, e^{\mathbf{i}(\frac{\pi}{2} - \alpha_{jk})}\right), \quad c \;:=\; [X_G]_{jj}\,[X_G]_{kk} - \left|[X_G]_{jk}\right|^2 \;>\; 0$$

Therefore setting $r_{jk} := \sqrt{b^2 + c} - b > 0$ yields an $\hat{X}_G$ that is 2×2 psd rank-1.

To show that $\hat{X}_G$ is feasible for SOCP (10.22) and has an equal or lower cost than

X_G, we have for $l = 0, 1, \ldots, L$,

$$\begin{aligned}
\operatorname{tr} C_l \hat{X}_G - \operatorname{tr} C_l X_G &= \operatorname{tr}\left(C_l\left(\hat{X}_G - X_G\right)\right) = \sum_{(j,k)\in E} [C_l]_{jk}\left([\hat{X}_G]_{jk} - [X_G]_{jk}\right)^{\mathsf{H}} \\
&= 2\sum_{j<k,(j,k)\in E} \operatorname{Re}\left([C_l]_{jk} \cdot r_{jk}\, e^{\mathbf{i}\left(\frac{\pi}{2}-\alpha_{jk}\right)}\right) \\
&= 2\sum_{\substack{j<k \\ (j,k)\in E}} \left|[C_l]_{jk}\right| r_{jk} \cos\left(\angle[C_l]_{jk} + \frac{\pi}{2} - \alpha_{jk}\right) \leq 0
\end{aligned}$$

where the last inequality follows because assumption C10.1 implies

$$\frac{\pi}{2} \leq \angle[C_l]_{jk} + \tfrac{\pi}{2} - \alpha_{jk} \leq \frac{3\pi}{2}$$

and therefore $\cos\left(\angle[C_l]_{jk} + \frac{\pi}{2} - \alpha_{jk}\right) \leq 0$. This completes the proof. □

Proof of Corollary 10.7.

C10.2 implies that the objective function of SOCP (10.22) is strictly convex and hence has a unique optimal solution. Suppose X_G is an optimal solution of SOCP (10.22) but $[X_G]_{jj}[X_G]_{kk} > |[X_G]_{jk}|^2$ for some (j,k), i.e., X_G is 2×2 psd but not 2×2 psd rank-1. Then the proof for Theorem 10.6 constructs another feasible solution $\hat{X}_G$ with equal cost. This contradicts the uniqueness of the optimal solution of SOCP (10.22), and hence X_G must be 2×2 psd rank-1. □

10.4 Exactness condition: small angle differences

The sufficient conditions in [125, 128, 129] require that the voltage angle difference across each line be small. We explain the intuition using a result in [128] for an OPF problem under the following simplifying assumptions. We assume $y^s_{jk} = y^s_{kj}$ (assumption 4.1) and $y^m_{jk} = y^m_{kj} := 0$ for all lines (j,k). We use the polar form power flow equation (4.27) of Chapter 4.3.2, instead of the complex form that we have been using in the previous sections. We ignore reactive power and assume voltage magnitudes $|V_j|$ are fixed. Let $V_j = |V_j|\, e^{\mathbf{i}\theta_j}$. Then the optimization over (s, V) in OPF reduces to an optimization over (p, θ) as well as real line flows P as an auxiliary variable. Under these assumptions, as long as the voltage angle difference is small, the power flow solutions form a locally convex surface that is the Pareto front of its relaxation. This implies that the relaxation is exact. The intuition extends to cases where some of these assumptions are relaxed though the clean geometric insight becomes more obscure.

10.4.1 Sufficient condition

Let $y^s_{jk} = y^s_{kj} =: g_{jk} + \mathbf{i} b_{jk}$ with $g_{jk} > 0, b_{jk} < 0$ for all lines (j,k). Consider

$$\min_{p,P,\theta} C(p) \tag{10.25a}$$

$$\text{s.t. } p^{\min}_j \le p_j \le p^{\max}_j, \qquad j \in \overline{N} \tag{10.25b}$$

$$\theta^{\min}_{jk} \le \theta_{jk} \le \theta^{\max}_{jk}, \quad (j,k) \in E \tag{10.25c}$$

$$p_j = \sum_{k:k\sim j} P_{jk}, \qquad j \in \overline{N} \tag{10.25d}$$

$$P_{jk} = |V_j|^2 g_{jk} - |V_j||V_k| g_{jk} \cos\theta_{jk} - |V_j||V_k| b_{jk} \sin\theta_{jk}, \quad (j,k) \in E \tag{10.25e}$$

where $\theta_{jk} := \theta_j - \theta_k$ are the voltage angle differences across lines (j,k). The constraint (10.25c) on θ_{jk} is equivalent to a limit on the apparent line power .

We comment on the constraint (10.25c) on angles θ_{jk}. When the voltage magnitudes $|V_i|$ are fixed, constraints on real power flows, branch currents, line losses, as well as stability constraints can all be represented in terms of θ_{jk}. Indeed a line flow constraint of the form $|P_{jk}| \le P^{\max}_{jk}$ becomes a constraint on θ_{jk} using the expression for P_{jk} in (10.25e) (or see Exercise 9.8). A current constraint of the form $|I_{jk}| \le I^{\max}_{jk}$ is also a constraint on θ_{jk} since $|I_{jk}|^2 = |y_{jk}|(|V_j|^2 + |V_k|^2 - 2|V_j V_k| \cos\theta_{jk})$. The line loss over $(j,k) \in E$ is equal to $P_{jk} + P_{kj}$ which is again a function of θ_{jk}. Stability typically requires $|\theta_{jk}|$ to stay within a small threshold. Therefore given constraints on branch power or current flows, losses, and stability, appropriate bounds $\theta^{\min}_{jk}, \theta^{\max}_{jk}$ can be determined to enforce these constraints, assuming $|V_j|$ are fixed.

We can eliminate the branch flows P_{jk} and angles θ_{jk} from (10.25). Since $|V_j|, j \in \overline{N}$, are fixed we assume without loss of generality that $|V_j| = 1$ pu. Define the injection region

$$\mathbb{P}_\theta := \left\{ p \in \mathbb{R}^n \;\middle|\; p_j = \sum_{k:k\sim j} \left(g_{jk} - g_{jk}\cos\theta_{jk} - b_{jk}\sin\theta_{jk} \right),\ j \in \overline{N},\ \theta^{\min}_{jk} \le \theta_{jk} \le \theta^{\max}_{jk},\ (j,k) \in E \right\}$$

Let $\mathbb{P}_p := \{p \in \mathbb{R}^n \mid p^{\min}_j \le p_j \le p^{\max}_j, j \in N\}$. Then (10.25) is:
OPF:

$$\min_p C(p) \text{ subject to } p \in \mathbb{P}_\theta \cap \mathbb{P}_p \tag{10.26}$$

This problem is hard because the set $\mathbb{P}_\theta$ is nonconvex. To avoid triviality we assume OPF (10.26) is feasible. For a set A let $\operatorname{conv} A$ denote the convex hull of A. Consider the following problem that relaxes the nonconvex feasible set $\mathbb{P}_\theta \cap \mathbb{P}_p$ of (10.26) to a convex superset:
OPF-socp:

$$\min_p C(p) \text{ s.t. } p \in \operatorname{conv}(\mathbb{P}_\theta) \cap \mathbb{P}_p \tag{10.27}$$

We will show below that (10.27) is indeed an SOCP. It is said to be *exact* if every optimal solution of (10.27) lies in $\mathbb{P}_\theta \cap \mathbb{P}_p$ and is therefore also optimal for (10.26).

We say that a point $x \in A \subseteq \mathbb{R}^n$ is a *Pareto optimal point* in A if there does not exist another $x' \in A$ such that $x' \leq x$ with at least one strictly smaller component $x'_j < x_j$. The *Pareto front of* A, denoted by $\mathbb{O}(A)$, is the set of all Pareto optimal points in A. The significance of $\mathbb{O}(A)$ is that, for any increasing function, its minimizer, if exists, is necessarily in $\mathbb{O}(A)$ whether A is convex or not. If A is convex then x^{opt} is a Pareto optimal point in $\mathbb{O}(A)$ if and only if there is a nonzero vector $c := (c_1, \ldots, c_n) \geq 0$ such that x^{opt} is a minimizer of $c^{\mathsf{T}}x$ over A [57, pp.179–180].

Assume

. C10.3: For all $(j,k) \in E$, $\tan^{-1} \frac{b_{jk}}{g_{jk}} < \theta_{jk}^{\min} \leq \theta_{jk}^{\max} < \tan^{-1} \frac{-b_{jk}}{g_{jk}}$.
. C10.4: $C(p)$ is strictly increasing in each p_j.

The following result, proved in [125, 128, 129] says that (10.27) is exact provided θ_{jk} are suitably bounded.

Theorem 10.9. Suppose G is a tree and C10.3–C10.4 hold.

1 $\mathbb{P}_\theta \cap \mathbb{P}_p = \mathbb{O}(\text{conv}(\mathbb{P}_\theta) \cap \mathbb{P}_p)$.
2 The problem (10.27) is equivalent to (i.e., can be reformulated as) an SOCP. Moreover it is exact.

Remark 10.4 (Strong exactness)**.** Condition C10.4 is needed to ensure that *every* optimal solution of OPF-socp (10.27) is optimal for OPF (10.26). If $C(p)$ is nondecreasing but not strictly increasing in all p_j, then $\mathbb{P}_\theta \cap \mathbb{P}_p \subseteq \mathbb{O}(\text{conv}(\mathbb{P}_\theta) \cap \mathbb{P}_p)$ and OPF-socp may not be exact according to our definition. Even in that case it is possible to recover an optimal solution of OPF from any optimal solution of OPF-socp (see Exercise 10.9). □

10.4.2 Proof: 2-bus network

We now illustrate the geometric insight by proving the theorem for the case of a single line (see [128] for proof for a tree network).

Proof of Theorem 10.9: 2-bus network.

Consider two buses j and k connected by a line with admittance $y^s_{jk} = y^s_{kj} = g_{jk} + \mathbf{i}b_{jk}$ with $g_{jk} > 0, b_{jk} < 0$. Recall that we assume voltage magnitudes $|V_j| = 1$ pu are fixed for buses $j = 1, 2$, zero charging admittances, and we ignore reactive powers. Since

$p_j = P_{jk}$ and $p_k = P_{kj}$ we will work with $P := (P_{jk}, P_{kj})$. Then (the power flow equation (4.27a) in polar form)

$$\begin{aligned} P_{jk} &:= P_{jk}(\theta_{jk}) := g_{jk} - g_{jk}\cos\theta_{jk} - b_{jk}\sin\theta_{jk} \\ P_{kj} &:= P_{kj}(\theta_{jk}) := g_{jk} - g_{jk}\cos\theta_{jk} + b_{jk}\sin\theta_{jk} \end{aligned}$$

where $\theta_{jk} := \theta_j - \theta_k$, or in vector form

$$P - g_{jk}\mathbf{1} = A\begin{bmatrix}\cos\theta_{jk}\\ \sin\theta_{jk}\end{bmatrix} \tag{10.28}$$

where $\mathbf{1} := [1\ 1]^\mathsf{T}$ and A is an invertible matrix (A is not necessarily negative definite because it is not symmetric, but AA^T is positive definite since A is nonsingular):

$$A := \begin{bmatrix}-g_{jk} & -b_{jk}\\ -g_{jk} & b_{jk}\end{bmatrix}$$

The proof will proceed in four steps:

1 We show that P traces out an ellipse in $\mathbb{R}^2$ as θ_{jk} ranges over $[-\pi, \pi]$. Since the feasible set is a subset of ellipse, it is nonconvex.
2 We show that condition C10.3 restricts the feasible set to the lower half of the ellipse.
3 We show that condition C10.4 implies that the Pareto front of the feasible set of the relaxed problem (10.27) coincides with the feasible set. This implies that the relaxation is exact.
4 Finally we reformulate the relaxation (10.27) as an SOCP.

Step 1: P that satisfies (10.28) is an ellipse. In general the set of points $x \in \mathbb{R}^k$ that satisfy

$$(x-c)^\mathsf{T} M (x-c) \;=\; \left\| M^{1/2}(x-c)\right\|_2^2 = 1$$

is an ellipse if $c \in \mathbb{R}^n$ and $M \succ 0$ is a real (symmetric) positive definite matrix. The center of the ellipsoid is c and the k principal axes are the k eigenvectors of M (see Exercise 10.4). To see that P describes an ellipse, write $v := [\cos\theta_{jk}\ \sin\theta_{jk}]^\mathsf{T} = A^{-1}\left(P - g_{jk}\mathbf{1}\right)$. Hence $\|v\|_2^2 = 1$, yielding

$$(P - g_{jk}\mathbf{1})^\mathsf{T}\left(AA^\mathsf{T}\right)^{-1}(P - g_{jk}\mathbf{1}) = 1 \tag{10.29}$$

As noted above, AA^T is positive definite, implying that P is an ellipse centered at $g_{jk}\mathbf{1}$. From (10.28), the ellipse P passes through the origin when $\theta_{jk} = 0$, as shown in Figures 10.4. Since the feasible set is a subset of the ellipse P (without the interior), it is nonconvex.

Step 2: condition C10.3 restricts the feasible set to the lower half of the ellipse. Let

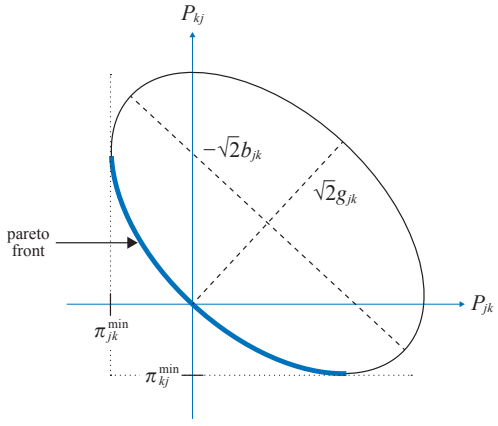


Figure 10.4 The feasible set of OPF (10.26) for the two-bus network is a subset of an ellipse without the interior, hence nonconvex. OPF-socp (10.27) includes the interior of the ellipse and is hence convex. If the cost function C is strictly increasing in (P_{jk}, P_{kj}) then the Pareto front of the SOCP feasible set will lie on the lower part of the ellipse, $\mathbb{O}(\mathbb{P}_\theta) = \mathbb{P}_\theta$, and hence OPF-socp is exact. The points $P := (P_{jk}(\theta_{jk}), P_{kj}(\theta_{kj})) = 0$ when $\theta_{jk} = 0$, $P_{jk} = \pi_{jk}^{\min}$ when $\theta_{jk} = \theta_{jk}^{\min}$, and $P_{kj} = \pi_{kj}^{\min}$ when $\theta_{jk} = \theta_{kj}^{\min}$.

$\pi_{jk}^{\min}$ denote the minimum $P_{jk}(\theta_{jk})$ and $\pi_{kj}^{\min}$ the minimum $P_{kj}(\theta_{jk})$ on the ellipse as shown in the figure. They are attained when θ_{jk} takes the values

$$\theta_{jk}^{\min} := \tan^{-1}\frac{b_{jk}}{g_{jk}} \text{ and } \theta_{kj}^{\min} := \tan^{-1}\frac{-b_{jk}}{g_{jk}}$$

respectively (Exercise 10.7). The condition $\theta_{jk}^{\min} \le \theta_{jk} \le \theta_{kj}^{\min}$ restricts $P(\theta_{jk})$ to the darkened segment of the ellipse in Figures 10.4. Recall the sets

$$\mathbb{P}_\theta := \{\, p \mid p = P, P \text{ satisfies (10.28) for } \theta_{jk}^{\min} \le \theta_{jk} \le \theta_{jk}^{\max} \,\}, \quad \mathbb{P}_p := \{p \mid p^{\min} \le p \le p^{\max}\}$$

and the feasible set $\mathbb{P}_\theta \cap \mathbb{P}_p$ of OPF (10.26). Condition C10.3 ensures $\theta_{jk}^{\min} \le \theta_{jk} \le \theta_{kj}^{\min}$ and hence restricts both $\mathbb{P}_\theta$ and the feasible set $\mathbb{P}_\theta \cap \mathbb{P}_p$ to the lower half of the ellipse.

The implication is that, under condition C10.4 that the cost function C is strictly increasing in the injections $(p_j, p_k) = (P_{jk}, P_{kj})$, the nonconvex feasible sets $\mathbb{P}_\theta$ and $\mathbb{P}_\theta \cap \mathbb{P}_p$ coincide with the Parento fronts of their respectively convex hulls, i.e.,

$$\mathbb{P}_\theta = \mathbb{O}(\text{conv}\,\mathbb{P}_\theta), \qquad \mathbb{P}_\theta \cap \mathbb{P}_p = \mathbb{O}(\text{conv}(\mathbb{P}_\theta \cap \mathbb{P}_p)) \tag{10.30}$$

Step 3: condition C10.4 implies that $\mathbb{P}_\theta \cap \mathbb{P}_p = \mathbb{O}(conv(\mathbb{P}_\theta) \cap \mathbb{P}_p)$. Unfortunately the convex hull $\text{conv}(\mathbb{P}_\theta \cap \mathbb{P}_p)$ in (10.30) of the intersection of two sets generally does not have a simple algebraic representation. The feasible set $\text{conv}(\mathbb{P}_\theta) \cap \mathbb{P}_p$ of the relaxation OPF-socp (10.27) is the intersection of two convex hulls and is more amenable to computation. It is however a superset of $\text{conv}(\mathbb{P}_\theta \cap \mathbb{P}_p)$. To illustrate their relation

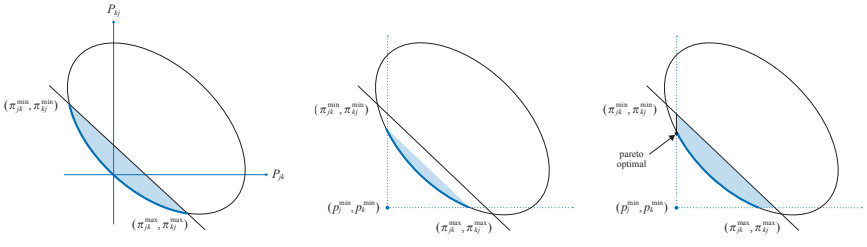
denote the points $P(\theta_{jk}) := \left(P_{jk}(\theta_{jk}), P_{kj}(\theta_{jk})\right)$ attained at $\theta_{jk}^{\min}$ and $\theta_{jk}^{\max}$ by

$$\left(\pi_{jk}^{\min}, \pi_{kj}^{\min}\right) \;:=\; P(\theta_{jk}^{\min}), \qquad \left(\pi_{jk}^{\max}, \pi_{kj}^{\max}\right) \;:=\; P(\theta_{jk}^{\max}) \tag{10.31}$$

The set $\mathbb{P}_\theta$ is the ellipse segment between these two points $\left(\pi_{jk}^{\min}, \pi_{kj}^{\min}\right)$ and $\left(\pi_{jk}^{\max}, \pi_{kj}^{\max}\right)$. As shown in Figure 10.5, the relation between these two convex sets is:

$$\operatorname{conv}(\mathbb{P}_\theta) \cap \mathbb{P}_p \;\supseteq\; \operatorname{conv}\left(\mathbb{P}_\theta \cap \mathbb{P}_p\right)$$

Even though these two sets are generally different, it is clear from the figure that, if



(a) $\operatorname{conv}(\mathbb{P}_\theta)$ (b) $\operatorname{conv}(\mathbb{P}_\theta \cap \mathbb{P}_p)$ (c) $\operatorname{conv}(\mathbb{P}_\theta) \cap \mathbb{P}_p$

Figure 10.5 (a) The set $\operatorname{conv}(\mathbb{P}_\theta)$ is the intersection of the ellipse, including its interior, and a half-space. (b)(c) $\operatorname{conv}(\mathbb{P}_\theta \cap \mathbb{P}_p) \subseteq \operatorname{conv}(\mathbb{P}_\theta) \cap \mathbb{P}_p$. If the cost $C(p)$ is strictly increasing in p_j but independent of p_k then the vertical darkened segment in (c) is part of the Pareto front of the relaxation but only the point on the ellipse is feasible, i.e., in $\mathbb{P}_\theta \cap \mathbb{P}_p$, and hence optimal.

the cost function $C(p)$ is strictly increasing in each p_j (condition C10.4), then they share the same Pareto front, i.e.,

$$\mathbb{O}(\operatorname{conv}(\mathbb{P}_\theta) \cap \mathbb{P}_p) \;=\; \mathbb{O}(\operatorname{conv}(\mathbb{P}_\theta \cap \mathbb{P}_p)) \;=\; \mathbb{P}_\theta \cap \mathbb{P}_p$$

where the last equality follows from (10.30). This proves the first claim of Theorem 10.9.

Step 4: (10.27) *is an SOCP and it is exact.* We now reformulate the feasible set $\operatorname{conv}(\mathbb{P}_\theta) \cap \mathbb{P}_p$ of OPF-socp as the intersection of a second-order cone with several affine sets. First, from (10.29), the solid ellipse including the interior is the set of P satisfying

$$1 \geq \left(P - g_{jk}\mathbf{1}\right)^{\mathsf{T}} (AA^{\mathsf{T}})^{-1} \left(P - g_{jk}\mathbf{1}\right)$$

This is a second-order cone $t^2 \geq (P - g_{jk}\mathbf{1})^{\mathsf{T}}(AA^{\mathsf{T}})^{-1}(P - g_{jk}\mathbf{1})$ intersecting with the affine set $t = 1$. Second the set $\operatorname{conv}(\mathbb{P}_\theta)$ is the intersection of this second-order cone with the following half space (see Figure 10.5(a)):

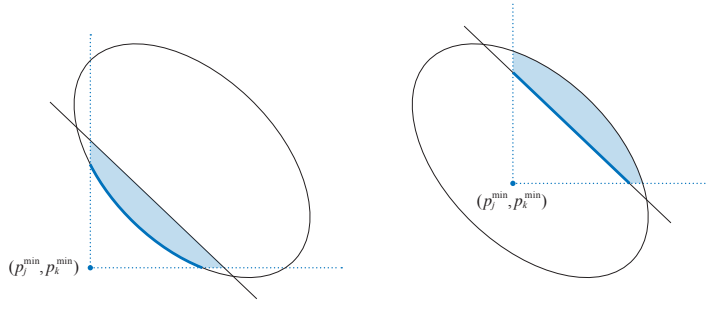
$$P_{kj} \leq \pi_{kj}^{\min} + \frac{\pi_{kj}^{\max} - \pi_{kj}^{\min}}{\pi_{jk}^{\max} - \pi_{jk}^{\min}} \left(P_{jk} - \pi_{jk}^{\min}\right)$$

where $(\pi_{jk}^{\min}, \pi_{kj}^{\min})$ and $(\pi_{jk}^{\max}, \pi_{kj}^{\max})$ are defined in (10.31). Finally intersecting this set with the affine set $\mathbb{P}_p$ produces the feasible set $\text{conv}(\mathbb{P}_\theta) \cap \mathbb{P}_p$ of OPF-socp. Hence the problem (10.27) is indeed an SOCP for the two-bus case.

In summary, the SOCP relaxation of OPF (10.26) enlarges the feasible set $\mathbb{P}_\theta \cap \mathbb{P}_p$ to the convex superset $\text{conv}(\mathbb{P}_\theta) \cap \mathbb{P}_p$. Under condition C10.4, *every* minimizer lies in its Pareto front and hence in the original nonconvex feasible set $\mathbb{P}_\theta \cap \mathbb{P}_p$, as proved in Step 3.

We have hence proved Theorem 10.9 for the two-bus case. □

We illustrate the purpose of condition C10.3. If there are no constraints on the injections p, then SOCP relaxation (10.27) is exact under condition C10.4 due to $\mathbb{P}_\theta = \mathbb{O}(\text{conv}\, \mathbb{P}_\theta)$ in (10.30). As illustrated in Figure 10.6, upper bounds $p^{\max}$ on power injections p do not affect exactness (as long as the problem remains feasible) whereas lower bounds $p^{\min}$ do. Specifically the lower half of the ellipse corresponds



(a) Exact relaxation (b) Inexact relaxation

Figure 10.6 Lower bounds $p^{\min}$ on injections affect exactness of relaxation.

to small $|\theta_{jk}|$ and the upper half of the ellipse corresponds to large $|\theta_{jk}|$ (Exercise 10.7). If the feasible set contains the lower half of the ellipse, as the shaded region in Figure 10.6(a) illustrates, then the Pareto front remains on the ellipse itself, $\mathbb{P}_\theta \cap \mathbb{P}_p = \mathbb{O}(\text{conv}(\mathbb{P}_\theta) \cap \mathbb{P}_p)$, and the relaxation is exact. On the other hand the upper half of the ellipse corresponds to large $|\theta_{jk}|$. The feasible set of OPF may include only the upper half of the ellipse if the lower bounds $p^{\min}$ are large (see Figure 10.6(b)), in which case the Pareto front does not lie on the ellipse and the relaxation is not exact. The purpose of condition C10.3 is to restrict the angle θ_{jk} in order to eliminate the upper half of the ellipse from $\mathbb{P}_\theta$.

We close this subsection with a remark on the importance of tree topology.

Remark 10.5 (Tree topology)**.** The tree topology allows the extension of the argument for a single line to a radial network with multiple lines, in two ways. First let $\mathbb{F}_\theta^{jk}$

denotes the set of branch power flows on each line $(j,k) \in E$:

$$\mathbb{F}_\theta^{jk} := \{\ (P_{jk}, P_{kj}) \mid (P_{jk}, P_{kj}) \text{ satisfies (10.28) for } \theta_{jk}^{\min} \le \theta_{jk} \le \theta_{jk}^{\max}\ \}$$

If the network is a tree, the set $\mathbb{F}_\theta$ of branch power flows on all lines is simply the product set, $\mathbb{F}_\theta = \prod_{(j,k)\in E} \mathbb{F}_\theta^{jk}$, because given any $(\theta_{jk}, (j,k) \in E)$ there is always a (unique up to a reference angle) $(\theta_j, j \in \overline{N})$ that satisfies $\theta_{jk} = \theta_j - \theta_k$. If the network has cycles then this is not possible for some vectors $(\theta_{jk}, (j,k) \in E)$ and $\mathbb{F}_\theta$ is no longer a product set of $\mathbb{F}_\theta^{jk}$.

Second the power injections p are related to the branch flows P by a linear transformation $\mathbb{P}_\theta = A\mathbb{F}_\theta$ for some $(N+1)\times 2M$ dimensional matrix A. Matrix A has full row rank and there is a bijection between P_θ and F_θ (after fixing the reference angle) using the fact that the graph is a tree. We can therefore freely work with either $p \in \mathbb{P}_\theta$ or the corresponding $P \in \mathbb{F}_\theta$ in the proof for a tree network (see [128]).

When the network is not radial or $|V_j|$ are not constants, then the feasible set can be much more complicated than ellipsoids and the simple geometric insight becomes obscure [27, 28, 29, 129]. □

10.5 Other convex relaxations

10.6 Bibliographical notes

Solving OPF through semidefinite relaxation in the bus injection model is first proposed in [130] as a second-order cone program (SOCP) for radial (tree) networks and in [131] as a semidefinite program (SDP) for general networks. The exactness of semidefinite relaxations is first studied in [69]. By defining a new set of variables $v_j := |V_j|^2$, $R_{jk} := |V_j||V_k|\cos(\theta_j - \theta_k)$, and $I_{jk} := |V_j||V_k|\sin(\theta_j - \theta_k)$ where $\theta_j := \angle V_j$, [130] rewrites the bus injection model (4.27) in the polar form as a set of linear equations in these new variables and the following quadratic equations:

$$v_j v_k = R_{jk}^2 + I_{jk}^2$$

Relaxing these equalities to $v_j v_k \ge R_{jk}^2 + I_{jk}^2$ enlarges the solution set to a second-order cone that is equivalent to $\mathbb{W}_G^+$ in this chapter. Partial matrices and their completions are studied in [112, 114, 115]. Exploiting graph sparsity to simplify the SDP relaxation of OPF through chordal extension is first proposed in [132, 133, 134] and analyzed in [113, 135, 33]. Theorem 10.1 is from [33] and Corollary 10.2 is from [113]). The sufficient condition on angle differences for exact SOCP relaxation in Chapter 10.4 is from [125, 128] and our proof mostly follows that in [128]. The result in Chapter 10.4 assumes the voltage magnitudes are fixed and ignores reactive powers. These

assumptions are relaxed in [129] although, without these assumptions, the feasible set may no longer be a convex surface that is the Pareto front of its relaxation.

The semidefinite relaxation of three-phase OPF in Chapter **??** follows the idea in [105, 136].

Simulations [78] show that the SDP relaxation of OPF is often exact and adding valid inequalities and bound tightening can further reduce the optimality gap to within 1%, though [118] also reports instances where the optimality gap of SDP relaxation is large.

10.7 Problems

Chapter 10.1

Exercise 10.1 (Solution recovery)**.** Given a partial matrix X_F in $\mathbb{X}_F$ defined in (10.7c) and a vector $x \in \mathbb{C}^n$ recovered from X_F using (10.10), show that x satisfies $[X_F]_{jk} = x_j \overline{x}_k$ and $x^{\mathsf{H}} C_l x \le b_l$, $l = 1, \ldots, L$.

Chapter 10.2

Exercise 10.2 (Loss minimization)**.** In this problem we formulate and solve a simple nonconvex loss minimization problem. A generator supplies a load through a transmission line modeled as a series admittance $y := g + \mathbf{i}b = 1/(r + \mathbf{i}x)$ with $g > 0$ and $b < 0$. The voltage at the generator (reference) bus is fixed at $V_0 := 1\angle 0^\circ$ p.u. The required load power is $s = p + \mathbf{i}q = |s|e^{\mathbf{i}\phi}$ with $p > 0$ specified, i.e., $-s$ is the power injection at the load bus. Let the load voltage be $V := ve^{\mathbf{i}\theta}$.

1 Show that the active line loss $r|I|^2 = g\left|1 - ve^{\mathbf{i}\theta}\right|^2$.
2 Fix v and p. Formulate OPF as minimization over (θ, ϕ) of the active line loss.
3 Reformulate OPF as an unconstrained minimization $\min_\phi f(\phi)$ over ϕ only.
4 Show that the unique minimizer of $f(\phi)$ over $(-\pi/2, \pi/2)$ is $\phi_{\min} := \tan^{-1}(-b(1 - v^2)/p)$, even though the original OPF problem in part 2 is nonconvex.
5 Suppose v is also an optimization variable and assume $p < g$. Show that

$$\phi_{\min} \;=\; \tan^{-1}(-b/g) \;=\; \tan^{-1}(x/r), \qquad\qquad v_{\min} \;=\; \sqrt{1 - p/g}$$

is an isolated local minimizer[3] of $f(\phi, v)$ over $\phi \in (-\pi/2, \pi/2)$ and $v > 0$, by showing $\nabla^2 f(\phi_{\min}, v_{\min})$ is positive definite.
6 Is $(\phi_{\min}, v_{\min})$ a global minimizer over $\phi \in (-\pi/2, \pi/2)$ and $v > 0$? (Hint: What is $f(\phi_{\min}, v_{\min})$ and the load voltage $v_{\min} e^{\mathbf{i}\theta_{\min}}$? Interpret.)

[3] There is a neighborhood of $(\phi_{\min}, v_{\min})$ that contains no other minimizer.

Chapter 10.3

Exercise 10.3 (Linear separability)**.** The linear separability condition C10.1' requires that some of power injections be unconstrained even though in practice they are always bounded. The next exercise shows that, for a convex problem, C10.1' is equivalent to requiring that the finite bounds on these power injections be inactive at optimality (as opposed to removing these finite bounds but optimal solutions of the unconstrained problem turn out to satisfy these bounds).

Consider the two problems:

$$\hat{x} \in \arg\min_{x \in X} f(x) \tag{10.32a}$$

$$x^* \in \arg\min_{x \in X} f(x) \quad \text{s. t.} \quad g(x) \leq 0 \tag{10.32b}$$

where $X \subseteq \mathbb{R}^n$ is convex and $g : \mathbb{R}^n \to \mathbb{R}^m$ is a convex function. We assume the minimizers $\hat{x}$ and x^* exist.

1 Suppose f is strictly convex. Show that $g(\hat{x}) < 0$ if and only if $g(x^*) < 0$ in which case $f(\hat{x}) = f(x^*)$.
2 Show that if f is nonconvex, then it is possible that both $g(x^*) < 0$ and $g(\hat{x}) > 0$, in which case $f(\hat{x}) < f(x^*)$.

Chapter 10.4

The next few problems use a two-bus example to illustrate the geometry of solutions to the polar form power flow equations, convex relaxation and its exactness [125, 128].

Exercise 10.4 (Ellipsoid)**.** An ellipsoid in $\mathbb{R}^k$ (without the interior) in standard form are the points $x \in \mathbb{R}^k$ that satisfy

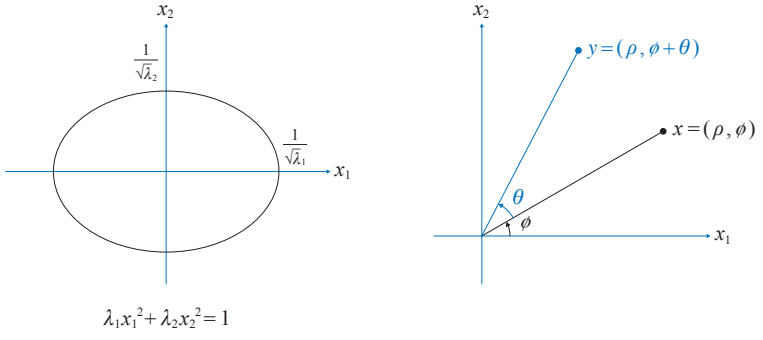
$$x^{\mathsf{T}} \Lambda x = 1 \tag{10.33a}$$

for a real positive definite diagonal matrix $\Lambda \succ 0$. The center of the ellipsoid is the origin 0 and the k principal axes are the coordinate axes. This is illustrated in Figure 10.7(a) for $k = 2$. In general the set of points $x \in \mathbb{R}^k$ that satisfy

$$(x-c)^{\mathsf{T}} M (x-c) = \left\| M^{1/2}(x-c) \right\|_2^2 = 1 \tag{10.33b}$$

is an ellipse if $c \in \mathbb{R}^n$ and $M \succ 0$ is a real (symmetric) positive definite matrix. The center of the ellipsoid is c and the k principal axes are the k eigenvectors of M. In this exercise, we show that a general ellipsoid (10.33b) can be obtained through simple transformations of the standard form ellipsoid (10.33a).

Given a standard form ellipsoid $x \in \mathbb{R}^k$ that satisfies (10.33a).



(a) Ellipsoid in $\mathbb{R}^2$ (b) Rotation in $\mathbb{R}^2$

Figure 10.7 Exercises 10.4 and 10.5.

1 *Translation*: Let $y := x + x_0 \in \mathbb{R}^k$. Show that y is a standard form ellipsoid with its center translated to x_0. Illustrate y for $k = 2$.
2 *Scaling*: Let $y := ax$ where $a \in \mathbb{R}$ is nonzero. Show that y is a standard form ellipsoid with its size scaled by a in all the k dimensions. Illustrate y for $k = 2$.
3 *Scaling and rotation*: Let $y := Ax$. Show that y is an ellipsoid as long as A is real and invertible, i.e., y satisfies (10.33b) with a real (symmetric) positive definite matrix M.
4 *Inverse scaling and rotation*: Show that a general ellipsoid y that satisfies (10.33b) with the origin $c = 0$ as its center is a standard form ellipsoid x scaled and rotated by a matrix U, i.e., $y = Ux$. Derive U.

Exercise 10.5 (Rotation in $\mathbb{R}^2$)**.** Show that $y = R(\theta)x$ is a rotation of x by an angle θ in $\mathbb{R}^2$ where

$$R(\theta) := \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

as illustrated in Figure 10.7(b).

1 Show that $R^{-1}(\theta) = R(-\theta) = R^{\mathsf{T}}(\theta)$.
2 Show that $R(\theta)$ is normal and find its spectral decomposition for $\theta \neq 0$.
3 Suppose x is a standard form ellipse in $\mathbb{R}^2$ that satisfies (10.33a). Show that $y := R(\theta)x$ is an ellipse, i.e., y satisfies (10.33b) with a real (symmetric) positive definite matrix M.

Exercise 10.6 (Geometric insight [125, 128])**.** Fix a line (j, k) so we can omit the

subscript in g_{jk}, b_{jk}. Show that (10.28) can be rewritten as

$$P = \begin{bmatrix} P_{jk} \\ P_{kj} \end{bmatrix} = \sqrt{2} \begin{bmatrix} \cos 45^\circ & \sin 45^\circ \\ -\sin 45^\circ & \cos 45^\circ \end{bmatrix} \cdot \hat{P} + g \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{10.34a}$$

where $\hat{P} \in \mathbb{R}^2$ satisfies

$$1 = \left\| \begin{bmatrix} \cos\theta_{jk} \\ \sin\theta_{jk} \end{bmatrix} \right\|^2 = \hat{P}^{\mathsf{T}} \begin{bmatrix} \frac{1}{b^2} & 0 \\ 0 & \frac{1}{g^2} \end{bmatrix} \hat{P} \tag{10.34b}$$

This says that $\hat{P}$ defined by (10.34) is a standard form ellipse centered at the origin with its major axis of length $2b$ on the x-axis and its minor axis of length $2g$ on the y-axis. P is the ellipse obtained from $\hat{P}$ by scaling it by $\sqrt{2}$, rotating it by -45°, and shifting its center to (g, g).

Exercise 10.7 (Geometric insight [125, 128])**.** Show that the two-bus network given by (10.28), reproduced here with subscript jk dropped:

$$p_1 = p_1(\theta) := g - g\cos\theta - b\sin\theta \tag{10.35a}$$

$$p_2 = p_2(\theta) := g - g\cos\theta + b\sin\theta \tag{10.35b}$$

We have shown that (p_1, p_2) forms an ellipse. Draw the ellipse and indicate on the ellipse values for θ where p_1 and p_2 attain minimum or maximum values. Conclude that the "lower half" of the ellipse corresponds to small $|\theta|$ and the "upper half" corresponds to large $|\theta|$.

Exercise 10.8 (Geometric insight [125, 128])**.** Consider the 2-bus network in Exercise 10.7. Let $x := (p_1, p_2, \theta)$. Let $c(p_1, p_2)$ be a cost function that is strictly increasing in (p_1, p_2), e.g., $c(p_1, p_2) := p_1 + p_2$.

1 Consider the OPF problem:

$$\min_x \; c(p_1, p_2) \quad \text{s.t.} \quad x \in X_1 \tag{10.36}$$

where the only constraint is the power flow equation:

$$X_1 := \{x := (p_1, p_2, \theta) \; : \; x \text{ satisfies (10.35)}\}$$

The feasible set is nonconvex because it is an ellipse without its interior. Consider the convex relaxation:

$$\min_x \; c(p_1, p_2) \quad \text{s.t.} \quad x \in \text{conv}(X_1) \tag{10.37}$$

Explain why the relaxation is exact, i.e., an optimal x^* for (10.37) is also optimal for (10.36).

2 Consider the constraints on injections (p_1, p_2) and constraints on θ:

$$X_2 := \{x := (p_1, p_2, \theta) \in \mathbb{R}^3 \ : \ \theta^{\min} \le \theta \le \theta^{\max}\}$$
$$X_3 := \{x := (p_1, p_2, \theta) \in \mathbb{R}^3 \ : \ p_j^{\min} \le p_j \le p_j^{\max}, \ j = 1,2\}$$

Consider the OPF:

$$\min_x \ c(p_1, p_2) \quad \text{s.t.} \quad x \in X_1 \cap X_2 \cap X_3 \tag{10.38}$$

and its convex relaxation:

$$\min_x \ c(p_1, p_2) \quad \text{s.t.} \quad x \in \text{conv}(X_1 \cap X_2) \cap X_3 \tag{10.39}$$

Indicate the feasible sets of (10.38) and (10.39) projected onto (p_1, p_2) plane, and explain why lower bounds $(p_1^{\min}, p_2^{\min})$ on the injections (p_1, p_2) affect the exactness of SOCP relaxation, but not the upper bounds $(p_1^{\max}, p_2^{\max})$.

3 Explain why limiting $|\theta|$ to $[\theta^{\min}, \theta^{\max}]$ can ensure exact relaxation as long as (recall that $g > 0, b < 0$)

$$\tan^{-1}\left(\frac{b}{g}\right) \ \le \ \theta^{\min} \ < \ \theta^{\max} \ \le \ \tan^{-1}\left(\frac{-b}{g}\right)$$

Exercise 10.9 (Condition C10.4 and Pareto front)**.** In general, a point x^* is Pareto optimal over a convex set $A \subseteq \mathbb{R}^k$ if and only if it $x^* = \arg\min_{x \in A} c^\mathsf{T} x$ for some nonzero $c \ge 0$.

1 Show that, for the two-bus network in Exercise 10.7, $\mathbb{O}(\text{conv}(\mathbb{P}_\theta \cap \mathbb{P}_p) \supseteq \mathbb{O}(\text{conv}(\mathbb{P}_\theta \cap \mathbb{P}_p))$ if condition C10.4 does not hold.
2 Show that if condition C10.4 holds, then we can define a Pareto optimal x^* as $x^* = \arg\min_{x \in A} c^\mathsf{T} x$ for some $c > 0$ and $\mathbb{O}(\text{conv}(\mathbb{P}_\theta) \cap \mathbb{P}_p) = \mathbb{O}(\text{conv}(\mathbb{P}_\theta \cap \mathbb{P}_p))$.

Exercise 10.10 (Convex hull and Pareto front)**.** Let $B, C \subseteq \mathbb{R}^k$ be arbitrary sets, $D := \{x \in \mathbb{R}^k | Mx \le c\}$ be an affine set, and M a matrix and b a vector of appropriate dimensions.

1 $\text{conv}(MB) = M\,\text{conv}(B)$ and $\text{conv}(B \times C) = \text{conv}(B) \times \text{conv}(C)$ where for any sets $A_1, A_2 \subseteq \mathbb{R}^k$, $(x^1, x^2) \in A_1 \times A_2$ if and only if $x^1 \in A_1$ and $x^2 \in A_2$.
2 Suppose B and C are convex and a point is Pareto optimal over a set if and only if it minimizes $c^\mathsf{T} x$ over the set for some nonzero $c \ge 0$. Then $\mathbb{O}(MB) = M\,\mathbb{O}(B)$ and $\mathbb{O}(B \times C) = \mathbb{O}(B) \times \mathbb{O}(C)$.
3 If $B = \mathbb{O}(\text{conv}\ B)$ then $B \cap D \subseteq \mathbb{O}(\text{conv}(B) \cap D)$.

Chapter **??**.

Exercise 10.11 (Lemma 18.1 [136])**.**

11 Semidefinite relaxations: BFM

In Chapter 10 we study the semidefinite relaxation of OPF in the bus injection model. In this chapter we continue our study in the branch flow model for radial networks. In Chapter 11.1 we formulate SOCP relaxation and prove its equivalence to the SOCP relaxation in BIM. In Chapters 11.2 and 11.3 we prove sufficient conditions for exact relaxation for radial networks.

11.1 SOCP relaxation

We first focus on the DistFlow model studied in Chapter 5.1.3 where $z^s_{jk} = z^s_{kj}$ and $z^m_{jk} = z^m_{kj} = 0$ for each line $(j,k) \in E$. We formulate SOCP relaxation of OPF under these two assumptions in Chapter 11.1.1 and prove its equivalence to the SOCP relaxation in the bus injection model in Chapter 11.1.2. Then we extend SOCP relaxation to general radial networks without these assumptions in Chapter 11.1.3.

11.1.1 DistFlow model

The DistFlow model of Chapter 5.1.3 assumes the series impedances $z^s_{jk} = z^s_{kj}$ of each line (j,k) are equal in each direction (assumption C5.1) and shunt admittances are zero $z^m_{jk} = z^m_{kj} = 0$. It is a reasonable model for single-phase radial networks, but requires approximations to incorporate transformer models (see discussions in Chapter 5.1.1). These two assumptions allow us to assume the network graph $G = (\overline{N}, E)$ is directed and includes branch variables in only one direction (see Chapter 5.1.3 for details). We denote a line in E from bus j to bus k either by $(j,k) \in E$ or $j \to k$. It is characterized by its series impedance $z_{jk} := z^s_{jk}$. Without loss of generality we take bus 0 as the root of the tree.

Consider a single-phase radial network $G = (\overline{N}, E)$ with $N+1$ buses and $M = N$ lines modeled by DistFlow equation (5.9) with up orientation (all lines point *towards*

bus 0), reproduced here:

$$S_{jk} = \sum_{i:i\to j} \left(S_{ij} - z_{ij}\ell_{ij}\right) + s_j, \qquad j \in \overline{N} \tag{11.1a}$$

$$v_j - v_k = 2\,\mathrm{Re}\left(\bar{z}^s_{jk} S_{jk}\right) - |z_{jk}|^2 \ell_{jk}, \qquad j \to k \in E \tag{11.1b}$$

$$v_j \ell_{jk} = |S_{jk}|^2, \qquad j \to k \in E \tag{11.1c}$$

where $k := k(j)$ in (11.1a) denotes the node adjacent to j on the unique path from bus j to bus 0, with the udnerstanding that $S_{jk} := 0$ when $j = 0$ and $S_{ij} = 0$, $\ell_{ij} = 0$ when j is a leaf node.[1] The injection, voltage and line limits are:

$$s_j^{\min} \le s_j \le s_j^{\max}, \quad v_j^{\min} \le v_j \le v_j^{\max}, \quad \ell_{jk} \le \ell_{jk}^{\max}, \quad j \in \overline{N},\ (j,k) \in E \tag{11.2}$$

The model (11.1) includes only voltage and power sources whose controllable variables are v_j and s_j respectively. See Remark 9.5 of Chapter 9.2 on how to incorporate current sources and impedances. Denote by $(s,v) := (s_j, v_j, j \in \overline{N}) \in \mathbb{R}^{3(N+1)}$ the bus injections and squared voltage magnitudes, and by $(\ell, S) := (\ell_{jk}, S_{jk}, j \to k \in E) \in \mathbb{R}^{3M}$ the squared line current magnitudes and line powers. The vector v includes v_0 and s includes s_0. Let $x := (s,v,\ell,S)$ in $\mathbb{R}^{3(2N+1)}$ since G is a tree.

Let $C(x)$ be a cost function. Let the feasible set be

$$\mathbb{X}_{\mathrm{df}} := \{x := (s,v,\ell,S) \in \mathbb{R}^{6N+3} \mid x \text{ satisfies (11.1)(11.2)}\} \tag{11.3a}$$

The OPF (9.20) formulated in Chapter 9.2 (but with a different graph orientation) is
OPF:

$$\min_x\ C(x) \ \text{ subject to } \ x \in \mathbb{X}_{\mathrm{df}} \tag{11.3b}$$

To avoid triviality we will assume unless otherwise specified that OPF (11.3) is feasible. The constraints (11.1a)(11.1b) are linear in x. The constraint (11.1c) is however quadratic in x, making the feasible set of OPF (11.3) nonconvex. Relaxing the equality in (11.1c) into inequality

$$v_j \ell_{jk} \ge |S_{jk}|^2, \qquad j \to k \in E \tag{11.4}$$

results in a (convex) second-order cone. Define

$$\mathbb{X}^+_{\mathrm{df}} := \{x : (s,v,\ell,S) \in \mathbb{R}^{6N+3} \mid x \text{ satisfies (11.1a)(11.1b)(11.4)(11.2)}\} \tag{11.5a}$$

Then an SOCP relaxation of OPF (11.3) is:
OPF-socp:

$$\min_x\ C(x) \ \text{ subject to } \ x \in \mathbb{X}^+_{\mathrm{df}} \tag{11.5b}$$

We say that OPF-socp (11.5) is *exact* if *every* optimal solution x^{socp} of (11.5) attains equalities in (11.4) and hence is an optimal solution of OPF (11.3). This is convenient because it ensures that any algorithm that solves an exact relaxation always produces

[1] A node $j \in N$ is a *leaf node* if there is no i such that $i \to j \in E$.

a globally optimal solution to the OPF problem. This notion of strong exactness is however unnecessary under the sufficient exactness conditions of Chapters 11.2 and 11.3 for radial networks; see Remark 11.1 after Theorem 11.3 and Remark 11.3 after Theorem 11.5. These conditions guarantee that an optimal solution to OPF can be recovered from *any* optimal solution x^{socp} of OPF-socp whether or not x^{socp} attains equalities in (11.4).

The next result from [34] shows that, when the SOCP relaxation (in fact, *any* convex relaxation) of (11.3) is exact in the strong sense defined above, then the optimal solution is unique.

Theorem 11.1 (Unique optimal of SOCP relaxation)**.** Suppose the network graph G is a tree and the cost C is a convex function. If OPF-socp (11.5) is exact then its optimal solution is unique.

Proof Suppose $\hat{x}$ and $\tilde{x}$ are distinct optimal solutions of the relaxation OPF-socp (11.5). Since the feasible set of OPF-socp is convex the point $x := (\hat{x} + \tilde{x})/2$ is also feasible for OPF-socp. Since the cost function C is convex and both $\hat{x}$ and $\tilde{x}$ are optimal for (11.5), x is also optimal for (11.5). The exactness of OPF-socp then implies that x attains equality in (11.4). This contradicts Theorem 5.1 that shows that if $\hat{x}$ and $\tilde{x}$ are feasible, then no convex combination of $\hat{x}$ and $\tilde{x}$ can be feasible. □

11.1.2 Equivalence

The single-phase OPF (11.3) is equivalent to the single-phase OPF problem (9.9) or (9.16) in the bus injection model because their feasible sets $\mathbb{X}_{\text{df}}$ and $\mathbb{V}$ respectively are equivalent by Theorem 5.2. In this section we show that their SOCP relaxations are equivalent as well by establishing a bijection between the feasible sets of these relaxations.

The equivalence of the SOCP relaxations in these two models rests on the equivalence of their feasible sets. Recall that any sets A and B are *equivalent*, denoted by $A \equiv B$, if there is a bijection between them. When there is a one-one correspondence $g : A \to B$ between their feasible sets, a feasible point x is optimal for one problem if and only if $g(x)$ is optimal for the other problem. We now make this precise.

Recall from Chapter 10.2.1 that the SOCP relaxation (10.20c) of OPF in BIM is the minimization of $C(W_G)$ over Hermitian partial matrices $W_G \in \mathbb{C}^{2M+N+1}$ subject to operational and 2×2 psd constraints. The operational constraints are the injection limits, voltage limits, and line limits. In terms of the partial matrix W_G, they are respectively: (substituting $|V_j|^2 = [W_G]_{jj}$ and $V_j V_k^{\mathsf{H}} = [W_G]_{jk}$ into (9.8) (9.4b)(9.4c)):

$$s_j^{\min} \;\le\; \sum_{k:j\sim k} \bar{y}_{jk}^s \left([W_G]_{jj} - [W_G]_{jk}\right) \;\le\; s_j^{\max}, \qquad j \in \overline{N} \tag{11.6a}$$

$$v_j^{\min} \;\le\; [W_G]_{jj} \;\le\; v_j^{\max}, \qquad j \in \overline{N} \tag{11.6b}$$

$$\left|y_{jk}^s\right|^2 \left([W_G]_{jj} + [W_G]_{kk} - [W_G]_{jk} - [W_G]_{kj}\right) \;\le\; \ell_{jk}^{\max}, \quad j \to k \in E \tag{11.6c}$$

The 2×2 psd constraint $W_G(j,k) \succeq 0$, $(j,k) \in E$, is equivalent to

$$[W_G]_{jk} = [W_G]_{kj}^{\mathsf{H}}, \quad [W_G]_{jj} > 0, \quad [W_G]_{kk} > 0, \quad [W_G]_{jj}[W_G]_{kk} \ge \left|[W_G]_{jk}\right|^2, \quad (j,k) \in \tag{11.6d}$$

Then the feasible set of the SOCP relaxation (10.20c) of OPF in BIM is

$$\mathbb{W}_G^+ \;:=\; \{\, W_G \in \mathbb{C}^{2M+N+1} \mid W_G \text{ satisfies (11.6)} \,\} \tag{11.7a}$$

and the SOCP relaxation is

$$\min_{W_G} \quad C(W_G) \qquad \text{s.t.} \qquad W_G \in \mathbb{W}_G^+ \tag{11.7b}$$

The feasible set of OPF-socp (11.5) in BFM is equivalent to that of (11.7) in BIM.

Theorem 11.2 (Equivalence of SOCPs)**.** $\mathbb{X}_{\text{df}}^+ \equiv \mathbb{W}_G^+$.

The theorem implies that there is a bijection $g : \mathbb{W}_G^+ \to \mathbb{X}_{\text{df}}^+$. If the cost function in the SOCP relaxation (11.5) in BFM and that in (11.7) in BIM are equivalent, i.e., $C(W_G) = C\left(g(W_G)\right)$, then these SOCP relaxations are equivalent problems in the sense that W_G^{opt} is optimal for (11.7) if and only if $x^{\text{opt}} := g(W_G^{\text{opt}})$ is optimal for (11.5).

The proof of Theorem 11.2 below constructs a linear mapping $g : \mathbb{W}_G^+ \to \mathbb{X}_{\text{df}}^+$, motivated by the factorization $W = VV^{\mathsf{H}}$ of the psd rank-1 completion W of the partial matrix W_G when W_G is psd rank-1. Define the linear mapping $g : \mathbb{W}_G^+ \to \mathbb{X}_{\text{df}}^+$ with $x := (s, v, \ell, S) = g(W_G)$ where

$$\begin{aligned} s_j \;&:=\; \sum_{k:j\sim k} \bar{y}_{jk}^s \left([W_G]_{jj} - [W_G]_{jk}\right), \\ &=\; \sum_{i:i\to j} \bar{y}_{ij}^s \left([W_G]_{jj} - [W_G]_{ji}\right) + \sum_{k:j\to k} \bar{y}_{jk}^s \left([W_G]_{jj} - [W_G]_{jk}\right), \quad j \in \overline{N} \end{aligned} \tag{11.8a}$$

$$v_j \;:=\; [W_G]_{jj}, \qquad j \in \overline{N} \tag{11.8b}$$

$$\ell_{jk} \;:=\; |y_{jk}^s|^2 \left([W_G]_{jj} + [W_G]_{kk} - [W_G]_{jk} - [W_G]_{kj}\right), \qquad j \to k \in E \tag{11.8c}$$

$$S_{jk} \;:=\; \bar{y}_{jk}^s \left([W_G]_{jj} - [W_G]_{jk}\right), \qquad j \to k \in E \tag{11.8d}$$

and the mapping $g^{-1} : \mathbb{X}_{\rm df}^{+} \to \mathbb{W}_{G}^{+}$ with $W_G = g^{-1}(x)$ where

$$[W_G]_{jj} := v_j, \qquad j \in \overline{N} \tag{11.9a}$$

$$[W_G]_{jk} := v_j - \bar{z}_{jk}^{s} S_{jk} = [W_G]_{kj}^{\mathsf{H}}, \qquad j \to k \in E \tag{11.9b}$$

Note that in (11.8a) the first summation over lines $i \to j$ is $[W_G]_{jj} - [W_G]_{ji}$, not $\left([W_G]_{ii} - [W_G]_{ij}\right)$. The proof below establishes that g and g^{-1} are indeed inverses of each other. By restricting these mappings g and g^{-1} to subsets $\mathbb{W}_G \subseteq \mathbb{W}_G^{+}$ and $\mathbb{X}_{\rm df} \subseteq \mathbb{X}_{\rm df}^{+}$, the theorem immediately implies the equivalence of $\mathbb{X}_{\rm df} \equiv \mathbb{W}_G$ and hence the equivalence of single-phase OPF (11.3) in BFM and the OPF (9.9) or (9.16) in BIM (since $\mathbb{W}_G \equiv \mathbb{V}$).

Since we assume $z_{jk}^{m} = z_{kj}^{m} = y_{jk}^{m} = y_{kj}^{m} = 0$, we often omit the superscript s in z_{jk}^{s} and y_{kj}^{s}.

Proof of Theorem 11.2.

We will prove that g and g^{-1} are indeed inverses of each other in three steps: (1) g maps every point $W_G \in \mathbb{W}_G^{+}$ to a point in $\mathbb{X}_{\rm df}^{+}$; (2) g^{-1} maps every point $x \in \mathbb{X}_{\rm df}^{+}$ to a point in $\mathbb{W}_G^{+}$; and (3) $g(g^{-1}(x)) = x$ and $g^{-1}(g(W_G)) = W_G$. This defines a bijection between $\mathbb{W}_G^{+}$ and $\mathbb{X}_{\rm df}^{+}$ and establishes $\mathbb{W}_G^{+} \equiv \mathbb{X}_{\rm df}^{+}$.

Step 1: $x := g(W_G) \in \mathbb{X}_{df}^{+}$. Given a $W_G \in \mathbb{W}_G^{+}$, we have to prove $x := g(W_G)$ satisfies (11.1a) (11.1b) (11.4) (11.2). Clearly (11.2) follows from (11.8) and (11.6). To prove (11.1a), we have for $j \in \overline{N}$

$$\begin{aligned}
&\sum_{i:i\to j} \left(S_{ij} - z_{ij}\ell_{ij}\right) + s_j \\
&= \sum_{i:i\to j} \left(\bar{y}_{ij}\left([W_G]_{ii} - [W_G]_{ij}\right) - \bar{y}_{ij}\left([W_G]_{ii} + [W_G]_{jj} - [W_G]_{ij} - [W_G]_{ji}\right)\right) + s_j \\
&= \sum_{i:i\to j} \left(-\bar{y}_{ij}\left([W_G]_{jj} - [W_G]_{ji}\right)\right) + \sum_{i:i\to j} \bar{y}_{ji}\left([W_G]_{jj} - [W_G]_{ji}\right) + \sum_{k:j\to k} \bar{y}_{jk}\left([W_G]_{jj} - [W_G]_{jk}\right) \\
&= \sum_{k:j\to k} S_{jk}
\end{aligned}$$

where the last equality uses $y_{ij} = y_{ji}$ by assumption C5.1. To prove (11.1b), we have for $j \to k \in E$

$$\begin{aligned}
2\,\mathrm{Re}\left(\bar{z}_{jk}S_{jk}\right) - |z_{jk}|^2\ell_{jk} &= 2\,\mathrm{Re}\left([W_G]_{jj} - [W_G]_{jk}\right) - \left([W_G]_{jj} + [W_G]_{kk} - [W_G]_{jk} - [W_G]_{kj}\right) \\
&= \left([W_G]_{jj} - [W_G]_{kk}\right) - [W_G]_{jk}^{\mathsf{H}} + [W_G]_{kj} \\
&= v_j - v_k
\end{aligned}$$

where the last equality follows because the partial matrix W_G is Hermitian. Finally to prove (11.4), for each $j \to \in E$, we have from (11.6d) $[W_G]_{jj}[W_G]_{kk} \geq |[W_G]_{jk}|^2$.

Hence

$$\begin{aligned} v_j \ell_{jk} &= \left|y_{jk}\right|^2 \, [W_G]_{jj} \left([W_G]_{jj} + [W_G]_{kk} - [W_G]_{jk} - [W_G]_{kj}\right) \\ &\geq \left|y_{jk}\right|^2 \left([W_G]_{jj}^2 + \left|[W_G]_{jk}\right|^2 - [W_G]_{jj}[W_G]_{jk} - [W_G]_{jj}[W_G]_{jk}^{\mathsf{H}}\right) \\ &= \left|S_{jk}\right|^2 \end{aligned} \tag{11.10}$$

as desired. Hence g maps every $W_G \in \mathbb{W}_G^+$ to an $x \in \mathbb{X}_{\mathrm{df}}^+$.

Step 2: $W_G := g^{-1}(x) \in \mathbb{W}_G^+$. Given an $x \in \mathbb{X}_{\mathrm{df}}^+$, we have to prove that $W_G := g^{-1}(x)$ satisfies (11.6). Clearly (11.9a) and the voltage limit in (11.2) implies (11.6b).

To prove (11.6a), we have for each $j \in N^+$

$$\begin{aligned} \sum_{k:(j,k)\in E} \bar{y}_{jk}^s \left([W_G]_{jj} - [W_G]_{jk}\right) &= \sum_{i:i\to j} \bar{y}_{ji}^s \left([W_G]_{jj} - [W_G]_{ji}\right) + \sum_{k:j\to k} \bar{y}_{jk}^s \left([W_G]_{jj} - [W_G]_{jk}\right) \\ &= \sum_{i:i\to j} \bar{y}_{ij}^s \left(v_j - \left(v_i - \bar{z}_{ij}^s S_{ij}\right)^{\mathsf{H}}\right) + \sum_{k:j\to k} \bar{y}_{jk}^s \left(v_j - \left(v_j - \bar{z}_{jk}^s S_{jk}\right)^{\mathsf{H}}\right) \\ &= \sum_{k:j\to k} S_{jk} - \sum_{i:i\to j} \bar{y}_{ij}^s \left(v_i - v_j - z_{ij}^s S_{ij}^{\mathsf{H}}\right) \\ &= \sum_{k:j\to k} S_{jk} - \sum_{i:i\to j} \bar{y}_{ij}^s \left(2\,\mathrm{Re}(\bar{z}_{ij}^s S_{ij}) - \left|z_{ij}^s\right|^2 \ell_{ij} - z_{ij}^s S_{ij}^{\mathsf{H}}\right) \end{aligned}$$

where the second equality follows from (11.9) and $y_{ji} = y_{ij}$ by assumption C5.1, and the last equality follows from (11.1b). But

$$\left(2\,\mathrm{Re}(\bar{z}_{ij}^s S_{ij}) - z_{ij}^s S_{ij}^{\mathsf{H}}\right) = \left(\bar{z}_{ij}^s S_{ij} + z_{ij}^s S_{ij}^{\mathsf{H}}\right) - z_{ij}^s S_{ij}^{\mathsf{H}} \;=\; \bar{z}_{ij}^s S_{ij}$$

and hence

$$\sum_{k:(j,k)\in E} \bar{y}_{jk}^s \left([W_G]_{jj} - [W_G]_{jk}\right) = \sum_{k:j\to k} S_{jk} - \sum_{i:i\to j} \left(S_{ij} - z_{ij}^s \ell_{ij}\right) \;=\; s_j$$

where the last equality follows from (11.1a). This and the injection limits in (11.2) imply (11.6a). To prove (11.6c), we have for each $(j,k) \in E$, from (11.9),

$$\begin{aligned} \left|y_{jk}\right|^2 \left([W_G]_{jj} + [W_G]_{kk} - [W_G]_{jk} - [W_G]_{kj}\right) &= \left|y_{jk}\right|^2 \left(v_j + v_k - \left(v_j - \bar{z}_{jk}^s S_{jk}\right) - \left(v_j - \bar{z}_{jk}^s S_{jk}\right)^{\mathsf{H}}\right) \\ &= \left|y_{jk}\right|^2 \left(-v_j + v_k + \bar{z}_{jk}^s S_{jk} + z_{jk}^s S_{jk}^{\mathsf{H}}\right) \\ &= \ell_{jk} \end{aligned}$$

where last equality follows from (11.1b). This and the line limit in (11.2) imply (11.6c). Finally to prove (11.6d), note that $[W_G]_{jk} = [W_G]_{kj}^{\mathsf{H}}$, $[W_G]_{jj} > 0$, and $[W_G]_{kk} > 0$

follow directly from (11.9). Furthermore

$$\begin{aligned}
[W_G]_{jj}[W_G]_{kk} - \left|[W_G]_{jk}\right|^2 &= v_j v_k - \left|v_j - \bar{z}^s_{jk} S_{jk}\right|^2 \\
&= v_j v_k - \left(v_j^2 + \left|z^s_{jk}\right|^2 |S_{jk}|^2 - 2 v_j \,\mathrm{Re}\left(\bar{z}^s_{jk} S_{jk}\right)\right) \\
&= v_j \left(v_k - v_j + 2\,\mathrm{Re}\left(\bar{z}^s_{jk} S_{jk}\right)\right) - \left|z^s_{jk}\right|^2 |S_{jk}|^2 \\
&= \left|z^s_{jk}\right|^2 \left(v_j \ell_{jk} - |S_{jk}|^2\right) \;\geq\; 0
\end{aligned}$$

where last equality follows from (11.1b) and the last inequality follows from (11.4). Therefore $W_G(j,k) \geq 0$ for all $(j,k) \in E$, as desired. This shows that g^{-1} maps every $x \in \mathbb{X}^+_{\rm df}$ to a $W_G \in \mathbb{W}^+_G$.

Step 3: $g(g^{-1}(x)) = x$ and $g^{-1}(g(W_G)) = W_G$. The proof uses (11.8)(11.9)(11.1a)(11.1b). It follows a similar argument used in Steps 1 and 2, and is omitted. This completes the proof that g and g^{-1} are indeed inverses of each other and establishes $\mathbb{W}^+_G \equiv \mathbb{X}^+_{\rm df}$.

This completes the proof of Theorem 11.2. □

11.1.3 General radial network

The OPF (11.3) and its SOCP relaxation (11.5) are based on the DistFlow model that assumes $y^s_{jk} = y^s_{kj}$ (assumption C5.1) and $y^m_{jk} = y^m_{kj} = 0$. OPF is also formulated in (9.22) of Chapter 9.2 without these assumptions, based on the branch flow model (5.1) that includes branch variables $\ell := \left(\ell_{jk}, \ell_{kj}, (j,k) \in E\right)$, $S := \left(S_{jk}, S_{kj}, (j,k) \in E\right)$ in both directions, reproduced here:

$$s_j = \sum_{k:j\sim k} S_{jk}, \qquad j \in \overline{N} \tag{11.11a}$$

$$\left|\alpha_{jk}\right|^2 v_j - v_k = 2\,\mathrm{Re}\left(\alpha_{jk} \bar{z}^s_{jk} S_{jk}\right) - \left|z^s_{jk}\right|^2 \ell_{jk}, \qquad (j,k) \in E \tag{11.11b}$$

$$\left|\alpha_{kj}\right|^2 v_k - v_j = 2\,\mathrm{Re}\left(\alpha_{kj} \bar{z}^s_{kj} S_{kj}\right) - \left|z^s_{kj}\right|^2 \ell_{kj}, \qquad (j,k) \in E \tag{11.11c}$$

$$\bar{\alpha}_{jk} v_j - \bar{z}^s_{jk} S_{jk} = \left(\bar{\alpha}_{kj} v_k - \bar{z}^s_{kj} S_{kj}\right)^{\mathsf{H}}, \qquad (j,k) \in E \tag{11.11d}$$

$$\left|S_{jk}\right|^2 = v_j \ell_{jk}, \qquad \left|S_{kj}\right|^2 = v_k \ell_{kj}, \qquad (j,k) \in E \tag{11.11e}$$

where

$$\alpha_{jk} := 1 + z^s_{jk} y^m_{jk}, \qquad \alpha_{kj} := 1 + z^s_{kj} y^m_{kj}$$

The feasible set is

$$\mathbb{X}_{\rm tree} := \{x : (s, v, \ell, S) \in \mathbb{R}^{9N+3} \mid x \text{ satisfies (11.11), (11.2)}\} \tag{11.12a}$$

and the OPF problem is:
OPF:

$$\min_x \; C(x) \quad \text{subject to} \quad x \in \mathbb{X}_{\text{tree}} \tag{11.12b}$$

Its SOCP relaxation replaces the quadratic equality constraint (11.11e) by second-order cones:

$$v_j \ell_{jk} \;\geq\; |S_{jk}|^2, \qquad v_k \ell_{kj} \;\geq\; |S_{kj}|^2, \qquad j \to k \in E \tag{11.13}$$

Then the feasible set is

$$\mathbb{X}_{\text{df}}^+ \;:=\; \{x : (s, v, \ell, S) \in \mathbb{R}^{6N+3} \mid x \text{ satisfies (11.11a)} - \text{(11.11d)}, \text{(11.13)}, \text{(11.2)}\} \tag{11.14a}$$

and
OPF-socp:

$$\min_x \; C(x) \qquad \text{subject to} \qquad x \in \mathbb{X}_{\text{df}}^+ \tag{11.14b}$$

We say that OPF-socp (11.14) is *exact* if *every* optimal solution x^{socp} of (11.14) attains equalities in (11.13) and hence is an optimal solution of OPF (11.12). We study exactness condition for (11.14) in Theorem 11.4 of Chapter 11.2.

11.2 Exactness condition: inactive injection lower bounds

11.2.1 DistFlow model

Consider first OPF (11.3) and its SOCP relaxation (11.5) in the DistFlow model. Assume

> C11.1: The cost function $C(x) = C(p, q, v, \ell)$ is independent of branch flows $S = (P, Q)$ and nondecreasing in (p, q, ℓ). Moreover it is strictly increasing in every component of $(\ell_{jk}, (j,k) \in E)$ or in every component of $(p_j, j \in \overline{N})$ or in every component of $(q_j, j \in \overline{N})$.
> C11.2: For $j \in \overline{N}$, $s_j^{\min} = -\infty - \mathbf{i}\infty$.

Popular cost functions in the literature include active power loss over the network or active power generations, both of which satisfy C11.1.

Theorem 11.3 (Inactive injection lower bounds)**.** Suppose the network graph G is a tree and C11.1, C11.2 hold. Then the SOCP relaxation (11.5) is exact, i.e., every optimal solution x^{socp} of (11.5) is optimal for OPF (11.3).

Remark 11.1 (Strong exactness and global optimality)**.** 1 If the cost function $C(x)$ in C11.1 is only nondecreasing, rather than strictly increasing, in ℓ, then C11.1,

C11.2 still guarantee that all optimal solutions of OPF (11.3) are optimal solutions of its relaxation OPF-socp (11.5), but OPF-socp may have an optimal solution x^{socp} that maintains a strict inequality in (11.4) and hence is infeasible for OPF. Even though OPF-socp is not exact in the strong sense of Definition 10.2, an optimal solution of OPF (11.3) can still be constructed from such a solution x^{socp}; see explanation immediately after the proof of Theorem 11.3 below.

2 Theorem 9.7 of Chapter 9.4.4 shows that C11.2 and a strengthened version of C11.1 (and other mild conditions) also guarantee that every local optimum of OPF (11.3) is a global optimum. □

Remark 11.2 (Convexity)**.** For exact relaxation, we do not require the cost function $C(x)$ to be convex in x; $C(x)$ needs to be convex for (11.5) to be a convex problem.

We can allow more general constraints on power injections s_j than $s_j \le s_j^{\max}$ assumed in Theorem 11.3. The injection s_j can be in an *arbitrary* set B_j that satisfies C11.2. In particular B_j need not be convex nor even connected for OPF-socp to be exact. It (only) needs to be convex to be efficiently computable. Such a general constraint on s is useful in many applications. For instance it allows constraints of the form $|s_j|^2 \le a, |\angle s_j| \le \phi_j$ that is useful for inverter control or $q_j \in \{0, a\}$ for capacitor configuration. □

Proof of Theorem 11.3.

Fix any optimal solution $x := (s, v, \ell, S) \in \mathbb{R}^{3(2N+1)}$ of OPF-socp (11.5). Since G is a tree, the cycle condition is vacuous and we only need to show that x attains equality in (11.4). For the sake of contradiction assume this is violated on line $j \to$k, i.e.,

$$v_j \ell_{jk} > |S_{jk}|^2 \tag{11.15}$$

We will construct an $\tilde{x}$ that is feasible for OPF-socp and attains a strictly lower cost, contradicting the optimality of x.

For an $\epsilon > 0$ to be determined below, consider the following $\tilde{x}$ obtained by modifying only the current ℓ_{jk} and power flow S_{jk} on line $j \to k$ and the injections s_j, s_k at two ends of line $j \to k$:

$$\tilde{\ell}_{jk} := \ell_{jk} - \epsilon \tag{11.16a}$$

$$\tilde{S}_{jk} := S_{jk} - z_{jk}\epsilon/2 \tag{11.16b}$$

$$\tilde{s}_j := s_j - z_{jk}\epsilon/2 \tag{11.16c}$$

$$\tilde{s}_k := s_k - z_{kj}\epsilon/2 \tag{11.16d}$$

and $\tilde{v} := v, \tilde{\ell}_{il} := \ell_{il}$ and $\tilde{S}_{il} := S_{il}$ for $(i,l) \neq (j,k)$, $\tilde{s}_i := s_i$ for $i \neq j, k$. In particular, no other variables than $(s_j, s_k, \ell_{jk}, S_{jk})$ associated with the single line $j \to k$ are modified.[2] By assumption C11.1 the cost function $C(x)$ is strictly increasing in every

[2] In the proof of Theorem 9.7 of Chapter 9.4.4 on global optimality of OPF, the adjustment (9.39) to x is the same as that in (11.16) but on all lines $i \to l \in E$ and all buses $i \in \overline{N}$, not just on $j \to k$, with individual $\epsilon_i = \epsilon_{il} = t\Delta_{il}$.

component of $(\ell_{jk}, (j,k) \in E)$ or in every component of $(p_j, j \in \overline{N})$ or in every component of $(q_j, j \in \overline{N})$. Hence $\tilde{x}$ has a strictly lower cost than x. It suffices to show that there exists an $\epsilon > 0$ such that $\tilde{x}$ is feasible for OPF-socp (11.5), i.e., $\tilde{x}$ satisfies (11.1a)(11.1b)(11.4)(11.2). Moreover we can choose $\epsilon > 0$ so that $\tilde{x}$ attains equalities in (11.4) and is therefore feasible for OPF.

Assumption C11.2 ensures that $\tilde{x}$ satisfies (11.2) since $z_{jk} > 0$ and $\epsilon > 0$. Further $\tilde{x}$ satisfies (11.1a) at buses $i \neq j,k$, and satisfies (11.1b)(11.4) over lines $(i,l) \neq (j,k)$. We now show that $\tilde{x}$ also satisfies (11.1a)(11.1b)(11.4) at buses j,k and over the line (j,k).

For (11.1a) at bus j, we have from (11.16b)(11.16c)

$$\tilde{S}_{jk} = S_{jk} - z_{jk}\frac{\epsilon}{2} \;=\; \sum_{i:i\to j} \left(S_{ij} - z_{ij}\ell_{ij}\right) + s_j \;-\; z_{jk}\frac{\epsilon}{2} \;=\; \sum_{i:i\to j} \left(\tilde{S}_{ij} - z_{ij}\tilde{\ell}_{ij}\right) + \tilde{s}_j$$

as desired (recall that no variables except those associated with line (j,k) are modified). For (11.1a) at bus k, on line $k \to l$ from k towards bus 0, we have from (11.16a)(11.16b)(11.16d)

$$\begin{aligned}\tilde{S}_{kl} = S_{kl} \;&=\; \left(S_{jk} - z_{jk}\ell_{jk}\right) + \sum_{i\neq j:i\to k} \left(S_{ik} - z_{ik}\ell_{ik}\right) + s_k \\ &= \left(\tilde{S}_{jk} - z_{jk}\tilde{\ell}_{jk} - z_{jk}\frac{\epsilon}{2}\right) + \sum_{i\neq j:i\to k} \left(\tilde{S}_{ik} - z_{ik}\tilde{\ell}_{ik}\right) + s_k \;=\; \sum_{i:i\to k} \left(\tilde{S}_{ik} - z_{ik}\tilde{\ell}_{ik}\right) + \tilde{s}_k\end{aligned}$$

as desired. This shows that $\tilde{x}$ satisfies (11.1a) at both buses j,k. For (11.1b) over line (j,k), we have from (11.16a)(11.16b)

$$\tilde{v}_j - \tilde{v}_k = v_j - v_k \;=\; 2\,\mathrm{Re}\left(z_{jk}^{\mathsf{H}} S_{jk}\right) - |z_{jk}|^2 \ell_{jk} \;=\; 2\,\mathrm{Re}\left(z_{jk}^{\mathsf{H}} \tilde{S}_{jk}\right) - |z_{jk}|^2 \tilde{\ell}_{jk}$$

as desired. For (11.4) over line (j,k), we have from (11.16a)(11.16b)

$$\tilde{v}_j \tilde{\ell}_{jk} - \left|\tilde{S}_{jk}\right|^2 = -\frac{\left|z_{jk}\right|^2}{4}\epsilon^2 \;-\; \left(v_j - \mathrm{Re}\left(z_{jk}^{\mathsf{H}} S_{jk}\right)\right)\epsilon \;+\; \left(v_j \ell_{jk} - \left|S_{jk}\right|^2\right)$$

Hence (11.15) implies that we can always choose an $\epsilon > 0$ such that $\tilde{v}_j \tilde{\ell}_{jk} = \left|\tilde{S}_{jk}\right|^2$.

This completes the proof of Theorem 11.3. □

Note that the construction of $\tilde{x}$ ensures that *equalities* are attained in (11.4) and therefore $\tilde{x}$ is feasible for OPF (11.3), not just for its SOCP relaxation. If the cost function $C(x)$ in C11.1 is only nondecreasing, rather than strictly increasing, in ℓ (or in p or q), then it is possible that $C(\tilde{x}) = C(x)$ and OPF-socp (11.5) has an optimal solution x that maintains a strict inequality in (11.4). Even in this case, the proof shows how to construct from such an x an optimal solution $\tilde{x}$ for OPF (11.3) under C11.1 and C11.2.

11.2.2 General radial network

Theorem 11.3 can be extended to general radial networks where z^s_{jk} and z^s_{kj} may not be equal and z^m_{jk}, z^m_{kj} may not be zero. The OPF and its SOCP relaxation for this general model are given in (11.12) and (11.14) respectively. Assume

> C11.3: For all $(j,k) \in E$, both $\mathrm{Re}(\alpha_{jk})$ and $\mathrm{Re}(\alpha_{kj})$ are positive; furthermore $z^s_{jk} = z^s_{kj}$.

For C11.3, $\mathrm{Re}(\alpha_{jk}) = 1 + \mathrm{Re}(z^s_{jk} y^m_{jk}) \geq 1 - |y^m_{jk}/z^s_{jk}|$. Since y^m_{jk} is typically much smaller in magnitude than y^s_{jk}, $\mathrm{Re}(\alpha_{jk})$ is usually strictly positive. The next theorem is proved in Exercise 11.1.

Theorem 11.4 (Inactive injection lower bounds)**.** Suppose the network graph G is a tree and C11.1, C11.2, C11.3 hold. Then the SOCP relaxation (11.14) is exact, i.e., every optimal solution x^{socp} of (11.14) is optimal for OPF (11.12).

11.3 Exactness condition: inactive voltage upper bounds

In this section we present a sufficient condition for exact SOCP relaxation of single-phase OPF on a radial network, when the operational constraint (11.2) is replaced by the following set of constraints:

$$v_j^{\min} \leq v_j \leq v_j^{\max}, \qquad j \in N \tag{11.17a}$$

$$s_j \in B_j \subseteq \{s_j \in \mathbb{C} \,|\, s_j \leq s_j^{\max}\}, \qquad j \in N \tag{11.17b}$$

for some given finite $s_j^{\max}$, $j \in N$. In particular we ignore line limits, but allow the injections $(s_j, j \in N)$ at non-root buses to be in an arbitrary set B_j that is bounded above (see Remark 11.2). We also assume v_0 is given and satisfies (11.17a) and s_0 is unconstrained.

Then OPF and its feasible set are:

$$\textbf{OPF:} \quad \min_x \; C(x) \qquad \text{s.t.} \qquad x \in \mathbb{X}_{\mathrm{df}} \tag{11.18a}$$

$$\text{where} \quad \mathbb{X}_{\mathrm{df}} := \{x : (s,v,\ell,S) \in \mathbb{R}^{6N+3} \mid x \text{ satisfies (11.1)(11.17)}\} \tag{11.18b}$$

Their SOCP relaxations are:

$$\textbf{OPF-socp:} \quad \min_x \; C(x) \qquad \text{s.t.} \qquad x \in \mathbb{X}^+_{\mathrm{df}} \tag{11.19a}$$

$$\text{where} \quad \mathbb{X}^+_{\mathrm{df}} := \{x : (s,v,\ell,S) \in \mathbb{R}^{6N+3} \mid x \text{ satisfies (11.1a)(11.1b)(11.4)(11.17)}\} \tag{11.19b}$$

OPF-socp (11.19) is *exact* if every optimal solution x^{socp} of (11.19) attains equality in (11.4) and is hence optimal for OPF (11.18).

11.3.1 Sufficient condition

We now state the sufficient condition for exact SOCP relaxation for radial networks and show that exactness implies uniqueness of the optimal solution. The main sufficient condition is that the voltage upper bounds are inactive at optimality. [3] Before presenting it we first explain a simple intuition using a two-bus network that motivates this condition.

Example 11.1 (Geometric insight)**.** Consider bus 0 and bus 1 connected by a line with impedance $z := r + \mathbf{i}x$ with $r, x > 0$. Without loss of generality, let the direction of the line be from bus 1 to bus 0. Let ℓ be the sending-end squared current magnitude from buses 1 to 0 (recall that $S_{01} := 0$ in (11.1a)). Suppose also without loss of generality that $v_0 = 1$ pu. The model in (11.1) reduces to (Exercise 11.2):

$$p_0 - r\ell = -p_1, \quad q_0 - x\ell = -q_1, \quad p_0^2 + q_0^2 = \ell \tag{11.20a}$$

$$v_1 - v_0 = 2(rp_1 + xq_1) - (r^2 + x^2)\ell \tag{11.20b}$$

Suppose s_1 is given (e.g., a constant power load). Then the variables are $x := (p_0, q_0, v_1, \ell)$ and the feasible set consists of solutions of (11.20), subject to operational constraints on x. The case without any constraint is instructive and shown in

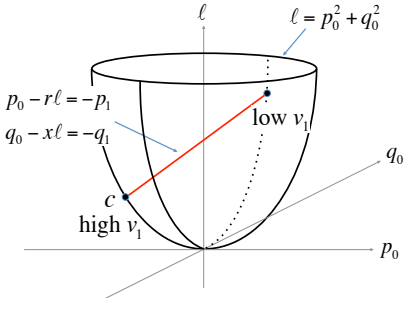


Figure 11.1 Feasible set of OPF for a two-bus network without any constraint. It consists of the (two) points of intersection of the line with the convex surface (without the interior), and hence is nonconvex. SOCP relaxation includes the interior of the convex surface and enlarges the feasible set to the line segment joining these two points. If the cost function C is increasing in ℓ or (p_0, q_0) then the optimal point over the SOCP feasible set (line segment) is the lower feasible point c, and hence the relaxation is exact.

Figure 11.1 (see explanation in the caption). The point c in the figure corresponds to a power flow solution with a large v_1 (normal operation) whereas the other intersection corresponds to a 3 solution with a small v_1 (fault condition). (See Example 5.3 of Chapter 5.1.5 for detailed calculations.) As explained in the caption, SOCP relaxation is exact if there is no voltage constraint and as long as constraints on (p_0, q_0, ℓ) do not remove the high-voltage solution c. Only when the system is stressed to a point

[3] Exercise 10.3 shows that, since SOCP is a convex problem, condition C11.5 that requires an upper bound of v_j be less than $v_j^{\max}$ is equivalent to requiring $v_j \le v_j^{\max}$ but the bound is inactive at optimality.

where the high-voltage solution becomes infeasible will relaxation lose exactness. This agrees with conventional wisdom that power systems under normal operations are well behaved.

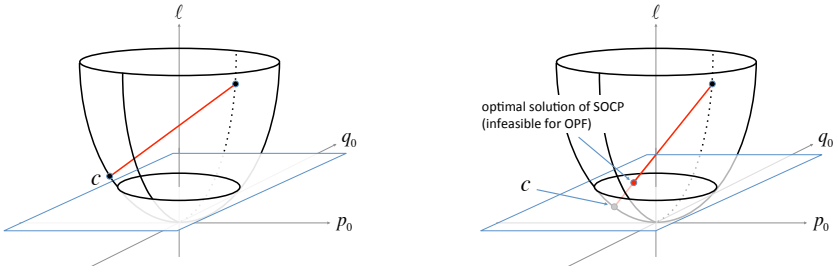
Consider now the voltage constraint $v_1^{\min} \le v_1 \le v_1^{\max}$. We have from (11.20b) and $v_0 = 1$

$$v_1 = (1 + 2rp_1 + 2xq_1) - |z|^2 \ell$$

translating the constraint on v_1 into a box constraint on ℓ:

$$\frac{1}{|z|^2}\left(2rp_1 + 2xq_1 + 1 - v_1^{\max}\right) \;\le\; \ell \;\le\; \frac{1}{|z|^2}\left(2rp_1 + 2xq_1 + 1 - v_1^{\min}\right)$$

Figure 11.1 shows that the lower bound $v_1^{\min}$ (corresponding to an upper bound on ℓ) does not affect the exactness of SOCP relaxation. The effect of upper bound $v_1^{\max}$ (corresponding to a lower bound on ℓ) is illustrated in Figure 11.2. As explained in the caption of the figure SOCP relaxation is exact if the upper bound $v_1^{\max}$ does not exclude the high-voltage solution c and is not exact otherwise.



(a) Voltage constraint not binding (b) Voltage constraint binding

Figure 11.2 Impact of voltage upper bound $v_1^{\max}$ on exactness. (a) When $v_1^{\max}$ (corresponding to a lower bound on ℓ) is not binding, the power flow solution c is in the feasible set of SOCP and hence the relaxation is exact. (b) When $v_1^{\max}$ excludes c from the feasible set of SOCP, the optimal solution is infeasible for OPF and the relaxation is not exact.

See Example 5.3 and Exercise 11.3 for details of feasibility and exactness of OPF-socp. □

To state the exactness condition for a general radial network, recall the linear approximation of BFM studied in Chapter 5.4.3.2, obtained by setting $\ell_{jk} = 0$ in (11.1). Given v_0 and the injections $\hat{s} := (\hat{p}, \hat{q}) := (p_j, q_j, j \in N)$ at non-root buses, the line flow vector $S^{\text{lin}}(s) := \left(S^{\text{lin}}_{jk}, (j,k) \in E\right)$ and the voltage vector $\hat{v}^{\text{lin}}(s) := (v_j^{\text{lin}}, j \in N)$ at non-root buses in the linearized model are explicitly given by (from Theorem 5.3):

$$S^{\text{lin}}(s) \;=\; \hat{C}^{-1}\hat{s}, \qquad \hat{v}^{\text{lin}}(s) \;=\; v_0\mathbf{1} + 2(R\hat{p} + X\hat{q}) \tag{11.21}$$

for some given invertible matrices $\hat{C}$, R and X. The key property we will use is, from

Corollary 5.5:

$$S_{jk} \le S_{jk}^{\text{lin}}(s) \text{ and } v_j \le v_j^{\text{lin}}(s), \ j \in N \tag{11.22}$$

Define the 2×2 matrix function

$$A_{jk}(S_{jk}, v_j) := \mathbb{I}_2 - \frac{2}{v_j} z_{jk} \left(S_{jk}\right)^{\mathsf{T}} \tag{11.23}$$

where $\mathbb{I}_2$ is the identity matrix of size 2, $z_{jk} := (r_{jk}, x_{jk})$ is the column vector of line impedance and $S_{jk} := (P_{jk}, Q_{jk})$ is the column vector of branch power flows, so that $z_{jk}\left(S_{jk}\right)^{\mathsf{T}}$ is a 2×2 matrix with rank less or equal to 1. As we will see below, the matrices $A_{jk}(S_{jk}, v_j)$ describe how changes in branch power flows propagate towards the root node 0. Evaluate the Jacobian matrix $A_{jk}(S_{jk}, v_j)$ at the boundary values:

$$\underline{A}_{jk} := A_{jk}\left(\left[S_{jk}^{\text{lin}}(s^{\max})\right]^+, v_j^{\min}\right) = \mathbb{I}_2 - \frac{2}{v_j^{\min}} z_{jk}\left(\left[S_{jk}^{\text{lin}}(s^{\max})\right]^+\right)^{\mathsf{T}} \tag{11.24}$$

Here $\left([a]^+\right)^{\mathsf{T}}$ is the row vector $[[a_1]^+ \ [a_2]^+]$ with $[a_j]^+ := \max\{a_j, 0\}$.

For a radial network, for $j \neq 0$, every line $j \to k$ identifies a unique node k and therefore, to simplify notation, we refer to a line interchangeably by (j,k) or j and use $A_j, \underline{A}_j, z_j$ etc. in place of $A_{jk}, \underline{A}_{jk}, z_{jk}$ etc. respectively. Assume

C11.4: The cost function is $C(x) := \sum_{j=0}^{N} C_j\left(p_j\right)$ with $C_0(p_0)$ strictly increasing in p_0. There is no constraint on s_0.
C11.5: The set B_j of injections satisfies $\hat{v}_j^{\text{lin}}(s) \le v_j^{\max}$, $j \in N$, where $\hat{v}_j^{\text{lin}}(s)$ is given by (11.21).
C11.6: For each leaf node $j \in N$ let the unique path from j to 0 have k lines and be denoted by $\mathsf{P}_j := ((i_k, i_{k-1}), \ldots, (i_1, i_0))$ with $i_k = j$ and $i_0 = 0$. Then $\underline{A}_{i_t} \cdots \underline{A}_{i_{t'}} z_{i_{t'+1}} > 0$ for all $1 \le t \le t' < k$, where $\underline{A}_j$ are defined in (11.24).

Theorem 11.5. Suppose the network graph G is a tree and C11.4–C11.6 hold. Then OPF-socp (11.19) is exact.

The proof of Theorem 11.5 is long and relegated to Appendix 11.3.2. It can be shown that Theorem 11.5 have the following simple and practical interpretation: OPF-socp is exact provided at least one of the following is satisfied:

- There are no reverse power flows in the network.
- The r/x ratios on all lines are equal.
- If the r/x ratios increase in the downstream direction from the substation (node 0) to the leaves then there are no reverse real power flows.
- If the r/x ratios decrease in the downstream direction then there are no reverse reactive power flows.

These properties are derived in [137, 138, 139] and are special cases of Theorem 11.5.

We now comment on the conditions C11.4–C11.6. C11.5 is affine in the injections $s := (p, q)$. It enforces the upper bounds on voltage magnitudes because of (11.22). C11.6 is a technical assumption and has a simple interpretation: the branch power flow S_{jk} on all branches should move in the same direction. Specifically, given a marginal change in the complex power on line $j \to k$, the 2×2 matrix $\underline{A}_{jk}$ is (a lower bound on) the Jacobian and describes the effect of this marginal change on the complex power on the line immediately upstream from line $j \to k$. The product of $\underline{A}_i$ in C11.6 propagates this effect upstream towards the root. C11.6 requires that a small change, positive or negative, in the power flow on a line affects *all* upstream branch powers in the same direction. This seems to hold with a significant margin in practice.

Remark 11.3 (Strong exactness)**.** Condition C11.4 requires that the cost functions C_j depend only on the injections p_j. For instance, if $C_j(p_j) = p_j$, then the cost is total active power loss over the network. It also requires that C_0 be strictly increasing but makes no assumption on $C_j, j > 0$, e.g., the total cost $C(x)$ can be $C_0(p_0)$. Common cost functions such as line loss or generation cost usually satisfy C11.4. If C_0 is only nondecreasing, rather than strictly increasing, in p_0 then C11.4–C11.6 still guarantee that all optimal solutions of OPF (11.18) are (effectively) optimal for OPF-socp (11.19), but OPF-socp may not be exact in our definition, i.e., it may also have an optimal solution that maintains a strict inequality in (11.4). In this case the proof of Theorem 11.5 can still construct from it another optimal solution that attains equalities in (11.4) and is hence optimal for OPF. □

11.3.2 Appendix: Proof of Theorem 11.5

Given an optimal solution $x := (s, v, \ell, S)$ that maintains a strict inequality in (11.4), $v_j \ell_{jk} > |S_{jk}|^2$, for some line $j \to k \in E$, the proof of Theorem 11.3 in Section 11.2 constructs another feasible solution $\hat{x}$ from x that incurs a strictly smaller cost, contradicting the optimality of x. The modification is over a single line over which x maintains a strict inequality. The proof of Theorem 11.5 is also by contradiction but, unlike that of Theorem 11.3, the construction of $\hat{x}$ from x involves modifications on multiple lines, propagating from the line that is closest to bus 0 where strictly inequality holds all the way to bus 0. The proof relies crucially on the recursive structure of the branch flow model (11.1).

Proof of Theorem 11.5 To simplify notation we only prove the theorem for the case of a linear network representing a primary feeder without laterals. The proof for a general tree network follows the same idea but with more cumbersome notations; see [34] for details. We adopt the graph orientation where every line points *towards* the root node 0. The notation for the linear network is explained in Figure 11.3 (we refer to a line $j \to k$ by j and index the associated variables $z_{jk}, S_{jk}, \ell_{jk}$ with j). With this

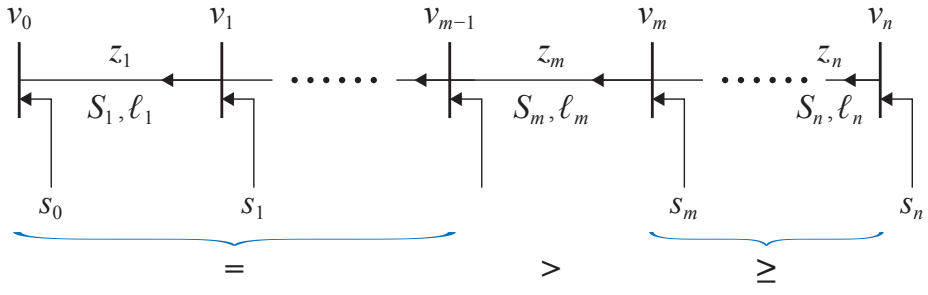


Figure 11.3 Linear network and notations. Line m in the proof is the line closest to bus 0 where the inequality in (11.26) is strict, i.e., (11.26) holds with equality at lines $j = 1, \ldots, m-1$, strict inequality at line m, and inequality at lines $j = m+1, \ldots, N$.

notation the branch flow model (11.1) is the following recursion:

$$S_{j-1} = S_j - z_j \ell_j + s_{j-1}, \qquad j = 1, \ldots, N \tag{11.25a}$$

$$v_{j-1} = v_j - 2\,\mathrm{Re}\left(z_j^{\mathsf{H}} S_j\right) + |z_j|^2 \ell_j, \quad j = 1, \ldots, N \tag{11.25b}$$

$$v_j \ell_j = |S_j|^2, \qquad j = 1, \ldots, N \tag{11.25c}$$

$$S_n = s_n, \qquad S_0 := 0 \tag{11.25d}$$

where v_0 is given. The SOCP relaxation of (11.25c) is:

$$v_j \ell_j \geq |S_j|^2, \qquad j = 1, \ldots, N \tag{11.26}$$

OPF on the linear network in Figure 11.3 then becomes (s_0 is unconstrained by assumption C11.4):

OPF:

$$\min_x \; C(x) := \sum_{j=0}^{N} C_j\left(p_j\right) \tag{11.27a}$$

$$\text{s.t. } (11.17)(11.25) \tag{11.27b}$$

and its SOCP relaxation becomes:

OPF-socp:

$$\min_x \; C(x) := \sum_{j=0}^{N} C_j\left(p_j\right)$$

$$\text{s.t. } (11.17), (11.25a)(11.25b)(11.25d), (11.26) \tag{11.28a}$$

For the linear network assumption C11.6 reduces:

C11.6': $\underline{A}_j \cdots \underline{A}_k \; z_{k+1} > 0$ for $1 \leq j \leq k < N$ where $\underline{A}_j$ are defined in (11.24).

Our goal is to prove OPF-socp (11.28) is exact, i.e., every optimal solution of (11.28) attains equality in (11.26) and hence is also optimal for OPF (11.27). Suppose on the

contrary that there is an optimal solution $x := (S, \ell, v, s)$ of OPF-socp (11.28) that violates (11.25c). We will construct another feasible point $\hat{x} := (\hat{S}, \hat{\ell}, \hat{v}, \hat{s})$ of OPF-socp (11.28) that has a strictly lower cost than x, contradicting the optimality of x.

Let $m := \min\{j \in N \mid v_j \ell_j > |S_j|^2\}$ be the closest line from bus 0 where (11.25c) is violated; see Figure 11.3. Pick any $\epsilon_m \in (0, \ell_m - |S_m|^2/v_m]$ and construct $\hat{x}$ as follows:

1 $\hat{s}_j := s_j$ for $j \neq 0$.
2 For $\hat{S}, \hat{\ell}, \hat{s}_0$:
- For $j = N, \ldots, m+1$: $\hat{S}_j := S_j$ and $\hat{\ell}_j := \ell_j$.
- For $j = m$: $\hat{S}_m := S_m$ and $\hat{\ell}_m := \ell_m - \epsilon_m$.
- For $j = m-1, \ldots, 1$:

$$\hat{S}_j := \hat{S}_{j+1} - z_{j+1}\hat{\ell}_{j+1} + \hat{s}_j$$
$$\hat{\ell}_j := \frac{|\hat{S}_j|^2}{v_j}$$

- $\hat{s}_0 := -\hat{S}_1 + z_1\hat{\ell}_1$.

3 $\hat{v}_0 := v_0$. For $j = 1, \ldots, N$,

$$\hat{v}_j := \hat{v}_{j-1} + 2\,\mathrm{Re}\left(z_j^{\mathsf{H}}\hat{S}_j\right) - |z_j|^2\hat{\ell}_j$$

Notice that the denomintor in $\hat{\ell}_j$ is defined to be v_j, not $\hat{v}_j$. This decouples the recursive construction of $(\hat{S}_j, \hat{\ell}_j)$ and $\hat{v}_j$ so that the former propagates from bus N towards bus 1 while the latter propagates in the opposite direction, as in backward forward sweep studied in Chapter 5.3.

By construction $\hat{x}$ satisfies (11.25a), (11.25b), (11.25d), and (11.17b). We only have to prove that $\hat{x}$ satisfies (11.17a) and (11.26). Hence the proof of Theorem 11.5 is complete after Lemma 11.6 is established, which asserts that $\hat{x}$ is feasible and has a strictly lower cost under assumptions C11.4, C11.5, C11.6'.

Lemma 11.6. Under the conditions of Theorem 11.5 $\hat{x}$ satisfies

1 $C(\hat{x}) < C(x)$.
2 $\hat{v}_j\hat{\ell}_j \geq \left|\hat{S}_j\right|^2$, $j \in N$.
3 $v_j^{\min} \leq \hat{v}_j \leq v_j^{\max}$, $j \in N$.

To simplify notation redefine $S_0 := -s_0$ and $\hat{S}_0 := -\hat{s}_0$. Then for $j \in \overline{N}$ define $\Delta S_j := \hat{S}_j - S_j$ and $\Delta v_j := \hat{v}_j - v_j$. The key result that leads to Lemma 11.6 is:

$$\Delta S_j \geq 0 \text{ and } \Delta v_j \geq 0, \qquad j \in \overline{N}$$

The first inequality is stated more precisely in Lemma 11.7 and proved after the proof of Lemma 11.6.

Lemma 11.7. Suppose $m > 1$ and C11.6' holds. Then $\Delta S_j \geq 0$ for $j \in \overline{N}$ with $\hat{S}_j > S_j$ for $j = 0, \ldots, m-1$. In particular $\hat{s}_0 < s_0$.

We now prove the second inequality together with Lemma 11.6 assuming Lemma 11.7 holds.

Proof of Lemma 11.6 1) If $m = 1$ then, by construction, $\hat{s}_0 = s_0 - z_1\epsilon_1 < s_0$ since $z_1 > 0$. If $m > 1$ then $\hat{s}_0 < s_0$ by Lemma 11.7. Since $\hat{s} = s$ and $\hat{s}_0 < s_0$ we have

$$C(\hat{x}) - C(x) = \sum_{j=0}^{N} \left(C_j\left(\hat{p}_j\right) - C_j\left(p_j\right)\right) \;=\; C_0\left(\hat{p}_0\right) - C_0\left(p_0\right) \;<\; 0$$

as desired, since C_0 is strictly increasing.

2) To avoid circular argument we will first prove using Lemma 11.7

$$\hat{v}_j \geq v_j, \qquad j \in N \tag{11.29}$$

We will then use this and Lemma 11.7 to prove $\hat{v}_j\hat{\ell}_j \geq |\hat{S}_j|^2$ for all $j \in N$. We then use assumption C11.5 to prove $v_j^{\min} \leq \hat{v}_j \leq v_j^{\max}$, $j \in N$. This shows that $\hat{x}$ satisfies (11.26) and (11.17a) (in addition to (11.25a)(11.25b)(11.25d) and (11.17b)).

To prove (11.29), note that both $\hat{v}$ and v satisfy (11.25b) and hence we have, for $j = 1, \ldots, N$,

$$\Delta v_{j-1} = \Delta v_j - 2\,\mathrm{Re}\left(z_j^{\mathsf{H}}\Delta S_j\right) + |z_j|^2\Delta\ell_j \tag{11.30}$$

where $\Delta\ell_j := \hat{\ell}_j - \ell_j$. From (11.25a) we have

$$z_j\Delta\ell_j = \Delta S_j - \Delta S_{j-1} + \Delta s_{j-1}$$

where $\Delta s_0 := \hat{s}_0 - s_0 < 0$ and $\Delta s_{j-1} = 0$ for $j > 1$. Multiplying both sides by z_j^{H} and noticing that both sides must be real, we conclude

$$|z_j|^2\Delta\ell_j = \mathrm{Re}\,\left(z_j^{\mathsf{H}}\Delta S_j - z_j^{\mathsf{H}}\Delta S_{j-1} + z_j^{\mathsf{H}}\Delta s_{j-1}\right)$$

Substituting into (11.30) we have for $j = 1, \ldots, N$

$$\Delta v_j - \Delta v_{j-1} = \mathrm{Re}\; z_j^{\mathsf{H}}\Delta S_j \;+\; \mathrm{Re}\; z_j^{\mathsf{H}}\Delta S_{j-1} \;-\; \mathrm{Re}\; z_j^{\mathsf{H}}\Delta s_{j-1}$$

But Lemma 11.7 implies that $\mathrm{Re}\; z_j^{\mathsf{H}}\Delta S_j = r_j\Delta P_j + x_j\Delta Q_j \geq 0$. Similarly every term on the right-hand side is nonnegative and hence

$$\Delta v_j \geq \Delta v_{j-1} \qquad \text{for } j = 1, \ldots, N$$

implying that $\Delta v_j \geq \Delta v_0 = 0$, proving (11.29).

We now use (11.29) to prove the second assertion of the lemma. By construction,

for $j = m+1, \ldots, N$,

$$\hat{\ell}_j = \ell_j \;\geq\; \frac{|S_j|^2}{v_j} \;\geq\; \frac{|\hat{S}_j|^2}{\hat{v}_j}$$

as desired, since $\hat{S}_j = S_j$ and $\hat{v}_j \geq v_j$. Similarly (11.26) holds for $\hat{x}$ for $j = m$ because of the choice of ϵ_m. For $j = 1, \ldots, m-1$, $\hat{v}_j \geq v_j$ again implies

$$\hat{\ell}_j = \frac{|\hat{S}_j|^2}{v_j} \;\geq\; \frac{|\hat{S}_j|^2}{\hat{v}_j}$$

3) The relation (11.29) means

$$\hat{v}_j \geq v_j \;\geq\; v_j^{\min}, \qquad j \in N$$

Assumption C11.5 and (11.22) imply that

$$\hat{v}_j \leq v_j^{\text{lin}}(s) \;\leq\; v_j^{\max}, \qquad j \in N$$

This proves $\hat{x}$ satisfies (11.17a) and completes the proof of Lemma 11.6. □

The remainder of this subsection is devoted to proving the key result Lemma 11.7.

Proof of Lemma 11.7 By construction $\Delta S_j = 0$ for $j = m, \ldots, n$. To prove $\Delta S_j > 0$ for $j = 0, \ldots, m-1$, the key idea is to derive a recursion on ΔS_j in terms of the Jacobian matrix $A_j(S_j, v_j)$. The intuition is that, when the branch current ℓ_m is reduced by ϵ_m to $\hat{\ell}_m$, loss on line m is reduced and all upstream branch powers S_j will be increased to $\hat{S}_j$ as a consequence.

This is proved in three steps, of which we now give an informal overview. First we derive a recursion (11.32) on ΔS_j. This motivates a collection of linear dynamical systems w in (11.34) that contains the process $(\Delta S_j,\ j = 0, \ldots, m-1)$ as a specific trajectory. Second we construct another collection of linear dynamical systems $\underline{w}$ in (11.35) such that assumption C11.6' implies $\underline{w} > 0$. Finally we prove an expression for the process $w - \underline{w}$ that shows $w \geq \underline{w}$ (in Lemmas 11.8, 11.9, 11.10). This then implies $\Delta S = w \geq \underline{w} > 0$ as desired. We now make these steps precise.

Since both x and $\hat{x}$ satisfy (11.25a) and $\hat{s}_j = s_j$ for all $j \in N$ we have (with the redefined $\Delta S_0 := -(\hat{s}_0 - s_0)$)

$$\Delta S_{j-1} = \Delta S_j - z_j \Delta \ell_j, \quad j = 1, 2, \ldots, N \tag{11.31}$$

where $\Delta \ell_j := \hat{\ell}_j - \ell_j$. For $j = 1, \ldots, m-1$ both x and $\hat{x}$ satisfy (11.25c). For these j, fix any $v_j \geq v_j^{\min}$ and consider $\ell_j := \ell_j(S_j)$ as functions of the real pair $S_j := (P_j, Q_j)$:

$$\ell_j(S_j) := \frac{P_j^2 + Q_j^2}{v_j}, \quad j = 1, \ldots, m-1$$

whose Jacobian are the row vectors:

$$\frac{\partial \ell_j}{\partial S_j}(S_j) = \frac{2}{v_j}[P_j \;\; Q_j] \;=\; \frac{2}{v_j} S_j^{\mathsf{T}}$$

The mean value theorem implies for $j = 1, \ldots, m-1$

$$\Delta \ell_j = \ell_j(\hat{S}_j) - \ell_j(S_j) \;=\; \frac{\partial \ell_j}{\partial S_j}(\tilde{S}_j)\, \Delta S_j$$

where $\tilde{S}_j := \alpha_j S_j + (1-\alpha_j)\hat{S}_j$ for some $\alpha_j \in [0,1]$. Substituting it into (11.31) we obtain the recursion, for $j = 1, \ldots, m-1$,

$$\Delta S_{j-1} = \tilde{A}_j\, \Delta S_j \tag{11.32a}$$

$$\Delta S_{m-1} = \epsilon_m z_m \;>\; 0 \tag{11.32b}$$

where the 2×2 matrix $\tilde{A}_j$ is the matrix function $A_j(S_j, v_j)$ defined in (11.23) evaluated at $(\tilde{S}_j, v_j)$:

$$\tilde{A}_j \;:=\; A_j(\tilde{S}_j, v_j) := \mathbb{I}_2 - \frac{2}{v_j} z_j \tilde{S}_j^{\mathsf{T}} \tag{11.33}$$

which depends on $(S_j, \hat{S}_j)$ through $\tilde{S}_j$.

Note that $\tilde{A}_j$ and ΔS_j are not independent since both are defined in terms of $(S_j, \hat{S}_j)$, and therefore strictly speaking (11.32) does not specify a *linear* system. Given an optimal solution x of the relaxation OPF-socp (11.28) and our modified solution $\hat{x}$, however, the sequence of matrices $\tilde{A}_j$, $j = 1, \ldots, m-1$, are fixed. We can therefore consider the following collection of discrete-time linear time-varying systems (one for each τ), whose state at time t (going backward in time) is $w(t;\tau)$, when it starts at time $\tau \geq t$ in the initial state $z_{\tau+1}$: for each τ with $0 < \tau < m$,

$$w(t-1;\tau) = \tilde{A}_t\, w(t;\tau), \quad t = \tau, \tau-1, \ldots, 1 \tag{11.34a}$$

$$w(\tau;\tau) = z_{\tau+1} \tag{11.34b}$$

Clearly $\Delta S_j = \epsilon_m\, w(j; m-1)$. Hence, to prove $\Delta S_j > 0$, it suffices to prove $w(j; m-1) > 0$ for all j with $0 \leq j \leq m-1$.

To this end we compare the system $w(t;\tau)$ with the following collection of linear time-variant systems: for each τ with $0 < \tau < m$,

$$\underline{w}(t-1;\tau) = \underline{A}_t\, \underline{w}(t;\tau), \quad t = \tau, \tau-1, \ldots, 1 \tag{11.35a}$$

$$\underline{w}(\tau;\tau) = z_{\tau+1} \tag{11.35b}$$

where $\underline{A}_t$ is defined in (11.24) and reproduced here:

$$\underline{A}_t := A_t\left(\left[S_t^{\mathrm{lin}}(s^{\max})\right]^+, \underline{v}_t\right) \;=\; \mathbb{I}_2 - \frac{2}{v_t^{\min}} z_t \left(\left[S_t^{\mathrm{lin}}(s^{\max})\right]^+\right)^{\mathsf{T}} \tag{11.36}$$

Note that $\underline{A}_t$ are *independent of* the OPF-socp solution x and our modified solution $\hat{x}$. Then assumption C11.6' is equivalent to

$$\underline{w}(t;\tau) > 0 \quad \text{for all } 0 \leq t \leq \tau < m \tag{11.37}$$

We now prove, in Lemmas 11.8, 11.9, 11.10, that $w(t;\tau) \geq \underline{w}(t;\tau)$ and hence C11.6' implies $\Delta S_j = \epsilon_m\, w(j;m-1) \geq \epsilon_m\, \underline{w}(j;m-1) > 0$, establishing Lemma 11.7.

Lemma 11.8. For each $t = m-1,\ldots,1$

$$\tilde{A}_t - \underline{A}_t = 2\, z_t\, \delta_t^{\mathsf{T}}$$

for some 2-dimensional vector $\delta_t \geq 0$.

Proof of Lemma 11.8 Fix any $t = m-1,\ldots,1$. We have $S_t \leq S_t^{\text{lin}}(s)$ from (11.22). Even though we have not yet proved $\hat{S}_t$ is feasible for OPF-socp we know $\hat{S}_t$ satisfies (11.25a) by construction of $\hat{x}$. The same argument as in Corollary 5.5 then shows $\hat{S}_t \leq S_t^{\text{lin}}(s)$. Hence $\tilde{S}_t := \alpha_t S_t + (1-\alpha_t)\hat{S}_t$, $\alpha_t \in [0,1]$, satisfies $\tilde{S}_t \leq S_t^{\text{lin}}(s)$. Hence

$$\tilde{S}_t \;\leq\; S_t^{\text{lin}}(s) \;\leq\; S_t^{\text{lin}}(s^{\max}) \;\leq\; \left[S_t^{\text{lin}}(s^{\max})\right]^+ \tag{11.38}$$

Using the definitions of $\tilde{A}_t$ in (11.33) and $\underline{A}_t$ in (11.36) we have $\tilde{A}_t - \underline{A}_t = 2\, z_t \delta_t^{\mathsf{T}}$ where

$$\delta_t^{\mathsf{T}} := \left[\frac{\left[P_t^{\text{lin}}(s^{\max})\right]^+}{v_t^{\min}} - \frac{\tilde{P}_t}{v_t} \quad \frac{\left[Q_t^{\text{lin}}(s^{\max})\right]^+}{v_t^{\min}} - \frac{\tilde{Q}_t}{v_t}\right]$$

Then (11.38) and $v_t \geq v_t^{\min}$ impy that $\delta_t \geq 0$. □

For each τ with $0 < \tau < m$ define the scalars $a(t;\tau)$ in terms of the solution $\underline{w}(t;\tau)$ of (11.35) and δ_t in Lemma 11.8:

$$a(t;\tau) := 2\,\delta_t^{\mathsf{T}}\, \underline{w}(t;\tau) \;\; > \;\; 0 \tag{11.39}$$

Lemma 11.9. Fix any τ with $0 < \tau < m$. For each $t = \tau, \tau-1,\ldots,0$ we have

$$w(t;\tau) - \underline{w}(t;\tau) = \sum_{t'=t+1}^{\tau} a(t';\tau)\, w(t;t'-1)$$

Proof of Lemma 11.9 Fix a τ with $0 < \tau < m$. We now prove the lemma by induction on $t = \tau, \tau-1,\ldots,0$. The assertion holds for $t = \tau$ since $w(\tau;\tau) - \underline{w}(\tau;\tau) = 0$. Suppose it holds for t. Then for $t-1$ we have from (11.34) and (11.35)

$$\begin{aligned}
w(t-1;\tau) - \underline{w}(t-1;\tau) &= \tilde{A}_t\, w(t;\tau) - \underline{A}_t\, \underline{w}(t;\tau) \\
&= \left(\tilde{A}_t - \underline{A}_t\right) \underline{w}(t;\tau) \;+\; \tilde{A}_t \left(w(t;\tau) - \underline{w}(t;\tau)\right) \\
&= a(t;\tau)\, z_t \;+\; \sum_{t'=t+1}^{\tau} a(t';\tau)\; \tilde{A}_t\, w(t;t'-1) \\
&= a(t;\tau)\, z_t \;+\; \sum_{t'=t+1}^{\tau} a(t';\tau)\, w(t-1;t'-1) \\
&= \sum_{t'=t}^{\tau} a(t';\tau)\, w(t-1;t'-1)
\end{aligned}$$

where the first term on the right-hand side of the third equality follows from Lemma 11.8 and the definition of $a(t;\tau)$ in (11.39), and the second term from the induction hypothesis. The last two equalities follow from (11.34). □

Lemma 11.10. Suppose C11.6' holds. Then for each τ with $0 < \tau < m$ and each $t = \tau, \tau-1, \ldots, 0$,

$$w(t;\tau) \geq \underline{w}(t;\tau) \; > \; 0 \tag{11.40}$$

Proof of Lemma 11.10 We prove the lemma by induction on (t,τ).

1 *Base case:* For each τ with $0 < \tau < m$, (11.40) holds for $t = \tau$, i.e., for t such that $\tau - t = 0$.
2 *Induction hypothesis:* For each τ with $0 < \tau < m$, suppose (11.40) holds for $t \leq \tau$ such that $0 \leq \tau - t \leq k-1$.
3 *Induction:* We will prove that, for each τ with $0 < \tau < m$, (11.40) holds for $t \leq \tau$ such that $0 \leq \tau - t \leq k$. For $t = \tau - k$ we have from Lemma 11.9

$$w(t;\tau) - \underline{w}(t;\tau) = \sum_{t'=t+1}^{\tau} a(t';\tau)\, w(t;t'-1)$$

But each $w(t;t'-1)$ in the summands satisfies $w(t;t'-1) \geq \underline{w}(t;t'-1)$ by the induction hypothesis. Hence, since $a(t';\tau) > 0$,

$$w(t;\tau) - \underline{w}(t;\tau) \geq \sum_{t'=t+1}^{\tau} a(t';\tau)\, \underline{w}(t;t'-1) \;\; > \;\; 0$$

where the last inequality follows from (11.37) and (11.39).

This completes our induction proof. □

Lemma 11.10 implies, for $j = 0, \ldots, m-1$, $\Delta S_j = \epsilon_m\, w(j;m-1) > 0$. This completes the proof of Lemma 11.7. □

This completes the proof of Theorem 11.5 for the linear network. For a general tree network the proof is almost identical, except with more cumbersome notations, by focusing on a path from the root to a first line m over which $v_j \ell_j > |S_j|^2$; see [34]. □

11.4 Bibliographical notes

SOCP relaxation of Chapter 11.1 for radial networks in the DistFlow model of [24, 25] is first proposed in [140, 31]. Theorem 11.2 is proved in [33] and the proof presented here

follows that in [39, Theorem 11]. Theorem 11.3 is from [31, Part I] which generalizes an earlier result in [140] to allow convex objective functions, shunt elements, and line limits. Theorems 11.5 and 11.1 are from [34]. The semidefinite relaxation of three-phase OPF in Chapter **??** follows the idea in [105, 136].

11.5 Problems

Chapter 11.2.

Exercise 11.1 (Exactness: general tree)**.** Let $x := (s, v, \ell, S) \in \mathbb{R}^{9N+3}$ be any optimal solution of OPF-socp (11.14).

1 Show that $v_j \ell_{jk} > |S_{jk}|^2$ if and only if $v_k \ell_{kj} > |S_{kj}|^2$.
2 Prove Theorem 11.4.

(Hint: Modify proof of Theorem 11.3.)

Chapter 11.3.

Exercise 11.2 (Geometric insight)**.** For the 2-bus network in Example 11.1 of Chapter 11.3, derive the model (11.20) from the DistFlow equation (11.1) in the up orientation.

Exercise 11.3 (Feasible set and relaxation)**.** This problem illustrates SOCP relaxation of OPF and its exactness. Consider the 2-bus network in Example 11.1 of Chapter 11.3. Suppose $q_1 = 0$ and $v_0 = r = x = 1$ pu, and suppose the injection p_1 is controllable. Let $w := (p_0, q_0, p_1, v_1, \ell)$. Consider the OPF problem:

$$\min_{w} \quad C(w) \qquad \text{s.t.} \qquad p_0 - \ell = -p_1, \quad q_0 - \ell = 0, \quad v_1 - 1 = 2p_1 - 2\ell, \quad p_0^2 + q_0^2 = \ell$$
$$0.9 \text{ pu} \le v_1 \le 1.1 \text{ pu}, \qquad p_1^{\min} \le p_1 \le p_1^{\max}$$

where the cost function $C(w)$ is strictly increasing in ℓ. Its SOCP relaxation replaces the quadratic equality constraint with the convex constraint $p_0^2 + q_0^2 \le \ell$.

1 Determine the largest range $R_1 := [p_1^{\min}, p_1^{\max}]$ over which the SOCP relaxation is exact.
2 Determine the largest range $R_2 := [p_1^{\min}, p_1^{\max}]$ over which the SOCP relaxation is inexact. Note that in this regime, bus 1 is generating power and causing a large amount of reverse power flow.
3 What happens if the range $[p_1^{\min}, p_1^{\max}]$ for injection p_1 overlaps with neither R_1 nor R_2?

Draw a diagram to illustrate your answers. (Hint: The power flow solution as a function of p_1 is computed in Example 5.3 of Chapter 5.1.5.)

12 Nonsmooth convex optimization

Consider an optimization problem

$$\min_{x \in \mathbb{R}^n} \quad f(x) \qquad \text{s.t.} \qquad x \in X$$

where f is a convex function and $X \subseteq \mathbb{R}^n$ is a convex set. We will develop a basic theory to answer the following questions:

Q1 How to characterize optimal solutions?
Q2 When will optimal solutions exist?

We study these two questions in Chapter 8.3 when the cost and constraint functions are continuously differentiable. In many applications, however, these functions are convex but not differentiable everywhere and may take infinite values. We will show in this chapter that the optimality results summarized in Table 8.1 hold in a nonsmooth setting. We will develop set theoretic tools that handles nonsmooth but convex functions. This basic machinery enables a more fundamental, and simpler, approach and reveals that smoothness is unimportant for the theory of convex optimization (though smoothness can be important for computation).

Optimality conditions and algorithms for convex optimizations are often based on the linear approximations of the cost and constraint functions, e.g., the KKT condition (8.38) or the Newton-Raphson algorithm (8.81)(8.82). In particular the stationarity condition in (8.38) says

$$-\nabla f(x^*) \;=\; \nabla g(x^*)\lambda^* + \nabla h(x^*)\mu^*, \tag{12.1}$$

i.e., a feasible point x^* is a minimizer if the negative gradient $-\nabla f(x^*)$ points away from a linear approximation of the feasible set at x^* defined by the gradients $\nabla g(x^*), \nabla h(x^*)$ of the constraint functions at x^*. In the nonsmooth setting a linear approximation of the feasible set is called a tangent cone and a feasible point x^* is a minimizer if there is a negative cost subgradient that points away from the tangent cone of the feasible set at x^*, i.e., the subgradient is contained in the normal cone $N_X(x^*)$ of the feasible set at x^*. To describe optimality conditions precisely we need the following generalizations:

- Generalize linear approximation of feasible set to a tangent cone $T_X(x^*)$, or equiva-

lently, a normal cone $N_X(x^*)$ corresponding to the right-hand side of (12.1). This is studied in Chapter 12.1.

- Generalize smooth functions to extended real-valued convex functions. We can then treat a constrained minimization of a real-valued function as an unconstrained minimization of an extended real-valued function. This is studied in Chapter 12.2.
- Generalize gradients $\nabla f(x), \nabla g(x), \nabla h(x)$ to subgradients $\partial f(x), \partial g(x), \partial h(x)$. A convex function is always continuous and subdifferentiable in the relative interior of its effective domain. This is studied in Chapter 12.3.

In the remainder of this chapter we use these convex analysis tools to generalize the optimality conditions of Chapter 8.3 to a nonsmooth setting by replacing gradients with subdifferentials. Specifically we answer Q1 in Chapters 12.4 (the Saddle Point Theorem) and 12.5 (the KKT Theorem), and Q2 in Chapters 12.6 (primal optimality) and 12.7 (strong duality and dual optimality). Finally in Chapter 12.8 we apply the general theory developed in Chapters 12.4–12.7 to special classes convex optimization problems widely used in applications.

The topic of nonsmooth convex optimization is extensive. We only summarize key concepts and techniques, mostly from [54, 141], and use them to answer these questions. We include some (but not all) of the proofs to illustrate common techniques useful for nonsmooth convex optimization. Nonsmooth problems arise in many contexts. For example the dual function of a constrained convex optimization may take infinite values and may not be differentiable (but always concave) even if the cost and constraint functions are real-valued and continuously differentiable. This is because the primal minimizer of the Lagrangian may not be unique. In Chapter 13 we study stochastic optimization where some parameters of an optimization problem may be uncertain or random. These problems are generally intractable, but some have convex reformulation. Many of these reformulated problems however may not be differentiable. For example the two-stage optimization with recourse studied in Chapter 13.4 takes the following form: $\inf_x f^1(x) + Q(x)$ s.t. $h^1(x) \le 0$ where $Q(x) := E_\omega \left(\inf_{y(\omega)} \left\{f^2(x, y(\omega)) : h^2(x, y(\omega)) \le 0\right\}\right)$ and E_ω denotes expectation with respect to a random variable ω. The function $Q(x)$ is generally nondifferentiable even if (f^1, h^1) and (f^2, h^2) are continuously differentiable; moreover $Q(x)$ may be $\pm\infty$ even if (f^1, h^1) and (f^2, h^2) are real-valued. When these functions are convex, however, so is $Q(x)$. (Like a dual function, $Q(x)$ is defined by a minimization.)

12.1 Normal cones of feasible sets

In this section we develop concepts that linearly approximate a set X as the smallest convex cone (tangent cone) that contains X and, equivalently, the convex set (normal cone) that is "most opposite" to this linear approximation. The normal cone is central in optimality conditions of convex optimization as we will see. In Chapter 12.1.1 we

define polar cone that formalizes the notion of the "most opposite" directions to a set X. We use it in Chapter 12.1.2 to define the tangent cone and normal cone of X. In Chapter 12.1.3 we study how normal cones are transformed as X undergoes affine transformation. This is used to derive the normal cones of second-order constraints in Chapter 12.1.4.

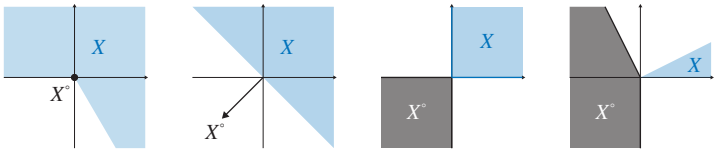
12.1.1 Polar cone

Recall the definition of relative interior, convex sets, closed convex cones, and second-order cones studied in Chapters 8.1.1, 8.1.2 and 8.2.1.

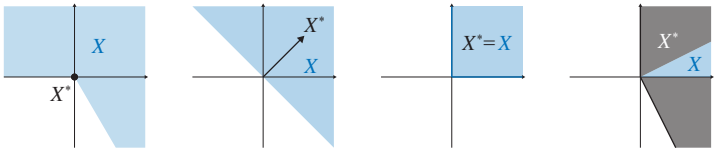
Definition 12.1 (Polar cone and dual cone)**.** Let $X \subseteq \mathbb{R}^n$ be a nonempty set.

1 The *polar cone of* X is $X^\circ := \{y \in \mathbb{R}^n : y^\mathsf{T} x \leq 0 \ \ \forall x \in X\}$.
2 The *dual cone of* X is $X^* := -X^\circ = \{y \in \mathbb{R}^n : y^\mathsf{T} x \geq 0 \ \ \forall x \in X\}$.
3 A cone K is called *self-dual* if $K^* = K$. □

It is clear that X° and X^* are indeed cones for arbitrary X, i.e., if y is in X° or X^*, so is γx for any $\gamma > 0$. Informally, the polar cone of X is the set of points that is "most opposite to the entire set X" or "most away from the entire set X". The dual cone of X is the set that is "most aligned with the entire set X" or "closest to the entire set X". The dual cone is used to define the dual problem of a conic program where the nonlinear constraint is specified abstractly by $x \in K$ for a general closed convex cone K; see Chapter 12.8.4. These cones are illustrated in Figure 12.1. The examples in the figure show that X^* can be a subset or a superset of X or equal to X. Some properties



(a) Polar cones X° of X



(b) Dual cones $X^* = -X^\circ$ of X

Figure 12.1 Polar cones and dual cones of $X \subseteq \mathbb{R}^n$. For the leftmost set X which is nonconvex, both its polar cone and dual cone contain only the origin. The other three sets X are closed convex cones and therefore $(X^\circ)^\circ = X$. Note that $(X^\circ)^\circ \neq -X^\circ = X^*$ unless X is self-dual.

of polar cones are given in the following result (see e.g. [54, Proposition 2.2.1, p.100]).

Proposition 12.1. Let $X \subseteq \mathbb{R}^n$ be a nonempty set.

1 Its polar cone X° is a closed convex cone.
2 $X^\circ = (\mathrm{cl}(X))^\circ = (\mathrm{conv}(X))^\circ = (\mathrm{cone}(X))^\circ$.
3 If $X \subseteq Y$ then $Y^\circ \subseteq X^\circ$.
4 If X is a cone then $(X^\circ)^\circ = \mathrm{cl}(\mathrm{conv}(X))$. If X is a closed convex cone then $(X^\circ)^\circ = X$.

□

Figure 12.1 shows the polar cones of sets X that contain the origin. For a set X whose closure $\mathrm{cl}(X)$ does not contain the origin, its polar cone X° is the same as the polar cone of $\mathrm{cone}(X)$ according to Proposition 12.1, as illustrated in Figure 12.2.

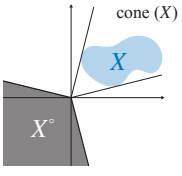


Figure 12.2 Polar cone $X^\circ = \mathrm{cone}^\circ(X)$ according to Proposition 12.1.

Example 12.1. Fix an $\bar{x} \in X^\circ$. By definition $\bar{x}^\mathsf{T} x \leq 0$ for all $x \in X$. Can there be an $x \in X^\circ$ such that $\bar{x}^\mathsf{T} x \leq 0$?

Solution. Yes if $X^\circ \not\subseteq X$. Consider $X := \{x \in \mathbb{R}^2 : x_1 > 0, x_2 = 0\}$. Then $X^\circ = \{x \in \mathbb{R}^2 : x_1 \leq 0\}$. An example is $\bar{x} := (0,-1) \in X^\circ$ and $x := (0,1) \in X^\circ$. □

12.1.2 Normal cone and tangent cone

Let $\bar{x} \in X \subseteq \mathbb{R}^n$. The *cone of feasible directions of X at $\bar{x}$* (or the radial cone) is, from Definition 8.5,

$$\mathrm{cone}(X - \bar{x}) \;:=\; \left\{\textstyle\sum_{i=1}^m \alpha_i (x_i - \bar{x}) : x_i \in X, \alpha_i \geq 0, \text{ integers } m > 0\right\}$$

It is the set of directions $x - \bar{x}$ and their convex combinations along which an infinitesimal step from $\bar{x}$ will stay in X. It is closed if and only if X is closed. The closure of $\mathrm{cone}(X - \bar{x})$ can be interpreted as a "linear approximation" of the set X at the point $\bar{x} \in X$ in that it is the smallest convex cone that contains all the feasible directions $x - \bar{x}$ at $\bar{x}$. For a smooth function f, the first-order Taylor expansion $\hat{f}(x) := f(\bar{x}) + \frac{\partial f}{\partial x}(\bar{x})(x - \bar{x})$ approximates f locally at $\bar{x}$ by a supporting hyperplane. For a "smooth" set X, the closed convex cone $\mathrm{cl}(\mathrm{cone}(X - \bar{x}))$, called a tangent cone, approximates the set X

locally at $\bar{x}$ by a halfspace associated with the supporting hyperplane at $\bar{x}$ (see Figure 12.4 below).

The notion of normal cone and tangent cone is fundamental to nonsmooth optimization. It suffices for our purposes to adopt the following definition.

Definition 12.2. Let $X \subseteq \mathbb{R}^n$ be a nonempty set and $\bar{x} \in X$.

1 The *tangent cone of X at $\bar{x}$* is the closure of the feasible direction cone of X at $\bar{x}$:

$$T_X(\bar{x}) := \mathrm{cl}\,(\mathrm{cone}(X-\bar{x}))$$

2 The *normal cone of X at $\bar{x}$* is the polar cone of the feasible direction cone of X at $\bar{x}$:

$$N_X(\bar{x}) := (\mathrm{cone}(X-\bar{x}))^\circ = (X-\bar{x})^\circ = \{y \in \mathbb{R}^n : y^\mathsf{T}(x-\bar{x}) \le 0\ \forall x \in X\}$$

□

Proposition 12.1 implies that the normal cone and the tangent cone are the polar cones of each other. The second equality in Definition 12.2 of normal cone also follows from Proposition 12.1. An equivalent definition for tangent cone of X at $\bar{x}$ is

$$T_X(\bar{x}) := \{0\} \cup \left\{ y \neq 0 : \exists x_k \in X \text{ s.t. } x_k \neq \bar{x}, x_k \to \bar{x}, \frac{x_k - \bar{x}}{\|x_k - \bar{x}\|} \to \frac{y}{\|y\|} \right\}$$

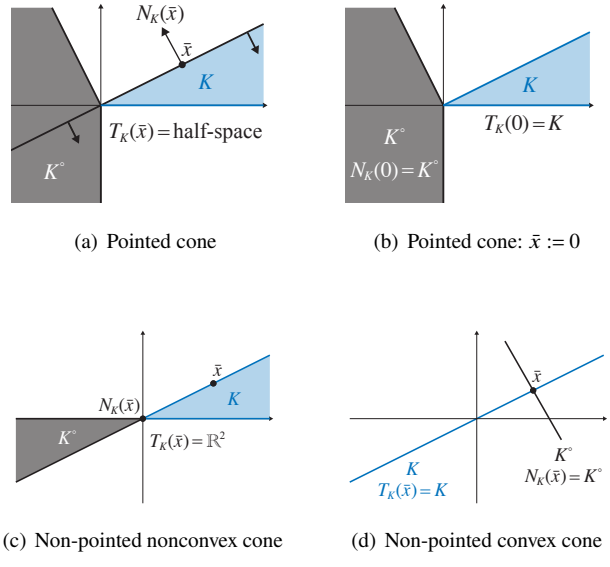
This definition is often used from which $T_X(\bar{x}) = \mathrm{cl}\,(\mathrm{cone}(X-\bar{x}))$ can be derived.

Proposition 12.2. Let $X \subseteq \mathbb{R}^n$ be a nonempty set and $\bar{x} \in X$.

1 The polar cone, dual cone, tangent cone, and normal cone are closed convex cones, even if X is neither closed nor convex.
2 $(T_X(\bar{x}))^\circ = N_X(\bar{x})$ and $T_X(\bar{x}) = (N_X(\bar{x}))^\circ$.
3 If $\bar{x} \in \mathrm{int}(X)$ then $N_X(\bar{x}) = \{0\}$ and $T_X(\bar{x}) = \mathbb{R}^n$. □

Proposition 12.2 is proved in Exercise 12.2. While a polar cone X° and a dual cone $X^* = -X^\circ$ are sets with respect to the entire set X, a normal cone $N_X(\bar{x})$ and a tangent cone $T_X(\bar{x})$ are set-valued functions whose values generally depend on their argument $\bar{x} \in X$. If $0 \in X$ then $X^\circ = N_X(0)$. Note that $T_X(\bar{x})$ is generally different from the dual cone $(X-\bar{x})^* = \{y \in \mathbb{R}^n : y^\mathsf{T}(x-\bar{x}) \ge 0, \forall x \in X\}$ (Exercise 12.2). If $\bar{x} \in \mathrm{ri}(K)$, instead of $x \in \mathrm{int}(X)$ as in Proposition 12.2, then $N_K(\bar{x})$ may not be $\{0\}$. For example, $K := \{(x_1, 0) \in \mathbb{R}^2 : x_1 \ge 0\}$ and $\bar{x} := (1,0) \in \mathrm{ri}(K)$ at which $N_K(\bar{x}) = \{(0, x_2) : x_2 \in \mathbb{R}\}$. The normal cones and tangent cones of three closed cones K at different boundary points $\bar{x}$ are illustrated in Figure 12.3.

Remark 12.1 (Linear approximation and optimality)**.** 1 A tangent cone $T_X(\bar{x}) = \mathrm{cl}(\mathrm{cone}(X-\bar{x}))$ locally approximates the set X at $\bar{x} \in X$ by the smallest closed convex cone containing all the feasible directions $x - \bar{x}$. Its polar cone, the normal



(a) Pointed cone (b) Pointed cone: $\bar{x} := 0$

(c) Non-pointed nonconvex cone (d) Non-pointed convex cone

Figure 12.3 Normal and tangent cones of closed cones $K \subseteq \mathbb{R}^2$ at a boundary point $\bar{x}$ (see Exercise 12.2 for derivation).

cone $N_X(\bar{x}) = (T_X(\bar{x}))^\circ$, specifies the directions $x - \bar{x}$ that are "most opposite to" or "most away from" the linear approximation $T_X(\bar{x})$.

2 If X is "smooth" at $\bar{x}$ then $T_X(\bar{x})$ is a halfspace associated with the supporting hyperplane at $\bar{x}$ and $N_X(\bar{x})$ is a singleton; see Figure 12.4.

3 For convex constrained optimization, the first order optimality condition says that x^* is a minimizer if and only if the direction of cost reduction aligns with $N_X(x^*)$, i.e., $-\nabla f(x^*) \in N_X(x^*)$. In a smooth setting, this takes the form $-\nabla f(x^*) = \nabla g(x^*)\lambda^* + \nabla h(x^*)\mu^*$, with the right-hand side being the singleton $N_X(x^*)$. We generalize this to the nonsmooth setting in Theorem 12.21 of Chapter 12.5. □

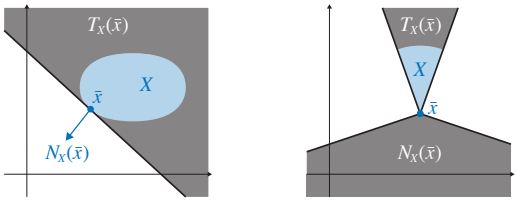


Figure 12.4 The tangent cones $T_X(\bar{x}) = \text{cl}(\text{cone}(X - \bar{x}))$ and the normal cones $N_X(\bar{x}) = \text{cone}^\circ(X - \bar{x})$ of X at $\bar{x}$. At $\bar{x}$ where the boundary of X is "smooth", the left panel illustrates the importance of "cl" in the definition of $T_X(\bar{x})$ and why $N_X(\bar{x})$ is a singleton.

Hyperplane, polyhedron, convex cone and convex set.

Recall from Chapter 8.1.2 that a hyperplane (or an intersection of hyperplanes) is a set $H_1 := \{x \in \mathbb{R}^n : Ax = b\}$ specified by a finite number of affine equalities with $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. A polyhedral set, or a *polyhedron*, is a set $H_2 := \{x \in \mathbb{R}^n : Ax \le b\}$ specified by a finite number of affine inequalities. A hyperplane H_1 is not a cone unless $b = 0$. Its normal cone $N_{H_1}(\bar{x})$ is independent of $\bar{x}$, unlike the normal cone of a polyhedron H_2 or a general convex cone. To avoid triviality we often assume implicitly these sets are nonempty.

The normal cones of hyperplanes, polyhedrons, general convex cones or convex sets specified by convex functions are particularly useful, so we derive them here. They give rise to optimal Lagrange multipliers in constrained convex optimization problems, as well as complementary slackness for inequality constraints at optimality. For specific convex programs in Chapter 12.8, substituting the expressions of the normal cones in Theorems 12.3 and 12.4 into the optimality condition in Theorem 12.21 leads to the KKT conditions for these convex programs. Recall that, for a matrix $A \in \mathbb{R}^{m\times n}$, $\text{cone}(A) := \{A\lambda : \lambda \in \mathbb{R}^n_+\} \subseteq \mathbb{R}^m$ is the set of nonnegative linear combinations of the columns of A.

Theorem 12.3 (Normal cones)**.** Given $A \in \mathbb{R}^{m\times n}$, let $H_1 := \{x \in \mathbb{R}^n : Ax = b\}$ be a hyperplane and $H_2 := \{x \in \mathbb{R}^n : Ax \le b\}$ be a polyhedron. Let $K_+ := \{x \in \mathbb{R}^n : x \ge 0\}$ be the nonnegative quadrant, and $K \subseteq \mathbb{R}^n$ a convex cone. Then

1 $N_{H_1}(\bar{x}) = \text{range}(A^\mathsf{T}) = \{A^\mathsf{T}\lambda \in \mathbb{R}^n : \lambda \in \mathbb{R}^m\}$ for any $\bar{x} \in H_1$.
2 $N_{H_2}(\bar{x}) = \text{cone}\left(A^\mathsf{T}_{I(\bar{x})}\right) = \{A^\mathsf{T}\lambda \in \mathbb{R}^n : \lambda \in \mathbb{R}^m_+,\ \lambda^\mathsf{T}(A\bar{x} - b)\}$ for any $\bar{x} \in H_2$ where $I(\bar{x}) := \{i : A\bar{x} = b\}$ is the set of active constraints. In particular $N_{H_2}(\bar{x}) = \{0\}$ if $A\bar{x} < b$.
3 $N_{K_+}(\bar{x}) = \{y \in \mathbb{R}^n : y \le 0,\ y^\mathsf{T}\bar{x} = 0\}$ for any $\bar{x} \in K_+$. In particular $N_{K_+}(\bar{x}) = \{0\}$ if $\bar{x} > 0$ and $N_{K_+}(0) = K_- := \{y \in \mathbb{R}^n : y \le 0\}$.
4 $N_K(\bar{x}) = \{y \in K^\circ : y^\mathsf{T}\bar{x} = 0\}$ for any $\bar{x} \in K$, where $K^\circ := \{y \in \mathbb{R}^n : y^\mathsf{T}x \le 0\ \ \forall x \in K\}$ is the polar cone of K. If $0 \in K$ then $N_K(0) = K^\circ$.

Proof 1 By definition

$$N_{H_1}(\bar{x}) \ = \ \{y \in \mathbb{R}^n : y^\mathsf{T}(x - \bar{x}) \le 0 \ \ \forall x \text{ s.t. } Ax = b\}$$

Since $x, \bar{x} \in H_1$, $A(x - \bar{x}) = 0$. Hence we can replace $x - \bar{x}$ for all $x \in H_1$ by all x in $\text{null}(A)$ to get

$$N_{H_1}(\bar{x}) \ = \ \{y \in \mathbb{R}^n : y^\mathsf{T}x \le 0 \ \ \forall x \text{ s.t. } Ax = 0\}$$

Since if $x \in \text{null}(A)$ then $-x \in \text{null}(A)$, we must have $y^\mathsf{T}x = 0$ for all $x \in \text{null}(A)$.[1] Hence $y \in \text{range}(A^\mathsf{T})$, i.e., $N_{H_1}(\bar{x}) = \{A^\mathsf{T}\lambda \in \mathbb{R}^n : \lambda \in \mathbb{R}^m\}$.

[1] More explicitly, for any $x \in H_1$ so that $A(x - \bar{x}) = 0$, the vector $x' := 2\bar{x} - x$ is also in H_1 since $Ax' = b$; moreover $A(x' - \bar{x}) = A(\bar{x} - x)$.

3 We prove parts 3 and 4 first. For $K_+ := \{x \in \mathbb{R}^n : x \geq 0\}$ we have

$$N_{K_+}(\bar{x}) \;=\; \{y \in \mathbb{R}^n : y^\mathsf{T}(x-\bar{x}) \leq 0 \;\; \forall x \geq 0\}$$

If $\bar{x} > 0$ (i.e., $\bar{x}$ is an interior point), then $x := \bar{x} + te_j$ for $t \in \mathbb{R}$ with small enough $|t|$ (where e_j is the unit vector with 1 in the jth entry and 0 elsewhere) ensures $y^\mathsf{T}(x-\bar{x}) = ty_j \leq 0$. As t can be negative or positive, we must have $y_j = 0$. Hence $N_{K_+}(\bar{x}) = \{0\}$ if $\bar{x} > 0$. If $\bar{x}$ is a boundary point of K_+ with $\bar{x}_j = 0$ for $j \in J \subseteq \{1,\ldots,n\}$ and $\bar{x}_j > 0$ for $j \notin J$, then the same reason implies $y \in N_{K_+}(\bar{x})$ will have $y_j = 0$ for $j \notin J$. For $j \in J$, using $x := te_j$ for any $t > 0$ gives $y^\mathsf{T}(x-\bar{x}) = ty_j \leq 0$, i.e., $y_j \leq 0$. Putting all this together we have $N_{K_+}(\bar{x}) := \{y \in \mathbb{R}^n : y \leq 0,\; y^\mathsf{T}\bar{x} = 0\}$.

4 For a general convex cone $K \subseteq \mathbb{R}^n$ (which includes K_+ as a special case if K is closed), we have

$$N_K(\bar{x}) \;:=\; \{y \in \mathbb{R}^n : y^\mathsf{T}(x-\bar{x}) \leq 0 \;\; \forall x \in K\}$$

Since K is a cone and $\bar{x} \in K$, $x := \gamma\bar{x} \in K$ for any $\gamma > 0$. Hence any $y \in N_K(\bar{x})$ must satisfy $y^\mathsf{T}(x-\bar{x}) = (\gamma-1)y^\mathsf{T}\bar{x} \leq 0$. Since γ can be chosen to be greater or smaller than 1 (as long as $\bar{x} \neq 0$) we must have $y^\mathsf{T}\bar{x} = 0$ (even if $\bar{x} = 0$). Then y satisfies $y^\mathsf{T}x \leq 0$ $\forall x \in K$, i.e., y is in the polar cone K° of K. This shows that $N_K(\bar{x}) \subseteq \{y \in K^\circ : y^\mathsf{T}\bar{x} = 0\}$. For the converse let $y \in K^\circ$ with $y^\mathsf{T}\bar{x} = 0$. Then clearly $y^\mathsf{T}(x-\bar{x}) \leq 0$ for all $x \in K$, i.e., $y \in N_K(\bar{x})$.

2 By definition

$$N_{H_2}(\bar{x}) \;:=\; \{y \in \mathbb{R}^n : y^\mathsf{T}(x-\bar{x}) \leq 0 \;\; \forall x \text{ s.t. } Ax \leq b\}$$

where $A\bar{x} \leq b$. Let

$$Y(\bar{x}) \;:=\; \{A^\mathsf{T}\lambda \in \mathbb{R}^n : \lambda \in \mathbb{R}^m_+,\; \lambda^\mathsf{T}(A\bar{x}-b) = 0\} \tag{12.2}$$

We will prove $N_{H_2}(\bar{x}) = Y(\bar{x})$. Suppose $y := A^\mathsf{T}\lambda \in Y(\bar{x})$. Then, for any x with $Ax \leq b$,

$$y^\mathsf{T}(x-\bar{x}) \;=\; \lambda^\mathsf{T}A(x-\bar{x}) \;=\; \lambda^\mathsf{T}(Ax-b) \;\leq\; 0$$

where the last inequality follows because $\lambda \geq 0$ and $Ax \leq b$. Therefore $y \in N_{H_2}(\bar{x})$.

Conversely suppose $y \in N_{H_2}(\bar{x})$. Let $I := I(\bar{x}) := \{i : a_i^\mathsf{T}\bar{x} = b_i\}$ where $a_i^\mathsf{T} \in \mathbb{R}^n$ are the rows of A and $a_i^\mathsf{T}\bar{x} < b_i$ for $i \notin I$. If $I = \emptyset$, i.e., $\bar{x} \in \text{int}\,(H_2)$, then the usual argument shows that $N_{H_2}(\bar{x}) = \{0\}$. Specifically there exists t with $|t| > 0$ such that $x := \bar{x} + te_i \in H_2$ and hence $y^\mathsf{T}(x-\bar{x}) = ty_i \leq 0$ implies $y_i = 0$ since t can be positive or negative, i.e., $N_{H_2}(\bar{x}) = \{0\}$. On the other hand, $\bar{x} \in \text{int}(H_2)$ implies that $\lambda = 0$ in the definition of $Y(\bar{x})$ and $Y(\bar{x}) = \{0\}$. Hence $N_{H_2}(\bar{x}) = Y(\bar{x})$.

We now prove $y \in Y(\bar{x})$ when $I \neq \emptyset$, by contradiction. Suppose $y \notin Y(\bar{x})$. We will construct a point $x := \bar{x} + \Delta x$ such that $Ax \leq b$ but $y^\mathsf{T}(x-\bar{x}) = y^\mathsf{T}\Delta x > 0$, contradicting that $y \in N_{H_2}(\bar{x})$ and proving $y \in Y(\bar{x})$. We first claim that there exist a nonzero $c \in \mathbb{R}^n$ such that

$$c^\mathsf{T}a_i \;\leq\; 0 \;<\; c^\mathsf{T}y, \qquad \forall i \in I \tag{12.3}$$

Indeed, since $Y(\bar{x})$ is a closed convex set, the Separating Hyperplane Theorem 8.10 says that there exists a nonzero $c \in \mathbb{R}^n$ such that (from (8.22b)):

$$c^\mathsf{T} A^\mathsf{T} \lambda \;<\; c^\mathsf{T} y, \qquad \forall \lambda \geq 0 \;\text{ s.t. }\; \lambda^\mathsf{T}(A\bar{x} - b) = 0 \tag{12.4}$$

Note that the unit vector $e_i \in \{0,1\}^m$ satisfies $e_i^\mathsf{T}(A\bar{x} - b) = 0$ for $i \in I$. Substituting into (12.4) we have

$$t_i c^\mathsf{T} a_i \;=\; c^\mathsf{T} A^\mathsf{T}(t_i e_i) \;<\; c^\mathsf{T} y, \qquad \forall t_i \geq 0,\; i \in I$$

where a_i^T are the ith row of A. Since this holds for all $t_i \geq 0$, (12.3) follows. (Also see Remark 12.2 for a more direct derivation of (12.3).)

Consider then $x(t) := \bar{x} + tc$. We have

$$Ax(t) \;=\; A\bar{x} + t(Ac) \;=\; \begin{bmatrix} A_I\bar{x} + t(A_I c) \\ A_{-I}\bar{x} + t(A_{-I} c) \end{bmatrix}$$

where for any matrix (or vector) M, M_I and M_{-I} denotes the submatrices of M consisting of its rows $i \in I$ and $i \notin I$ respectively. For $i \in I$, (12.3) implies that $A_I\bar{x} + t(A_I c) \leq A_I\bar{x} \leq b_I$. For $i \notin I$, since $A_{-I}\bar{x} < b_{-I}$, there exists small enough $t > 0$ such that $A_{-I}\bar{x} + t(A_{-I} c) \leq b_{-I}$. Hence $Ax(t) \leq b$. Yet, $y^\mathsf{T}(x(t) - \bar{x}) = ty^\mathsf{T} c > 0$ from (12.3), contradicting that $y \in N_{H_2}(\bar{x})$. This completes the proof of part 2.

□

For a general cone K, if $\bar{x} \in \text{int}(K)$, then $N_K(\bar{x}) = \{0\}$ for the same reason as in the proof above for $N_{K_+}(\bar{x})$. Part 3 is a special case of part 2 with $A = -\mathbb{I}$ and $b = 0$. It is also a special case of part 4 with $K = K_+$ and $K_+^\circ = \{y \in \mathbb{R}^n : y \leq 0\}$.

Remark 12.2 (Farkas Lemma (Theorem 8.12))**.** Note that $Y(\bar{x})$ is a convex *cone* because (12.2) is equivalent to

$$Y(\bar{x}) \;:=\; \{A_I^\mathsf{T}\lambda_I \in \mathbb{R}^n : \lambda_I \in \mathbb{R}_+^{|I|}\}$$

where $I := I(\bar{x}) := \{i : a_i^\mathsf{T}\bar{x} = b_i\}$ is the set of active constraints and A_I is the submatrix of A consisting of rows in I. If $y \notin Y(\bar{x})$, then (12.3) follows directly from the Farkas Lemma (Theorem 8.12). We derive (12.3) from the Separating Hyperplane Theorem (which underlies the Farkas Lemma) because, while Farkas Lemma applies to a convex cone, the Separating Hyperplane Theorem applies more broadly to a convex set (see Remark 12.5). □

Example 12.2 ($N_{H_1}(\bar{x})$ and $N_{H_2}(\bar{x})$)**.** Let

$$A \;:=\; \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad\qquad b \;:=\; \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Then $Ax = b$ defines the hyperplane $H_1 := \{x \in \mathbb{R}^3 : x_1 + x_2 = 1,\; x_3 = 2\}$. Its normal cone is the span of the columns of A^T independent of $\bar{x} \in H_1$:

$$N_{H_1}(\bar{x}) \;=\; \{y \in \mathbb{R}^3 : y = A^\mathsf{T}\lambda \text{ for some } \lambda \in \mathbb{R}^2\} \;=\; \left\{ \begin{bmatrix} \lambda_1 \\ \lambda_1 \\ \lambda_2 \end{bmatrix} : \lambda_i \in \mathbb{R} \right\}$$

Consider the polyhedron $H_2 := \{x \in \mathbb{R}^3 : x_1 + x_2 \le 1,\ x_3 \le 2\}$ and $\bar{x} := (0.5, 0.5, 0) \in H_2$. Then $I := I(\bar{x}) = \{1\}$. According to Theorem 12.3 its normal cone is in the cone of the columns of A^T with complementary slackness:

$$N_{H_2}(\bar{x}) \;=\; \{y \in \mathbb{R}^3 : y = A^\mathsf{T}\lambda \text{ for some } \lambda_1 \ge 0, \lambda_2 = 0\} \;=\; \left\{ \begin{bmatrix} \lambda_1 \\ \lambda_1 \\ 0 \end{bmatrix} : \lambda_1 \ge 0 \right\}$$

□

The proof of part 2 of Theorem 12.3 for polyhedron H_2 constructs a feasible point $x(t) := \bar{x} + tc$ in order to prove $y \in N_{H_2}(\bar{x}) \Rightarrow y \in Y(\bar{x})$. This relies on the fact that, when $h(x) := Ax - b$ is affine, $(\nabla h_i(\bar{x}))^\mathsf{T}(tc) = t(a_i^\mathsf{T} c) \le 0$ for all $t > 0$ and hence $h(x(t)) \le 0$ (the corresponding feature for the proof of part 4 for a convex cone K is that $\gamma\bar{x} \in K$ for any $\gamma > 0$). When $C := \{x \in \mathbb{R}^n : h(x) \le 0\}$ is defined by a nonlinear convex function $h(x)$, $x(t) := \bar{x} + tc$ may no longer be adequate because, for $i \in I(\bar{x})$, $(\nabla h_i(\bar{x}))^\mathsf{T} c$ may be zero and $h(x(t)) \le 0$ may not hold, as the next example illustrates. It explains why constraint qualification is needed for nonlinear constraints in convex optimization, but unnecessary for polyhedral constraints.

Example 12.3 (Inadequacy of $\bar{x} + tc$)**.** Let $C := \{x \in \mathbb{R}^n : h(x) \le 0\}$ where $h : \mathbb{R}^2 \to \mathbb{R}^2$ are given by $h_i(x_1, x_2) := \frac{1}{2}\left(x_1^2 + a_i x_2^2\right) - b_i$, $i = 1, 2$ with $a_i > 0, b_i > 0$ and $b_1/a_1 < b_2/a_2$; see Figure 12.5. Let $\bar{x} := \left(0, \sqrt{2b_1/a_1}\right)$. Then $h_1(\bar{x}) = 0$, $h_2(\bar{x}) < 0$,

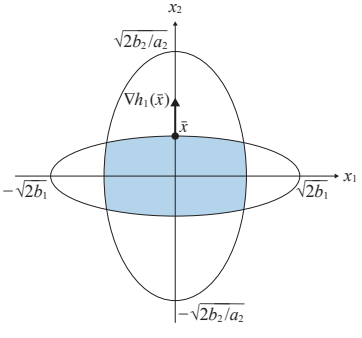


Figure 12.5 Example 12.3.

$I := I(\bar{x}) = \{1\}$, and $\nabla h_1(\bar{x}) := (\bar{x}_1, a_1\bar{x}_2) = (0, \sqrt{2a_1b_1})$. Consider $y := (y_1, y_2)$. Suppose $y \in N_C(\bar{x})$, but $y \notin Y(\bar{x})$ where

$$Y(\bar{x}) \;:=\; \left\{ \sum_{i \in I(\bar{x})} \lambda_i \nabla h_i(\bar{x}) : \lambda_i \ge,\ i \in I(\bar{x}) \right\} \;=\; \{y := (0, \lambda) : \lambda \ge 0\}$$

To derive a contradiction, the separating hyperplane argument in the proof of Theorem 12.3 for H_2 shows that there exists a nonzero $c \in \mathbb{R}^n$ such that (as for (12.3))

$$c^\mathsf{T} \nabla h_i(\bar{x}) \;\le\; 0 \;<\; c^\mathsf{T} y, \qquad \forall i \in I$$

Since $Y(\bar{x})$ is a closed convex cone, the vector c that defines the separating hyperplane is (from (8.23a)):

$$c \;:=\; \frac{y-\hat{y}}{\|y-\hat{y}\|_2} \;\neq\; 0$$

where $\hat{y}$ is the projection of y onto $Y(\bar{x}) := \{\sum_{i\in I} \lambda_i \nabla h_i(\bar{x}) : \lambda_i \geq 0, i \in I\}$:

$$\hat{y} \;=\; \sum_{i\in I} \frac{y^{\mathsf{T}} \nabla h_i(\bar{x})}{\|\nabla h_i(\bar{x})\|_2^2} \nabla h_i(\bar{x}) \;=:\; \sum_{i\in I} \hat{\lambda}_i \frac{\nabla h_i(\bar{x})}{\|\nabla h_i(\bar{x})\|_2}$$

i.e., the coefficient $\hat{\lambda}_i$ of the unit vector $\nabla h_i(\bar{x})/\|\nabla h_i(\bar{x})\|_2$ is the projection of y onto this unit vector. Therefore c, being $y-\hat{y}$ normalized, is a unit vector that is orthogonal to $Y(\bar{x}) = \{(0,\lambda) : \lambda \geq 0\}$, i.e., $c = (1,0)$. Since $c^{\mathsf{T}} \nabla h_1(\bar{x}) = 0$, for any $t > 0$,

$$h_1(x(t)) \;=\; h_1(\bar{x}) + t \frac{\partial h_1}{\partial x}(\bar{x})c + \frac{t^2}{2} c^{\mathsf{T}} \frac{\partial^2 h_1}{\partial x^2}(x(s))c \;=\; \frac{t^2}{2} c^{\mathsf{T}} \frac{\partial^2 h_1}{\partial x^2}(x(s))c \;>\; 0$$

for some $s \in [0,t]$. Hence $x(t) := \bar{x} + tc \notin Y(\bar{x})$ for any $t > 0$; see Figure 12.5.

Exercise 12.3 derives the normal cone $N_C(\bar{x})$ for this example. It also shows that constraint qualification is sufficient but not necessary for the existence of λ. □

When $C := \{x : h(x) \leq 0\}$ is a non-polyhedral convex set, a constraint qualification is needed to derive the normal cone $N_C(\bar{x})$. We next derive $N_C(\bar{x})$ under the *linear independence constraint qualification* (LICQ) discussed in Chapter 8.3.4:

$$\text{columns of } \nabla h_I(\bar{x}) \in \mathbb{R}^{n\times|I|} \text{ are linearly independent} \tag{12.5}$$

where $I := I(\bar{x}) := \{i : h_i(\bar{x}) = 0\}$ is the set of active constraints and $h_I := (h_i : i \in I)$ consists of constraint functions in I. The proof of Theorem 12.4 has three features that are useful in other applications: (i) It uses the Farkas Lemma (Theorem 8.12) or the Separating Hyperplane Theorem 8.10 to derive the inequality (12.6). (ii) It uses linear program (LP) duality to find a direction Δx for a contradiction argument. (iii) It illustrates the role of LICQ in the LP duality argument.

Theorem 12.4 (Normal cone of C)**.** Let $C := \{x \in \mathbb{R}^n : h(x) \leq 0\}$ be the convex set defined by a real-valued twice continuously differentiable function $h : \mathbb{R}^n \to \mathbb{R}^m$ that is convex on $\mathbb{R}^n$. If $\bar{x} \in C$ satisfies (12.5) then (denoting $I := I(\bar{x}) := \{i : h_i(\bar{x}) = 0\}$)

1 $N_C(\bar{x}) = \text{cone}\,(\nabla h_I(\bar{x})) = \{\nabla h(\bar{x})\lambda \in \mathbb{R}^n : \lambda \in \mathbb{R}^m_+,\ \lambda^{\mathsf{T}} h(\bar{x}) = 0\}$ for any $\bar{x} \in C$.
2 Every $y \in N_C(\bar{x})$ has a unique representation in terms of $\nabla h_I(\bar{x})$, i.e., for every $y \in N_C(\bar{x})$, there exists a unique $\lambda_I \in \mathbb{R}^{|I|}_+$ such that $y = \nabla h_I(\bar{x})\lambda_I$.

Proof By definition

$$N_C(\bar{x}) \;=\; \{y \in \mathbb{R}^n : y^{\mathsf{T}}(x-\bar{x}) \leq 0 \;\; \forall x \text{ s.t. } h(x) \leq 0\}$$

where $h(\bar{x}) \leq 0$. Define the closed convex cone

$$Y(\bar{x}) \;:=\; \{\nabla h(\bar{x})\lambda \in \mathbb{R}^n : \lambda \in \mathbb{R}^m_+,\; \lambda^{\mathsf{T}} h(\bar{x}) = 0\} \;=\; \left\{ \sum_{i \in I} \lambda_i \nabla h_i(\bar{x}) : \lambda_i \geq 0 \right\}$$

where $h_i(\bar{x}) = 0$ for $i \in I$ and $h_i(\bar{x}) < 0$ for $i \notin I$. We will show that $N_C(\bar{x}) = Y(\bar{x})$.

Suppose $y := \nabla h(\bar{x})\lambda \in Y(\bar{x})$. Since h is convex on $\mathbb{R}^n$ we have

$$h(x) - \Big(h(\bar{x}) + \nabla^{\mathsf{T}} h(\bar{x})(x - \bar{x}) \Big) \;\geq\; 0 \qquad \forall x \;\text{ s.t. }\; h(x) \leq 0$$

Multiplying both sides by $\lambda \geq 0$, we get, since $\lambda^{\mathsf{T}} h(\bar{x}) = 0$,

$$(\nabla h(\bar{x})\lambda)^{\mathsf{T}} (x - \bar{x}) \;\leq\; \lambda^{\mathsf{T}} h(x) \;\leq\; 0 \qquad \forall x \;\text{ s.t. }\; h(x) \leq 0$$

where the last inequality follows since $h(x) \leq 0$ and $\lambda \geq 0$. Hence $y \in N_C(\bar{x})$.

Conversely suppose $y \in N_C(\bar{x})$. As for $N_{H_2}(\bar{x})$, if $I = \emptyset$, then $\bar{x} \in \text{int}(C)$ and both $N_C(\bar{x})$ and $Y(\bar{x})$ are equal to $\{0\}$. Suppose then $I \neq \emptyset$ and $y \notin Y(\bar{x})$. We will show that there exists an $x(t) := \bar{x} + t\Delta x$ such that $h(x(t)) \leq 0$ but $y^{\mathsf{T}}(x(t) - \bar{x}) = t y^{\mathsf{T}} \Delta x > 0$, contradicting $y \in N_C(\bar{x})$ and proving that $y \in Y(\bar{x})$, in three steps.

Step 1: There exists c with $c^{\mathsf{T}} y > 0$. The same argument in Theorem 12.3 that derives (12.3) for the polyhedron H_2 shows that there exists a nonzero $c \in \mathbb{R}^n$ with

$$c^{\mathsf{T}} \nabla h_i(\bar{x}) \;\leq\; 0 \;<\; c^{\mathsf{T}} y, \qquad \forall i \in I \tag{12.6}$$

This is a consequence of the Farkas Lemma (Theorem 8.12) since $Y(\bar{x})$ is a convex cone (see Remark 12.2).

Step 2: Bound second-order term. For the polyhedron H_2 in Theorem 12.3, the required $\Delta x := c$. For nonlinear $h(x)$, however, $x(t) := \bar{x} + tc$ is inadequate because of the second-order term in the Taylor expansion of $h(x(t))$, as explained in Example 12.3. A more sophisticated argument is needed that uses linear programming duality in Theorem 8.23 of Chapter 8.4.2.

For each $i = 1, \ldots, m$, we have

$$h_i(x(t)) \;=\; h_i(\bar{x} + t\Delta x) \;=\; h_i(\bar{x}) + t \frac{\partial h_i}{\partial x}(\bar{x})\Delta x + \frac{t^2}{2} \Delta x^{\mathsf{T}} \frac{\partial^2 h_i}{\partial x^2}(x(s_i))\Delta x \tag{12.7}$$

for some $s_i \in [0, t]$. The last term depends on t through $x(s_i)$, but can be upper bounded by:

$$\alpha_i(\bar{x}, \Delta x) \;:=\; \max_{s_i \in [0,1]} \Delta x^{\mathsf{T}} \frac{\partial^2 h_i}{\partial x^2}(x(s_i))\Delta x$$

which is finite and independent of t, given $(\bar{x}, \Delta x)$, since h_i is twice continuously differentiable and s_i is in a compact set $[0, 1]$. Then (12.7) implies, for $i = 1, \ldots, m$,

$$h_i(x(t)) \;\leq\; h_i(\bar{x}) + t \left(\frac{\partial h_i}{\partial x}(\bar{x})\Delta x + \frac{t}{2} \alpha_i(\bar{x}, \Delta x) \right) \qquad \text{for } t \in [0, 1]$$

Hence if we can find a direction Δx such that

$$\frac{\partial h_i}{\partial x}(\bar{x})\Delta x \;<\; 0 \qquad \text{for } i \in I \tag{12.8a}$$

then, since $h_i(\bar{x}) = 0$ for $i \in I$ and $h_i(\bar{x}) < 0$ for $i \notin I$, there exists a sufficiently small $t > 0$ such that $h_i(x(t)) \le 0$ for all $i = 1, \ldots, m$. If Δx also satisfies

$$y^{\mathsf{T}}(x(t) - \bar{x}) \;=\; t y^{\mathsf{T}} \Delta x \;>\; 0 \tag{12.8b}$$

then $x(t)$ contradicts $y \in N_C(\bar{x})$ and thus proves $y \in Y(\bar{x})$.

Step 3: There exists Δx that satisfies (12.8). To find such a Δx, denote by $h_I := (h_i : i \in I)$ the vector of constraint functions h_i with $h_i(\bar{x}) = 0$. Consider the linear program

$$z^*(\epsilon) \;:=\; \min_{(\Delta x, z) \in \mathbb{R}^{n+1}} z \qquad \text{s.t.} \qquad \frac{\partial h_I}{\partial x}(\bar{x})\Delta x \le z\mathbf{1}, \;\; y^{\mathsf{T}}\Delta x \ge \epsilon \tag{12.9}$$

where the parameter $\epsilon > 0$ is to be determined, $\frac{\partial h_I}{\partial x}(\bar{x})$ and y are fixed, and $\mathbf{1}$ denotes the vector of all 1s of size $|I|$. An optimal solution $(\Delta x^*, z^*(\epsilon))$ with $z^*(\epsilon) < 0$ exists for some $\epsilon > 0$ if and only if $x^*(t) := \bar{x} + t\Delta x^*$ satisfies (12.8). We claim that the linear program (12.9) is feasible for a sufficiently small $\epsilon > 0$, because

$$\Delta x \;:=\; c, \qquad z \;:=\; \max_{i \in I} \frac{\partial h_i}{\partial x}(\bar{x}) c$$

satisfies the constraints in (12.9), where c is the vector in (12.6). Fix an $\epsilon > 0$ such that (12.9) is feasible.

We now show that the LICQ (12.5) implies that the dual of (12.9) is infeasible. Let $\lambda \in \mathbb{R}_+^{|I|}$ and $\mu \in \mathbb{R}_+$ denote the dual variables associated with the constraints in (12.9). The dual problem of (12.9) is (see Chapter 8.4.2 for details):

$$d^*(\epsilon) \;:=\; \max_{(\lambda, \mu) \in \mathbb{R}^{|I|+1}} \epsilon\mu \quad \text{s.t.} \quad \mathbf{1}^{\mathsf{T}}\lambda = 1, \;\; \nabla h_I(\bar{x})\lambda = \mu y, \;\; (\lambda, \mu) \ge 0 \tag{12.10}$$

where $\nabla h_I(\bar{x}) = \left(\frac{\partial h_I}{\partial x}(\bar{x})\right)^{\mathsf{T}}$. Suppose $(\lambda, \mu) \ge 0$ is feasible for (12.10). Then $\lambda \ne 0$. Since $\nabla h_I(\bar{x})$ has linearly independent columns (constraint qualification), $\nabla h_I(\bar{x})\lambda \ne 0$ and hence $\mu > 0$. Therefore we can write

$$y \;=\; \sum_{i \in I} \frac{\lambda_i}{\mu} \nabla h_i(\bar{x})$$

i.e., $y \in Y(\bar{x})$, contradicting the assumption that $y \notin Y(\bar{x})$. Hence the dual problem (12.10) is infeasible.

Since the primal problem is feasible but the dual problem is infeasible, linear programming duality implies that $z^*(\epsilon) = d^*(\epsilon) = -\infty$ (see Theorem 8.23 of Chapter 8.4.2). This means there exists a (finite) Δx that is feasible for (12.9) and that satisfies (12.8). This establishes the existence of an $x(t) := \bar{x} + t\Delta x$ such that $h(x(t)) \le 0$ but $y^{\mathsf{T}}(x(t) - \bar{x}) = t y^{\mathsf{T}}\Delta x > 0$, contradicting $y \in N_C(\bar{x})$ and proving that $y \in Y(\bar{x})$.

Finally, for any $y \in N_C(\bar{x})$, $y = \nabla h_I(\bar{x})\lambda_I$ for some $\lambda_I \in \mathbb{R}_+^{|I|}$. If there is another

distinct $\hat{\lambda}_I$ with $y = \nabla h_I(\bar{x})\hat{\lambda}_I$ then $\nabla h_I(\bar{x})(\hat{\lambda}_I - \lambda_I) = 0$, contradicting LICQ (12.5). Hence y has a unique representation. □

The uniqueness of λ_I in Theorem 12.4 underlies the property that LICQ (12.5) implies the uniqueness of dual optimal solution in constrained convex optimization. Constraint qualification however is sufficient, but not necessary, for the existence of λ_I in $N_C(\bar{x}) = \{\nabla h_I(\bar{x})\lambda_I \in \mathbb{R}^n : \lambda_I \in \mathbb{R}_+^{|I|}\}$; see Exercise 12.3.

Example 12.4 (Nonlinear equality constraint $g(x) = 0$)**.** Let the components g_i, $i = 1,\ldots,m$, of $g : \mathbb{R}^n \to \mathbb{R}^m$ be twice continuously differentiable convex functions. Let $X := \{x : g(x) = 0\}$. Note that X is *not* convex unless g is affine. Suppose $\bar{x} \in X$ satisfies LICQ (12.5), i.e., the columns of $\nabla g(\bar{x})$ are linearly independent. Show that

$$N_X(\bar{x}) \;=\; \text{range}(\nabla g(\bar{x})) \;:=\; \{\nabla g(\bar{x})\lambda : \lambda \in \mathbb{R}^m\}$$

Moreover for every $y \in N_X(\bar{x})$ there is a unique λ such that $y = \nabla g(\bar{x})\lambda$.

Solution. Write $g(x) = 0$ as $g(x) \le 0$ and $-g(x) \le 0$. Theorem 12.4 then implies that

$$\begin{aligned} N_X(\bar{x}) \;&=\; \left\{ \begin{bmatrix} \nabla g(\bar{x}) & -\nabla g(\bar{x}) \end{bmatrix} \begin{bmatrix} \lambda \\ \mu \end{bmatrix} \in \mathbb{R}^n : \lambda, \mu \in \mathbb{R}^m_+ \right\} \\ &=\; \left\{ (\lambda - \mu)\nabla g(\bar{x}) : \lambda, \mu \in \mathbb{R}^m_+ \right\} \;=\; \text{range}(\nabla g(\bar{x})) \end{aligned}$$

The unique representation of every $y \in N_X(\bar{x})$ also follows directly from Theorem 12.4. □

Theorems 12.3 and 12.4 derive the normal cones of common convex sets. The next result says that the normal cone of the intersection of convex sets is the sum of their individual normal cones. It is useful in deriving the normal cone of multiple constraints in an optimization problem from the normal cones of individual constraints. It is proved in Exercise 12.13 using Theorem 12.18 below (whose proof does not rely on Lemma 12.5 so there is no circular argument).

Lemma 12.5 (Normal cone of set intersection)**.** Consider polyhedral sets $C_i \subseteq \mathbb{R}^n$, $i = 1,\ldots,\bar{m}$, and convex sets $C_i \subseteq \mathbb{R}^n$, $i = \bar{m}+1,\ldots,m$, and let $C := \bigcap_{i=1}^m C_i$. If

$$\left(\cap_{i=1}^{\bar{m}} C_i\right) \bigcap \left(\cap_{i=\bar{m}+1}^{m} \text{ri}(C_i)\right) \neq \emptyset$$

then

$$N_C(\bar{x}) = \sum_i N_{C_i}(\bar{x}), \qquad \forall \bar{x} \in C$$

Summary. Theorems 12.3 and 12.4 and Example 12.4 are summarized in Table 12.1 (see Exercise 12.4 for derivation of the tangent cones). We will use these results together with Lemma 12.5 to derive KKT conditions in Chapter 12.8 for convex optimization problems widely used in applications. The intuition is explained in Remark 12.1: x^* is a minimizer if the negative cost gradient $-\nabla f(x^*)$ is in the normal cone $N_X(x^*)$ of the feasible set X at x^*. If the feasible set $X := \cap_i C_i$ is specified by multiple

Set $X \subseteq \mathbb{R}^n$	Normal cone $N_X(\bar{x}) \subseteq \mathbb{R}^n$	Tangent cone $T_X(\bar{x}) \subseteq \mathbb{R}^n$
$\{x : Ax = b\}$	$\text{range}(A^\mathsf{T}) := \{A^\mathsf{T}\lambda : \lambda \in \mathbb{R}^m\}$	$\text{null}(A) := \{y : Ay = 0\}$
$\{x : \text{convex } h(x) = 0\}$	$\text{range}(\nabla h(\bar{x})) := \{\nabla h(\bar{x})\lambda : \lambda \in \mathbb{R}^m\}$	$\text{null}(\nabla^\mathsf{T} h(\bar{x})) := \{y : \nabla^\mathsf{T} h(\bar{x})$
$\{x : Ax \le b\}$	$\text{cone}\left(A_I^\mathsf{T}\right) = \{A^\mathsf{T}\lambda \in \mathbb{R}^n : \lambda \in \mathbb{R}^m_+,\ \lambda^\mathsf{T}(A\bar{x} - b)\}$	$\{y : A_I y \le 0\}$
$\{x : \text{convex } h(x) \le 0\}$	$\text{cone}(\nabla h_I(\bar{x})) := \{\nabla h(\bar{x})\lambda : \lambda \in \mathbb{R}^m_+,\ \lambda^\mathsf{T} h(\bar{x}) = 0\}$	$\{y : \nabla^\mathsf{T} h_I(\bar{x}) y \le 0\}$
$\text{cone}\,\{x : x \ge 0\}$	$\{y \le 0 : y^\mathsf{T}\bar{x} = 0\}$	$\{y : \bar{x}_i = 0 \Rightarrow y_i \ge 0\}$
$\text{cone}\,K$	$\{y \in K^\circ : y^\mathsf{T}\bar{x} = 0\}$	$\text{cl}\{\sum_i \alpha_i(x_i - \bar{x}) : x_i \in K,\ \alpha_i$

Table 12.1 The tangent cones and normal cones of common sets. The function h is assumed to be twice continuously differentiable and convex and constraint qualification is satisfied at $\bar{x} \in X$.

constraints C_i, the optimality condition takes the form $-\nabla f(x^*) \in \sum_i N_{C_i}(x^*)$, e.g., $-\nabla f(x^*) = \nabla g(x^*)\lambda^* + \nabla h(x^*)\mu^*$ as in (12.1). The condition $y^\mathsf{T}\bar{x} = 0$ in Table 12.1 give rise to complementary slackness in KKT conditions, as we will see in Chapter 12.8. Theorem 12.4 underlies the need for constraint qualification and the uniqueness of the dual optimal solution under LICQ.

12.1.3 Affine transformation

We have derived in the previous subsection the normal cones of common sets. In this subsection we study how the normal cones are transformed when these sets undergo affine transformations. They will be applied in Chapter 12.1.4 to derive the normal cones of SOC constraints.

Linear transformation.

Consider the linear mapping $A \in \mathbb{R}^{m\times n}$, and the image $Y \subseteq \mathbb{R}^m$ and the pre-image $X \subseteq \mathbb{R}^n$ under A. We will study the relation between the normal cones $N_X(\bar{x})$ and $N_Y(A\bar{x})$. The main conclusion is that if X is the pre-image of an arbitrary set Y then $N_X(\bar{x}) = A^\mathsf{T} N_Y(A\bar{x})$. By Proposition 12.1, $X^\circ = [\text{cl}(\text{cone}(X))]^\circ = N_X(0)$ when $0 \in X$. Hence this also implies that $X^\circ = A^\mathsf{T} Y^\circ$.

Specifically given a nonempty set $X \subseteq \mathbb{R}^n$, its image under $A \in \mathbb{R}^{m\times n}$ is the set

$$Y \;:=\; AX \;:=\; \{Ax \in \mathbb{R}^m : x \in X\}$$

By definition of Y, the mapping $A : X \to Y$ is surjective, i.e., every $y \in Y$ satisfies $y = Ax$ for some $x \in X$. It is injective if A is of full column rank. In the following X° and $N_X(\bar{x})$ denote the polar cone and the normal cone of X at $\bar{x}$. The next result is used in Corollary 12.11 to derive the normal cone of a rotated second-order cone from that of a standard second-order cone.

Theorem 12.6 (Image of linear transformation)**.** Let $X \subseteq \mathbb{R}^n$ be a nonempty set and $Y := AX$ where $A \in \mathbb{R}^{m\times n}$. Suppose $\bar{x} \in X$ and $\bar{y} = A\bar{x} \in Y$. Then

1 The polar cone Y° and the normal cone $N_Y(\bar{y})$ of Y at $\bar{y}$ are the pre-images of the polar cone and the normal cone of X at $\bar{x}$ respectively under A^T:

$$Y^\circ = \{y \in \mathbb{R}^m : A^\mathsf{T} y \in X^\circ\}, \qquad N_Y(\bar{y}) = \{y \in \mathbb{R}^m : A^\mathsf{T} y \in N_X(\bar{x})\}$$

Hence $A^\mathsf{T} Y^\circ \subseteq X^\circ$ and $A^\mathsf{T} N_Y(\bar{y}) \subseteq N_X(\bar{x})$.

2 If $\operatorname{rank}(A) = n$ (full column rank) then $A^\mathsf{T} Y^\circ = X^\circ$ and $A^\mathsf{T} N_Y(\bar{y}) = N_X(\bar{x})$.

Proof For the normal cone $N_Y(\bar{y})$ we have, for any $\tilde{y} \in N_Y(\bar{y})$, $\tilde{y}^\mathsf{T}(y - \bar{y}) \le 0$ for $y = Ax \in Y$ for all $x \in X$. Then

$$\tilde{y}^\mathsf{T} A(x - \bar{x}) \le 0 \qquad \forall x \in X$$

i.e., $A^\mathsf{T}\tilde{y} \in N_X(\bar{x})$. This implies that $A^\mathsf{T} N_Y(\bar{y}) \subseteq N_X(\bar{x})$. Suppose now $\operatorname{rank}(A) = n$ so that $X = A^\dagger Y$ with $A^\dagger = (A^\mathsf{T}A)^{-1}A^\mathsf{T}$ (if $m = n$, then $A^\dagger = A^{-1}$); see Corollary A.20 of Chapter A.7. If $\tilde{x} \in N_X(\bar{x})$ then $\tilde{x}^\mathsf{T}(x - \bar{x}) = \tilde{x}^\mathsf{T} A^\dagger (y - \bar{y}) \le 0$ for all $y \in Y$, i.e., $A(A^\mathsf{T}A)^{-\mathsf{T}}\tilde{x} \in N_Y(\bar{y})$. Therefore $A(A^\mathsf{T}A)^{-\mathsf{T}}\tilde{x} = \tilde{y}$ or $\tilde{x} = A^\mathsf{T}\tilde{y}$ for some $\tilde{y} \in N_Y(\bar{y})$ because $A^\mathsf{T}\tilde{y} = A^\mathsf{T}A(A^\mathsf{T}A)^{-\mathsf{T}}\tilde{x} = \tilde{x}$. This shows that $A^\mathsf{T} N_Y(\bar{y}) \supseteq N_X(\bar{x})$ and hence $A^\mathsf{T} N_Y(\bar{y}) = N_X(\bar{x})$.

For the polar cone Y°, substitute $\bar{x} := 0$ into the above argument (whether or not $0 \in X$) to conclude that $A^\mathsf{T} Y^\circ \subseteq X^\circ$, and $A^\mathsf{T} Y^\circ = X^\circ$ if $\operatorname{rank}(A) = n$. □

Theorem 12.6 is illustrated in Figure 12.6 for the case when $\operatorname{rank}(A) = n$ so that $X^\circ = A^\mathsf{T} Y^\circ$. See Example 12.5 for a case when A is singular and $X^\circ \supsetneq A^\mathsf{T} Y^\circ$.

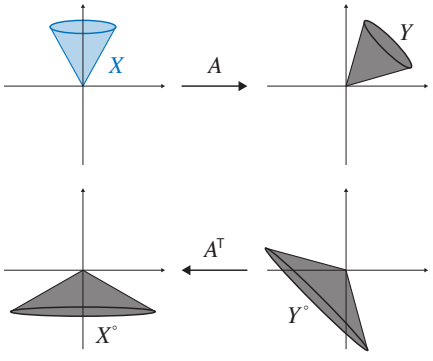


Figure 12.6 Theorem 12.6 when $\operatorname{rank}(A) = n$: linear transformation Y of a convex cone X and their polar cones $Y^\circ = N_Y(0)$ and $X^\circ = N_X(0)$ respectively.

Given a nonempty set $Y \subseteq \mathbb{R}^m$, its pre-image under $A \in \mathbb{R}^{m \times n}$ is the set

$$X := \{x \in \mathbb{R}^n : Ax \in Y\}$$

The mapping $A : X \to Y$ is not necessarily surjective, i.e., $AX \subseteq Y$ and AX can be a strict subset of Y. Moreover, if Y is the image of a given set X then Theorem 12.6 says that $N_X(\bar{x}) \subseteq A^\mathsf{T} N_Y(\bar{y})$ unless $\operatorname{rank}(A) = n$, but if X is the pre-image of a given set

Y then $N_X(\bar{x}) = A^\mathsf{T} N_Y(\bar{y})$ for *arbitrary* A, as Theorem 12.7 shows. This is because the pre-image X always contains $\mathrm{null}(A)$ whereas a given X may overlap with $\mathrm{null}(A)$ but not contain it (unless $\mathrm{rank}(A) = n$ in which case $\mathrm{null}(A) = \{0\}$). The next example illustrates this difference; see also (12.12) below.

Example 12.5 (Image vs pre-image)**.** Let $A := \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ be a singular matrix. We give a set X whose image $Y = AX$ satisfies $A^\mathsf{T} N_Y(\bar{y}) \subsetneq N_X(\bar{x})$ at $\bar{x} = 0$, and another set Y whose pre-image X satisfies $A^\mathsf{T} N_Y(\bar{y}) = N_X(\bar{x})$ at all $\bar{x} \in X$ and $\bar{y} = A\bar{x} \in Y$.

1 Consider the set X and its image Y under A:

$$X := \{x \in \mathbb{R}^2 : x \geq 0\}, \qquad Y := AX = \{Ax : x \geq 0\} = \left\{\alpha \begin{bmatrix} 1 \\ 1 \end{bmatrix} : \alpha \geq 0\right\}$$

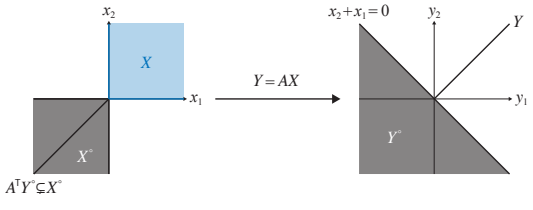
The polar cone of X is $X^\circ = \{x \in \mathbb{R}^2 : x \leq 0\}$. From Theorem 12.6 the polar cone Y° is the pre-image of X° under A^T:

$$Y^\circ = \{y \in \mathbb{R}^2 : A^\mathsf{T} y \in X^\circ\} = \{y \in \mathbb{R}^2 : y_1 + y_2 \leq 0\}$$

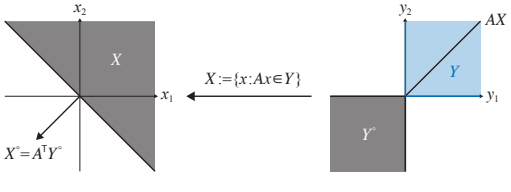
We hence have

$$A^\mathsf{T} N_Y(0) = A^\mathsf{T} Y^\circ = \{A^\mathsf{T} y : y \in Y^\circ\} = \left\{\alpha \begin{bmatrix} 1 \\ 1 \end{bmatrix} : \alpha \leq 0\right\} \subsetneq X^\circ = N_X(0)$$

as proved in Theorem 12.6. These sets are illustrated in Figure 12.7(a).



(a) Image of X



(b) Pre-image of Y

Figure 12.7 Example 12.5. Since A is singular, (a) $A^\mathsf{T} Y^\circ \subsetneq X^\circ$; (b) $A^\mathsf{T} Y^\circ = X^\circ$; moreover $A^\mathsf{T} N_Y(\bar{y}) = N_X(\bar{x})$.

2 Consider the set Y and its pre-image X under A:

$$Y := \{y \in \mathbb{R}^2 : y \geq 0\}, \qquad X := \{x \in \mathbb{R}^2 : Ax \geq 0\} = \{x : x_1 + x_2 \geq 0\}$$

Note that $AX = \{Ax : x_1 + x_2 \geq 0\} = \{\alpha(1,1) : \alpha \geq 0\} \subsetneq Y$. We have $Y^\circ = \{y : y \leq 0\}$ and

$$X^\circ \;:=\; \{x : \tilde{x}_1 + \tilde{x}_2 \geq 0 \Rightarrow x_1\tilde{x}_1 + x_2\tilde{x}_2 \leq 0\} \;=\; \{x : x_1 = x_2,\; x \leq 0\}$$

Hence, even though A is singular, $A^\mathsf{T} Y^\circ = \{A^\mathsf{T} y : y \leq 0\} = \{\alpha(1,1) : \alpha \leq 0\} = X^\circ$; see Figure 12.7(b).

Moreover all boundary points $\bar{x}$ of X, defined by $x_1 + x_2 = 0$, are mapped to $\bar{y} = A\bar{x} = 0$ which is the unique boundary point of Y. For this pair of $(\bar{x}, \bar{y})$, $A^\mathsf{T} N_Y(\bar{y}) = A^\mathsf{T} Y^\circ = X^\circ = N_X(\bar{x})$ from Theorem 12.3. On the other hand, any non-boundary point $\bar{x}$ with $x_1 + x_2 > 0$ is in $\mathrm{int}(X)$ and the corresponding $\bar{y} = A\bar{x} > 0$ is in $\mathrm{int}(Y)$, and hence $A^\mathsf{T} N_Y(\bar{y}) = N_X(\bar{x}) = \{0\}$. Therefore $A^\mathsf{T} N_Y(\bar{y}) = N_X(\bar{x})$ at any $\bar{x} \in X$ and $\bar{y} = A\bar{x}$ (including $\bar{x} = 0$). □

Part 2 of Example 12.5 shows $A^\mathsf{T} N_Y(\bar{y}) = N_X(\bar{x})$ for a specific A. Exercise 12.6 gives another example for arbitrary A but for the cone $Y = \{y \in \mathbb{R}^m : y \leq 0\}$, proved using the Farkas Lemma (Theorem 8.12). The next theorem shows that $A^\mathsf{T} N_Y(\bar{y}) = N_X(\bar{x})$ holds for an arbitrary set Y and arbitrary A. It is proved using the following property of the pseudo-inverse $A^\dagger$ from Theorem A.19 of Appendix A.7.

Consider an arbitrary real matrix $A \in \mathbb{R}^{m\times n}$ and let its singular value decomposition be $A = V\Sigma W^\mathsf{T} = V_r \Sigma_r W_r^\mathsf{T}$ where $\mathrm{rank}(A) = r$, $V \in \mathbb{R}^{m\times m}$ and $W \in \mathbb{R}^{n\times n}$ are unitary matrices partitioned so that their first r columns correspond to the r positive singular values of A (see Appendix A.7 for details)

$$V \;=\; \begin{bmatrix} V_r & V_{m-r} \end{bmatrix}, \qquad \Sigma \;=\; \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix}, \qquad W \;=\; \begin{bmatrix} W_r & W_{n-r} \end{bmatrix}$$

Its pseudo-inverse is the real matrix $A^\dagger := W\Sigma^\dagger V^\mathsf{T} = W_r \Sigma_r^{-1} V_r^\mathsf{T}$. Given an arbitrary set $X \subseteq \mathbb{R}^n$, let $Y := AX \subseteq \mathbb{R}^m$ be its image under A. Since the columns of W form an orthonormal basis of $\mathbb{R}^n$, $x = W_r \left(W_r^\mathsf{T} x\right) + W_{n-r}\left(W_{n-r}^\mathsf{T} x\right)$. This implies that every $x \in X$ has a unique orthogonal decomposition (using $A^\dagger A = W_r W_r^\mathsf{T}$):

$$x \;=\; A^\dagger \underbrace{(Ax)}_{y(x)} + W_{n-r} \underbrace{\left(W_{n-r}^\mathsf{T} x\right)}_{\beta(x)} \;=\; A^\dagger y(x) + W_{n-r}\beta(x) \tag{12.11}$$

for some unique $y(x) \in Y$ and unique $W_{n-r}\beta(x)$ in $\mathrm{null}(A) = \mathrm{range}(W_{n-r})$. This is illustrated in Figures 12.8. The first term $A^\dagger y = W_r \left(W_r^\mathsf{T} x\right) \in \mathrm{range}(W_r)$ is the projection of x onto $\mathrm{range}(A^\mathsf{T})$ and is orthogonal to the second term (Theorem A.19). As x takes values in X we write (12.11) as

$$X \;=\; A^\dagger Y + W_{n-r} B(X) \quad \text{with} \quad B(X) \;:=\; \left\{W_{n-r}^\mathsf{T} x : x \in X\right\} \subseteq \mathbb{R}^{n-r} \tag{12.12a}$$

Since $B(x)$ can be a strict subset of $\mathbb{R}^{n-r}$, $W_{n-r}B(X)$ can be a strict subset of $\mathrm{span}(W_{n-r}) = \mathrm{null}(A)$. On the other hand, given an arbitrary set $Y \subseteq \mathbb{R}^m$, let $X := \{x : Ax \in Y\} \subseteq \mathbb{R}^n$ be its pre-image under A. Then every $x \in X$ still decomposes uniquely into its orthogonal components along $\mathrm{range}(A^\mathsf{T})$ and $\mathrm{null}(A)$, but the

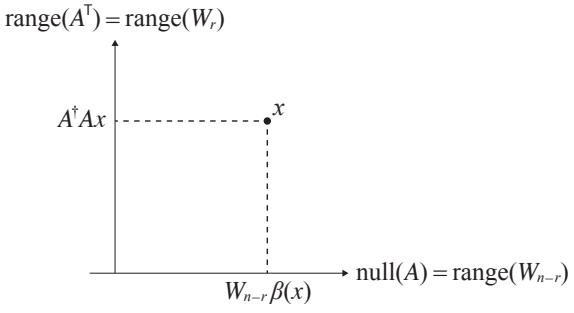


Figure 12.8 Orthogonal decomposition of $\mathbb{R}^n$ using singular value decomposition of matrix A.

set X in terms of Y becomes

$$X \;=\; A^\dagger Y + W_{n-r}\mathbb{R}^{n-r} \tag{12.12b}$$

i.e., the pre-image of each $y \in Y$ consists of $A^\dagger y$ plus the whole subspace null(A) = range(W_{n-r}). In contrast the pre-images of $y \in Y$ in (12.12a) consists of $A^\dagger y$ plus a subset of null(A). This underlies the difference between Theorems 12.6 and 12.7. When A has a full column rank, null$(A) = \{0\}$ and $X = A^\dagger Y$ in both (12.12a) and (12.12b).

Theorem 12.7 (Pre-image of linear transformation)**.** Let $Y \subseteq \mathbb{R}^m$ be a nonempty set and $X := \{x \in \mathbb{R}^n : Ax \in Y\}$ be its pre-image under $A \in \mathbb{R}^{m\times n}$. Then $N_X(\bar{x}) = A^\mathsf{T} N_Y(\bar{y})$ for any $\bar{x} \in X$ and $\bar{y} = A\bar{x} \in Y$.

Proof Given any $\tilde{y} \in N_Y(\bar{y})$, $\tilde{y}^\mathsf{T}(y - \bar{y}) \le 0$ for all $y \in Y$. In particular $\tilde{y}^\mathsf{T}(y - \bar{y}) \le 0$ for all $y = Ax \in AX \subseteq Y$. Therefore $\tilde{y}^\mathsf{T}A(x - \bar{x}) \le 0$ for all $x \in X$, i.e., $A^\mathsf{T}\tilde{y} \in N_X(\bar{x})$. This shows that $A^\mathsf{T}N_Y(\bar{y}) \subseteq N_X(\bar{x})$.

Conversely suppose $\tilde{x} \in N_X(\bar{x})$, i.e. $\tilde{x}^\mathsf{T}(x - \bar{x}) \le 0$ for all $x \in X$. Use (12.12b) to write $x - \bar{x} = A^\dagger(y - \bar{y}) + W_{n-r}(\beta - \bar{\beta})$ where $\bar{y} = A\bar{x} + W_{n-r}\bar{\beta}$ is fixed. Then

$$\tilde{x}^\mathsf{T}A^\dagger(y - \bar{y}) + \tilde{x}^\mathsf{T}W_{n-r}(\beta - \bar{\beta}) \le 0, \quad \forall y \in Y,\ \forall \beta \in \mathbb{R}^{n-r}$$

Since this holds for all $y \in Y$ and all $\beta \in \mathbb{R}^{n-r}$, it can be satisfied if and only if (setting $y = \bar{y}$ and then $\beta = \bar{\beta}$)

$$\tilde{x}^\mathsf{T}A^\dagger(y - \bar{y}) \;\le\; 0 \qquad \forall y \in Y \tag{12.13a}$$

$$\tilde{x}^\mathsf{T}W_{n-r}(\beta - \bar{\beta}) \;\le\; 0 \qquad \forall \beta \in \mathbb{R}^{n-r} \tag{12.13b}$$

The second inequality (12.13b) implies $\tilde{x}^\mathsf{T}W_{n-r} = 0$ (take $\beta = \bar{\beta} \pm e_j$) and hence $\tilde{x} \in$ range(W_r) according to Theorem A.19. The first inequality (12.13a) implies $(A^\dagger)^\mathsf{T}\tilde{x} \in N_Y(\bar{y})$, i.e., there exists $\tilde{y} \in N_Y(\bar{y})$ such that

$$(A^\dagger)^\mathsf{T}\tilde{x} \;=\; \tilde{y}$$

Multiplying both sides by A^T and using $A^\dagger A = W_r W_r^\mathsf{T}$ is symmetric we have

$$(A^\dagger A)^\mathsf{T}\tilde{x} \;=\; (A^\dagger A)\tilde{x} \;=\; \tilde{x} \;=\; A^\mathsf{T}\tilde{y}$$

where the second equality follows because, from Theorem A.19, $A^\dagger A = W_r W_r^\mathsf{H}$ projects $\tilde{x}$ onto range(W_r) but $\tilde{x}$ is already in range(W_r). This shows that $N_X(\bar{x}) \subseteq A^\mathsf{T} N_Y(\bar{y})$. This completes the proof of $N_X(\bar{x}) = A^\mathsf{T} N_Y(\bar{y})$. □

The same argument shows $X^\circ = A^\mathsf{T} Y^\circ$ (whether or not $0 \in X$).

Affine transformation.

We now generalize Theorem 12.6 and 12.7 to an affine transformation $f(x) = Ax + b$ where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. Given a nonempty set $X \subseteq \mathbb{R}^n$ let the image of X under the affine transformation be

$$Y_b \;:=\; AX + b \;\subseteq\; \mathbb{R}^m$$

i.e., $y \in Y_b$ if and only if $y = Ax + b$ for some $x \in X$. The next result shows that the normal cone of Y_b is independent of the translation by b (except for the relation $\bar{y} = A\bar{x} + b$). It reduces to Theorem 12.6 when $b = 0$.

Corollary 12.8 (Image of affine transformation)**.** Let $X \subseteq \mathbb{R}^n$ be a nonempty set and $Y_b := AX + b$ where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. Let $\bar{y}_b = A\bar{x} + b \in Y$.

1 The polar cone $(Y_b - b)^\circ$ of $Y_b - b = AX$ is the pre-image of X° under A^T:

$$(Y_b - b)^\circ \;=\; \{y \in \mathbb{R}^m : A^\mathsf{T} y \in X^\circ\}$$

Hence $A^\mathsf{T}(Y_b - b)^\circ \subseteq X^\circ$. If rank$(A) = n$ then $A^\mathsf{T}(Y_b - b)^\circ = X^\circ$

2 The normal cone $N_{Y_b}(\bar{y}_b)$ is independent of b and is the pre-image of $N_X(\bar{x})$ under A^T:

$$N_{Y_b}(\bar{y}_b) \;=\; N_{AX}(A\bar{x}) \;=\; \{y \in \mathbb{R}^m : A^\mathsf{T} y \in N_X(\bar{x})\}$$

Hence $A^\mathsf{T} N_{Y_b}(\bar{y}_b) \subseteq N_X(\bar{x})$. If rank$(A) = n$ then $A^\mathsf{T} N_{Y_b}(\bar{y}_b) = N_X(\bar{x})$.

Proof Since $\hat{Y} := Y_b - b := AX$, Theorem 12.6 implies that $A^\mathsf{T}\hat{Y}^\circ = A^\mathsf{T}(Y_b - b)^\circ \subseteq X^\circ$, and $A^\mathsf{T}(Y - b)^\circ = X^\circ$ if rank$(A) = n$.

For part 2, $\tilde{y} \in N_{Y_b}(\bar{y}_b)$ if and only if $\tilde{y}^\mathsf{T}(y - \bar{y}_b) \le 0$ for $y = Ax + b$ for all $x \in X$, i.e.,

$$\tilde{y} \in N_{Y_b}(\bar{y}_b) \qquad \Longleftrightarrow \qquad \tilde{y}^\mathsf{T} A(x - \bar{x}) \le 0 \quad \forall x \in X$$

This implies $N_{Y_b}(\bar{y}_b)$ is independent of b, and therefore $N_{Y_b}(\bar{y}_b) = N_Y(A\bar{x})$ at $b := 0$ with $Y := AX$. Theorem 12.6 then implies that $A^\mathsf{T} N_{Y_b}(\bar{y}_b) \subseteq N_X(\bar{x})$ in general and $A^\mathsf{T} N_{Y_b}(\bar{y}_b) = N_X(\bar{x})$ when rank$(A) = n$. □

Suppose X in Corollary 12.8 is a convex cone. If $0 \in X$, then the polar cone Y_b° is the intersection of the pre-image of X° under A^T and a halfspace:

$$Y_b^\circ \;=\; \{y \in \mathbb{R}^m : A^\mathsf{T} y \in X^\circ,\; y^\mathsf{T} b \le 0\} \tag{12.14}$$

To show this, we have $Y_b^\circ = \{y \in \mathbb{R}^m : y^\mathsf{T}(Ax+b) \le 0 \;\forall x \in X\}$. Since $0 \in X$, $b \in Y$ and $y \in Y_b^\circ$ implies $y^\mathsf{T} b \le 0$. Therefore

$$Y_b^\circ \;=\; \left\{y \in \mathbb{R}^m : \left(A^\mathsf{T} y\right)^\mathsf{T} x + y^\mathsf{T} b \le 0 \;\forall x \in X\right\} \cap H_-(b)$$

where $H_-(b) := \{y \in \mathbb{R}^m : y^\mathsf{T} b \le 0\}$ is a halfspace. We now show that $\left(A^\mathsf{T} y\right)^\mathsf{T} x + y^\mathsf{T} b \le 0$ for all $x \in X$ implies that $\left(A^\mathsf{T} y\right)^\mathsf{T} x \le 0$ for all $x \in X$. Suppose for the sake of contradiction that there exists $\bar{y} \in Y_b^\circ$ and $\bar{x} \in X$ such that $(A^\mathsf{T}\bar{y})^\mathsf{T}\bar{x} > 0$. Since $\gamma\bar{x} \in X$ for any $\gamma > 0$ we have $\lim_{\gamma\to\infty}(A^\mathsf{T}\bar{y})^\mathsf{T}(\gamma\bar{x}) \to \infty$, contradicting $\left(A^\mathsf{T}\bar{y}\right)^\mathsf{T}(\gamma\bar{x}) + \bar{y}^\mathsf{T} b \le 0$. Hence, for any $y \in Y^\circ$, $\left(A^\mathsf{T} y\right)^\mathsf{T} x \le 0$ for all $x \in X$, i.e., $A^\mathsf{T} y \in X^\circ$, as desired.

Theorem 12.8 is illustrated in the next example.

Example 12.6 (Image of affine transformation)**.** Consider the convex cone X and its affine transformation Y_b:

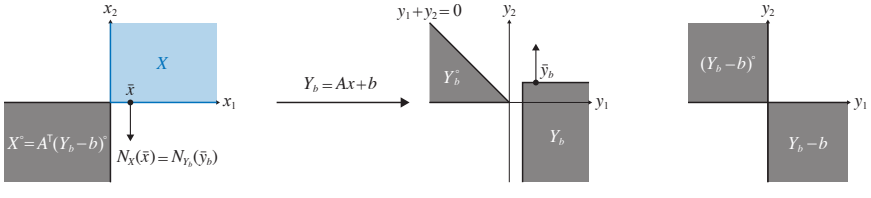
$$X \;:=\; \{x \in \mathbb{R}^2 : x \ge 0\}, \qquad A \;:=\; \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \qquad b \;:=\; \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$Y_b \;:=\; AX + b \;=\; \{y \in \mathbb{R}^2 : y_1 \ge 1,\; y_2 \le 1\}$$

The polar cone $X^\circ = \{x \in \mathbb{R}^2 : x \le 0\}$. Since $0 \in X$, (12.14) implies that the polar cone of Y_b is

$$Y_b^\circ \;=\; \{y \in \mathbb{R}^2 : A^\mathsf{T} y \in X^\circ,\; y^\mathsf{T} b \le 0\} \;=\; \{y \in \mathbb{R}^2 : y_1 \le 0,\; y_2 \ge 0,\; y_1 + y_2 \le 0\}$$

This is illustrated in Figure 12.9. It can be seen that Y_b is not a cone (since $b \neq 0$)



(a) Image of affine transformation (b) $Y_b - b$

Figure 12.9 Example 12.6: while X is a cone, Y_b is not. $X^\circ = A^\mathsf{T}(Y_b - b)^\circ$ and $N_X(\bar{x}) = A^\mathsf{T} N_{Y_b}(\bar{y}_b)$ because A is nonsingular (Theorem 12.8).

but Y_b° is a closed convex cone. Moreover $Y_b - b$ shifts the origin to b and is a convex cone with $(Y_b - b)^\circ = \{y \in \mathbb{R}^2 : A^\mathsf{T} y \in X^\circ\} \;= \{y \in \mathbb{R}^2 : y_1 \le 0,\; y_2 \ge 0\}$. Since A is nonsingular, it can be verified that $A^\mathsf{T}(Y_b - b)^\circ = X^\circ$.

At $\bar{x} = (1,0)$ and $\bar{y}_b = A\bar{x} + b = (2,1)$, the normal cone of the convex cone X is, from Theorem 12.3,

$$N_X(\bar{x}) \;=\; \{x \in X^\circ : x^\mathsf{T}\bar{x} = 0\} \;=\; \{x \in \mathbb{R}^2 : x_1 = 0,\; x_2 \le 0\}$$

The normal cone of Y_b is its pre-image, from Theorem 12.8,

$$N_{Y_b}(\bar{y}_b) \;=\; \{y : A^\mathsf{T} y \in N_X(\bar{x})\} = \{y : y_1 = 0,\; y_2 \ge 0\}$$

At $\bar{x} = 0$ and $\bar{y}_b = A\bar{x} + b = (1,1)$, $N_X(\bar{x}) = X^\circ$ and $N_{Y_b}(\bar{y}_b) = \{y : y_1 \le 0,\; y_2 \ge 0\}$. Since A is nonsingular, it can be verified that $A^\mathsf{T} N_{Y_b}(\bar{y}_b) = N_X(\bar{x})$ in both cases. □

Given a nonempty set $Y \subseteq \mathbb{R}^m$ let its pre-image under an affine map be

$$X_b \;:=\; \{x \in \mathbb{R}^n : Ax + b \in Y\}$$

where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. We will use this affine transformation in Chapter 12.1.4 to study the normal cone of the convex set defined by a second-order cone constraint where Y is a convex cone. Similar to Corollary 12.8, the following relation follows from Theorem 12.7 and is used to derive the normal cone of a SOC constraint (12.19) from the normal cone of a standard second-order cone.

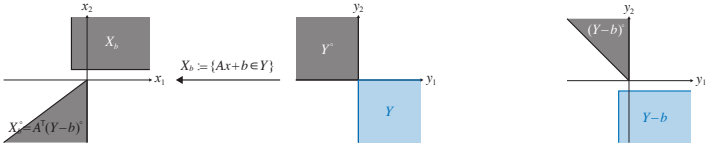
Corollary 12.9 (Pre-image of affine transformation)**.** Let $Y \subseteq \mathbb{R}^m$ be a nonempty set and $X_b := \{x \in \mathbb{R}^n : Ax + b \in Y\}$ be its pre-image under an affine transformation. Suppose $\bar{x} \in X_b$ and $\bar{y}_b = A\bar{x} + b \in Y$. Then $X_b^\circ = A^\mathsf{T}(Y - b)^\circ$ and $N_{X_b}(\bar{x}) = A^\mathsf{T} N_Y(\bar{y}_b)$.

Theorem 12.9 is verified in the next example (compared with Example 12.6).

Example 12.7 (Pre-image of affine transformation)**.** Consider the convex cone Y and its pre-image X_b under an affine transformation:

$$\begin{aligned} Y \;&:= \{y \in \mathbb{R}^2 : y_1 \ge 0,\; y_2 \le 0\} \\ X_b \;&:=\; \{x \in \mathbb{R}^2 : Ax + b \in Y\} \;=\; \{x \in \mathbb{R}^2 : x_1 \ge -1,\; x_2 \ge 1\} \end{aligned}$$

where A, b are the same as those in Example 12.6; see Figure 12.10. By definition,



(a) Pre-image of affine transformation (b) $Y - b$

Figure 12.10 Example 12.7: while Y is a cone, X_b is not. $X_b^\circ = A^\mathsf{T}(Y-b)^\circ$ and $N_{X_b}(\bar{x}) = A^\mathsf{T} N_Y(\bar{y}_b)$ (Theorem 12.9).

$y \in (Y - b)^\circ$ if and only if $y^\mathsf{T}\tilde{y} = \tilde{y}_1 y_1 + \tilde{y}_2 y_2 \le 0$ for all $\tilde{y} \in Y - b$, i.e., for all $\tilde{y}$ with

$\tilde{y}_1 \geq -1, \tilde{y}_2 \leq -1$. It can then be checked that $(Y-b)^\circ$ is (consider $\tilde{y} := (-1,-1) \in Y-b$, $\tilde{y}_1 \to \infty$ and $\tilde{y}_2 \to -\infty$)

$$(Y-b)^\circ \;=\; \{y \in \mathbb{R}^2 : y_1 + y_2 \geq 0,\; y_1 \leq 0,\; y_2 \geq 0\}$$

which is a closed convex cone even though $Y-b$ is not a cone (Proposition 12.2). Theorem 12.9 implies that $X_b^\circ = A^\mathsf{T}(Y-b)^\circ$, which we verify directly as follows. For $y \in (Y-b)^\circ$, $x := A^\mathsf{T} y = (y_1, -y_2)$ and hence

$$A^\mathsf{T}(Y-b)^\circ \;=\; \{x \in \mathbb{R}^2 : x_1 - x_2 \geq 0,\; x_1 \leq 0,\; x_2 \leq 0\}$$

On the other hand, $x \in X_b^\circ$ if and only if $x^\mathsf{T}\tilde{x} = \tilde{x}_1 x_1 + \tilde{x}_2 x_2 \leq 0$ for all $\tilde{x} \in X_b$, i.e., for all $\tilde{x}$ with $\tilde{x}_1 \geq -1$, $\tilde{x}_2 \geq 1$. It can then be checked that X_b° is (consider $\tilde{x} := (-1,1) \in X_b$, $\tilde{x}_1 \to \infty$ and $\tilde{x}_2 \to \infty$)

$$X_b^\circ \;=\; \{x \in \mathbb{R}^2 : x_1 - x_2 \geq 0,\; x_1 \leq 0,\; x_2 \leq 0\}$$

which equals $A^\mathsf{T}(Y-b)^\circ$; see Figure 12.10.

At $\bar{y}_b = (1,0)$ and $\bar{x} = A^{-1}(\bar{y}_b - b) = (0,1)$. Theorem 12.9 implies $N_{X_b}(\bar{x}) = A^\mathsf{T} N_Y(\bar{y}_b)$, which can be verified as follows. Since Y is a convex cone we can apply Theorem 12.3 to obtain $N_Y(\bar{y}_b) = \{y \in Y^\circ : y^\mathsf{T} y_b = 0\} = \{y \in \mathbb{R}^2 : y_1 = 0,\, y_2 \geq 0\}$. Hence $A^\mathsf{T} N_Y(\bar{y}_b) = \{x \in \mathbb{R}^2 : x_1 = 0,\, x_2 \leq 0\}$. Since X_b is not a cone we cannot apply Theorem 12.3 to obtain $N_{X_b}(\bar{x})$. By definition $x \in N_{X_b}(\bar{x})$ if and only if $x^\mathsf{T}(\tilde{x} - \bar{x}) \leq 0$ for all $\tilde{x} \in X_b$, i.e.,

$$\tilde{x}_1 x_1 + (\tilde{x}_2 - 1) x_2 \;\leq\; 0 \qquad \text{for all } \tilde{x} \text{ with } \tilde{x}_1 \geq -1, \tilde{x}_2 \geq 1$$

Taking $\tilde{x} = (-1,1)$ and $\tilde{x} = (1,1)$ yields $x_1 = 0$. Hence $x_2 \leq 0$. This shows that $N_{X_b}(\bar{x}) = A^\mathsf{T} N_Y(\bar{y}_b)$, verifying Theorem 12.9. □

12.1.4 Second-order cones and SOC constraints

Second-order cones.

The normal cone $N_K(\bar{x}, \bar{s})$ of the second-order cone K_{soc} defined in (8.16) can be derived explicitly. It is the polar cone K_{soc}° at the origin, the origin at an interior point, and, at a boundary point, the line segment in the intersection of the "lower cone" K_{soc}° and the hyperplane with normal $(\bar{x}/\|\bar{x}\|_2, 1)$.

Theorem 12.10 (Second-order cone)**.** Let $K_{\text{soc}} := \{(x,s) \in \mathbb{R}^{n+1} : \|x\|_2 \leq s\}$ be the standard second-order cone. Then

1 K_{soc} is a closed convex cone.
2 Its polar cone is $K_{\text{soc}}^\circ = \{(y,t) \in \mathbb{R}^{n+1} : \|y\|_2 \leq -t\}$.

3 Its normal cone $N_K(\bar{x},\bar{s})$ at an $(\bar{x},\bar{s}) \in K_{\text{soc}}$ is

$$N_K(\bar{x},\bar{s}) \;=\; \begin{cases} K_{\text{soc}}^\circ & \text{if } (\bar{x},\bar{s}) = (0,0) \\ \{(0,0) \in \mathbb{R}^{n+1}\} & \text{if } \|\bar{x}\|_2 < \bar{s} \\ \{\mu(\bar{x},-\bar{s}) \in \mathbb{R}^{n+1} : \mu \geq 0\} & \text{if } \|\bar{x}\|_2 = \bar{s} > 0 \end{cases}$$

Proof Part 1 is left as Exercise 8.11. To verify that $K_{\text{soc}}^\circ = \{(y,t) \in \mathbb{R}^{n+1} : \|y\|_2 \leq -t\}$, take any $(x,s) \in K_{\text{soc}}$ and (y,t) such that $\|y\|_2 \leq -t$. Then

$$x^\mathsf{T} y + st \;\leq\; \|x\|_2 \|y\|_2 + st \;\leq\; s(-t) + st \;=\; 0 \tag{12.15}$$

where the first inequality follows from the Cauchy-Schwarz inequality and the second inequality follows from definition of K_{soc}. Hence $(y,t) \in K_{\text{soc}}^\circ$. This shows that $K_{\text{soc}}^\circ \supseteq \{(y,t) \in \mathbb{R}^{n+1} : \|y\|_2 \leq -t\}$. Conversely let $(y,t) \in K_{\text{soc}}^\circ$, i.e., $x^\mathsf{T} y + st \leq 0$ for all $\|x\|_2 \leq s$. Clearly $(0,0) \in K_{\text{soc}}^\circ$ since K_{soc}° is a closed convex cone, so let $s \geq \|x\|_2 > 0$. Then $x^\mathsf{T} y + \|x\|_2 t \leq 0$ and hence

$$\frac{x^\mathsf{T}}{\|x\|_2} y + t \;\leq\; 0$$

Since this holds for all x (because there always exists some $s > 0$ such that $(x,s) \in K_{\text{soc}}$), we can take $x = y$ to conclude $\|y\|_2 + t \leq 0$. This proves part 2. Indeed K_{soc} is the "upper" cone in Figure 8.8(b) and K_{soc}° is the "lower" cone.

For part 3, application of Theorem 12.3 to part 2 yields

$$N_K((\bar{x},\bar{s})) \;=\; \{(y,t) \in \mathbb{R}^{n+1} : \|y\|_2 \leq -t,\; \bar{x}^\mathsf{T} y + \bar{s} t = 0\} \tag{12.16}$$

Hence if $(\bar{x},\bar{s}) = (0,0)$ then $N_K((\bar{x},\bar{s})) = K_{\text{soc}}^\circ$. If $\|\bar{x}\|_2 < \bar{s}$ then $(\bar{x},\bar{s})$ is in the interior of K_{soc} and hence $N_K(\bar{x},\bar{s}) = \{(0,0) \in \mathbb{R}^{n+1}\}$. Consider then $\|\bar{x}\|_2 = \bar{s} \neq 0$. The requirement that $\bar{x}^\mathsf{T} y + \bar{s} t = 0$ means that the two inequalities in (12.15) must hold with equality which is possible if and only if

$$y \;=\; \mu\bar{x} \quad \text{for any } \mu \in \mathbb{R}_+, \qquad\qquad \|x\|_2 \;=\; \bar{s}, \qquad\qquad \|y\|_2 \;=\; -t$$

Hence $-t = \|y\|_2 = \mu\|\bar{x}\|_2 = \mu\bar{s}$. This proves $(y,t) = \mu(\bar{x},-\bar{s})$. This is illustrated in Figure 12.11. □

We know from Theorem 12.3 that the normal cone $N_K(\bar{x},\bar{s})$ of a convex cone K are vectors in its polar cone K° where complementary slackness holds. Theorem 12.10 describes these vectors in more detail when K is explicitly specified as the second-order cone (note that the vector $\mu(\bar{x},-\bar{s}) \in K^\circ$).

Recall the relation $K_{\text{soc}} = AK_{\text{rsoc}}$ between a rotated second-order cone K_{rsoc} defined in (8.17) and a standard second-order cone K_{soc}, where A is a nonsingular matrix defined in (8.18), reproduced here:

$$A \;=\; \begin{bmatrix} 2\mathbb{I}_n & 0_n & 0_n \\ 0_n^\mathsf{T} & 1 & -1 \\ 0_n^\mathsf{T} & 1 & 1 \end{bmatrix} \tag{12.17}$$

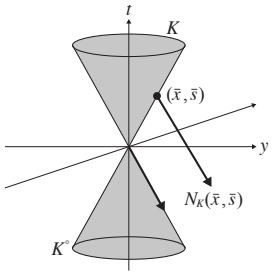


Figure 12.11 Theorem 12.10: The normal cone $N_K((\bar{x},\bar{s}))$ is the line segment on the boundary of the lower cone $K^\circ_{\rm soc}$ in the direction of $\bar{x}$. (April 19, 2025: Change $K \to K_{\rm soc}$ and $K^\circ \to K^\circ_{\rm soc}$.)

For an $x \in \mathbb{R}^n$, we use x^m, $m \le n$, to denote the subvector $x^m := (x_1, \ldots, x_m)$ of the first m entries of x. Since A is nonsingular, the application of Theorem 12.6 to Theorem 12.10 leads to the following result on rotated second-order cone.

Corollary 12.11 (Rotated second-order cone)**.** Let $K_{\rm rsoc} := \{x \in \mathbb{R}^{n+2} : \|x^n\|_2^2 \le x_{n+1}x_{n+2}, x_{n+1} \ge 0, x_{n+2} \ge 0\}$ be a rotated second-order cone. Let $K_{\rm soc} := AK_{\rm rsoc}$ where A is defined in (12.17) and $K^\circ_{\rm soc}$ denote its polar cone.

1 $K_{\rm rsoc}$ is a closed convex cone.
2 Its polar cone is

$$K^\circ_{\rm rsoc} \;=\; A^{\mathsf T}K^\circ_{\rm soc} \;=\; \{A^{\mathsf T}x \in \mathbb{R}^{n+2} : \|x^{n+1}\|_2 \le -x_{n+2}\}$$

3 Its normal cone $N_{K_r}(\bar{x}) = A^{\mathsf T}N_K(A\bar{x})$ at an $\bar{x} \in K_{\rm rsoc}$ is

$$N_{K_r}(\bar{x}) \;=\; \begin{cases} A^{\mathsf T}K^\circ_{\rm soc} & \text{if } A\bar{x} = 0 \\ \{(0,0) \in \mathbb{R}^{n+2}\} & \text{if } \|[A\bar{x}]^{n+1}\|_2 < [A\bar{x}]_{n+2} \\ \{\mu([A\bar{x}]^{n+1}, -[A\bar{x}]_{n+2}) \in \mathbb{R}^{n+2} : \mu \ge 0\} & \text{if } \|[A\bar{x}]^{n+1}\|_2 = [A\bar{x}]_{n+2} > 0 \end{cases}$$

SOC constraint.

Consider the convex set C defined by second-order cone constraint in (8.19), reproduced here:

$$C \;:=\; \{x \in \mathbb{R}^n : (Ax+b, c^{\mathsf T}x+d) \in K_{\rm soc}\} \;=\; \{x \in \mathbb{R}^n : \|Ax+b\|_2 \le c^{\mathsf T}x+d\} \quad (12.18)$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$, $d \in \mathbb{R}$, and $K_{\rm soc}$ is the standard second-order cone defined in (8.16). Then C is the pre-image of $K_{\rm soc}$ under the affine transformation

$$C \;=\; \{x \in \mathbb{R}^n : \tilde{A}x + \tilde{b} \in K_{\rm soc}\} \qquad \text{where} \qquad \tilde{A} := \begin{bmatrix} A \\ c^{\mathsf T} \end{bmatrix}, \quad \tilde{b} := \begin{bmatrix} b \\ d \end{bmatrix}$$

The convex set C reduces to the standard second-order cone $K_{\rm soc}$ if $A = \begin{bmatrix} \mathbb{I}_{n-1} & 0 \\ 0 & 0 \end{bmatrix}$, $c = e_n$, $b = 0$, $d = 0$. It may not be a cone, e.g., $C = \{x : \|b\|_2 \le c^{\mathsf T}x + d)\}$ is a halfspace

if $A = 0$. The mapping $f : C \to K_{\rm soc}$ defined by $f(x) = \tilde{A}x + \tilde{b}$ is generally neither surjective nor injective. For instance if $\tilde{A}$ is singular then f is not injective; if $b \neq 0$ then $f(x) = \tilde{A}x + \tilde{b} \neq 0$ for any x and hence f is not surjective. Theorem 12.9 allows us to derive the normal cone of C from that of the standard second-order cone $K_{\rm soc}$: for any $\bar{x} \in C$ and $\bar{y} = \tilde{A}\bar{x} + \tilde{b}$,

$$N_C(\bar{x}) \;=\; \tilde{A}^{\mathsf{T}} N_K(\bar{y}) \tag{12.19}$$

where $N_K(\bar{y})$ is given by Theorem 12.10.

Example 12.8. Consider the case where $A = 0 \in \mathbb{R}^{m\times n}$ and $C := \{x \in \mathbb{R}^n : \|b\|_2 \leq c^{\mathsf{T}}x + d\}$ is a halfspace. We know from Theorem 12.3 that its normal cone is, for any $\bar{x}$ with $-c^{\mathsf{T}}\bar{x} \leq d - \|b\|_2$,

$$N_C(\bar{x}) \;=\; \left\{-\lambda c : \lambda \in \mathbb{R} \text{ such that } \lambda \geq 0 \text{ with } \lambda = 0 \text{ if } -c^{\mathsf{T}}\bar{x} < d - \|b\|_2\right\} \tag{12.20}$$

Theorem 12.9 shows that $N_C(\bar{x}) = \tilde{A}^{\mathsf{T}} N_K(\tilde{A}\bar{x} + \tilde{b})$ where

$$\tilde{A} \;:=\; \begin{bmatrix} 0 \\ c^{\mathsf{T}} \end{bmatrix}, \qquad \tilde{b} := \begin{bmatrix} b \\ d \end{bmatrix}$$

and $N_K(\bar{y}) \subseteq \mathbb{R}^{m+1}$ is given by Theorem 12.10 as, writing $y \in \mathbb{R}^{m+1}$ as $y =: (y^m, y_{m+1})$ with $y^m \in \mathbb{R}^m$,

$$N_K(\tilde{A}\bar{x} + \tilde{b}) \;=\; \begin{cases} K_{\rm soc}^\circ & \text{if } (b, c^{\mathsf{T}}\bar{x} + d) = (0,0) \\ \{(0,0)\} & \text{if } \|b\|_2 < c^{\mathsf{T}}\bar{x} + d \\ \left\{\mu\left(b, -(c^{\mathsf{T}}\bar{x} + d)\right) \in \mathbb{R}^{m+1} : \mu \geq 0\right\} & \text{if } \|b\|_2 = c^{\mathsf{T}}\bar{x} + d > 0 \end{cases}$$

and $K_{\rm soc}^\circ = \{y \in \mathbb{R}^{m+1} : \|y^m\|_2 \leq -y_{m+1}\}$. (If $b \neq 0$ then $N_K(\tilde{A}\bar{x} + \tilde{b}) \neq K_{\rm soc}^\circ$ for any $\bar{x}$.)

We now verify that $N_C(\bar{x}) = \tilde{A}^{\mathsf{T}} N_K(\tilde{A}\bar{x} + \tilde{b})$. Indeed $\tilde{A}^{\mathsf{T}} N_K(\tilde{A}\bar{x} + \tilde{b})$ is, noting that $y_{m+1} \leq 0$,

$$\tilde{A}^{\mathsf{T}} N_K(\tilde{A}\bar{x} + \tilde{b}) \;=\; \begin{cases} \{y_{m+1}c : y_{m+1} \in \mathbb{R}_-\} & \text{if } (b, c^{\mathsf{T}}\bar{x} + d) = (0,0) \\ \{-\mu\|b\|_2 c : \mu \in \mathbb{R}_+\} & \text{if } \|b\|_2 = c^{\mathsf{T}}\bar{x} + d > 0 \\ \{0 \in \mathbb{R}^{n+1}\} & \text{if } \|b\|_2 < c^{\mathsf{T}}\bar{x} + d \end{cases}$$

which is equal to $N_C(\bar{x})$ in (12.20), as desired. □

12.2 CPC functions

When we allow extended real-valued and discontinuous functions we can treat constrained optimization as unconstrained optimization and develop a unified theory that covers both. In this section we define an important class of such functions, the set of closed proper convex (CPC) functions, that we will use extensively in deriving optimality conditions in later sections.

12.2.1 Extended real-valued function

A *real-valued* function $f : \mathbb{R}^n \to \mathbb{R}$ maps a finite vector $x \in \mathbb{R}^n$ to a finite value $f(x) \in \mathbb{R}$. An *extended real-valued* function $f : \mathbb{R}^n \to [-\infty, \infty]$ can take values $-\infty$ and ∞. For a function $f : X \to [-\infty, \infty]$ defined on $X \subseteq \mathbb{R}^n$, X is called the *domain* of f. The *effective domain* of f is the set $\mathrm{dom}(f) := \{x \in X : f(x) < \infty\}$. The *epigraph* of f is the set $\mathrm{epi}(f) := \{(x, y) \in X \times \mathbb{R} : y \geq f(x)\} \subseteq \mathbb{R}^{n+1}$. In particular if $(x, y) \in \mathrm{eip}(f)$ then $y \notin \{-\infty, \infty\}$ by definition. Therefore $x \in \mathrm{dom}(f)$ if and only if there exists $y = y(x) \in \mathbb{R}$ such that $(x, y) \in \mathrm{epi}(f)$, i.e., $\mathrm{dom}(f)$ is the projection of $\mathrm{epi}(f)$ onto $\mathbb{R}^n$.

For the purpose of minimization, a function $f : X \to [-\infty, \infty]$ defined on $X \subseteq \mathbb{R}^n$ can always be extended to $\mathbb{R}^n$ by defining

$$f_X(x) := \begin{cases} f(x) & \text{if } x \in X \\ \infty & \text{if } x \in \mathbb{R}^n \setminus X \end{cases} \tag{12.21}$$

The *epigraph* of f_X is then the set $\mathrm{epi}(f_X) := \{(x, y) \in \mathbb{R}^n \times \mathbb{R} : y \geq f_X(x)\} \subseteq \mathbb{R}^{n+1}$ (we reiterate that y is finite in $\mathrm{epi}(f_X)$ by definition). Therefore we often treat real-valued functions f on X as extended real-valued functions f_X on $\mathbb{R}^n$ whose effective domain $\mathrm{dom}(f_X)$ may be a subset of $\mathbb{R}^n$.

Consider an extended real-valued function $f : X \to [-\infty, \infty]$ where its domain $X \subseteq \mathbb{R}^n$. We say that f is *lower semicontinuous (lsc) at* $x \in X$ if

$$f(x) \leq \liminf_k f(x_k) \tag{12.22}$$

for every sequence $\{x_k\} \subseteq X$ with $x_k \to x$, and that f is *lower semicontinuous* (on X) if it is lsc at every $x \in X$. A function f is called *upper semicontinuous* (usc) if $-f$ is lsc. A function is continuous if and only if it is both lsc and usc.

Definition 12.3 (Closed proper convex (CPC) f)**.** Consider $f : X \to [-\infty, \infty]$ with $X \subseteq \mathbb{R}^n$.

1 The function f is *closed* if $\mathrm{epi}(f)$ is a closed set in $\mathbb{R}^{n+1}$.
2 The function f is *proper* if there exists $\bar{x} \in X$ such that $f(\bar{x}) < \infty$ (so that $\mathrm{epi}(f)$ is nonempty) and $f(x) > -\infty$ for all $x \in X$. In particular a real-valued function $f : X \to \mathbb{R}$ is proper.
3 Suppose X is convex. Then f is *convex* if $\mathrm{epi}(f)$ is a convex subset of $\mathbb{R}^{n+1}$. □

The convexity definition in terms of $\mathrm{epi}(f)$ implies that $\mathrm{dom}(f)$ is a convex set in $\mathbb{R}^n$. It reduces to the usual definition of convexity for real-valued functions. If a closed convex function f is not proper then f cannot take any finite value: $f(x) = -\infty$ if $x \in \mathrm{dom}(f)$ and $f(x) = \infty$ otherwise. We therefore consider only proper functions $f : X \to (-\infty, \infty]$. A proper and convex function is continuous, except possibly on its relative boundary. Moreover it is Lipschitz continuous over a compact set with the norm of a maximum subgradient as its Lipschitz constant; see Lemma 12.15.

A common mistake in the literature is to claim that if f is lsc, then $\mathrm{dom}(f)$ is a closed set or that f is a closed function.[2] The subtle relation between lsc, closed f (closed $\mathrm{epi}(f)$) and closed $\mathrm{dom}(f)$ is explained in the next remark.

Remark 12.3 (lsc, closed f, closed epi(f), closed $\mathrm{dom}(f)$)**.** Consider an extended real-valued function $f : X \to [-\infty, \infty]$ where $X \subseteq \mathbb{R}^n$ is the domain of f.

1 *lsc and X*. Whether f is lsc (or continuous) depends on its domain X. Take the indicator function $\delta_C(x) := 0$ if $x \in C \subseteq \mathbb{R}^n$ and ∞ if $x \notin C$. Suppose C is open in $\mathbb{R}^n$. If the domain X of the extended real-valued function is taken to be the closure $\mathrm{cl}(C)$ of C (or $\mathbb{R}^n$), then $\delta_C(x)$ is not lsc on $\mathrm{cl}(C)$ (or $\mathbb{R}^n$) because (12.22) is not satisfied at $x \in X$ on the boundary of C. If $X = C$, however, $\delta_C(x)$ is lsc on X because in the test (12.22) for lsc, x must be in the open set X.

2 *Continuity and X*. Consider the extended real-valued function $f(x) := 1/x$ defined on $X := [0, 1]$; in particular $f(0) := \infty$. Then f is lsc at $x = 0 \in X$ because $\liminf_k f(x_k)$ (and $\limsup_k f(x_k)$) can take $\pm\infty$ value by definition if the sequence $\{x_k\} \subseteq X$ is unbounded. In contrast, f is not continuous at $x = 0$ because continuity means that $f(x_k)$ converges to a *finite* value $y \in \mathbb{R}$ for every sequence $\{x_k\} \subseteq X$ with $\lim_k x_k = x \in X$ (x is also finite).[3] If the domain of f is taken to be $X' := (0, 1]$ instead, f is continuous on X' because the test sequence $\{x_k\}$ cannot converge to a boundary point not in X'.

3 *lsc and closedness of f: $X = \mathbb{R}^n$*. If $X = \mathbb{R}^n$, then f is lsc on $\mathbb{R}^n$ if and only if $\mathrm{epi}(f)$ is a closed set in $\mathbb{R}^{n+1}$ (f is closed). See [54, Propositions 1.1.2 and 1.1.3, p.10] (Exercise 12.8).

4 *lsc and closedness of f: $X \subsetneq \mathbb{R}^n$*. If $X \subsetneq \mathbb{R}^n$, however, lsc and closedness of f are not equivalent. If the effective domain $\mathrm{dom}(f) := \{x \in X : f(x) < \infty\}$ is closed in X and f is lsc on $\mathrm{dom}(f)$, then $\mathrm{epi}(f)$ is a closed set in $X \times \mathbb{R}$ (f is closed).

The converse may not hold. It is possible that f is lsc on $\mathrm{dom}(f)$ but $\mathrm{dom}(f)$ is not closed in X, and yet, f is closed. An example is the function $f(x) := 1/x$ on $X := [0, 1]$ defined above where f is lsc (in fact continuous) on $\mathrm{dom}(f) = (0, 1]$, but $\mathrm{dom}(f)$ is not closed in X (or in $\mathbb{R}$). To see that f is a closed function, consider any sequence $\{(x_k, y_k)\} \in \mathrm{epi}(f) \subseteq X \times \mathbb{R}$ such that $(x_k, y_k) \to (\bar{x}, \bar{y}) \in X \times \mathbb{R}$. By definition $\bar{y}$ is finite and therefore $\bar{x}$ cannot be 0 (i.e., $(\bar{x}, \bar{y}) \neq (0, \infty)$). Moreover $(\bar{x}, \bar{y}) \in \mathrm{epi}(f)$ because

$$f(\bar{x}) \;\leq\; \liminf_k f(x_k) \;\leq\; \liminf_k y_k \;=\; \bar{y}$$

where the first inequality follows from lsc of f on $\mathrm{dom}(f)$, the second inequality follows because $(x_k, y_k) \in \mathrm{epi}(f)$, and the equality follows because $y_k \to \bar{y}$. In general, the closedness of $\mathrm{dom}(f)$ ensures that for any sequence $\{(x_k, y_k)\} \in$

[2] Such a claim has been made on the recourse function $Q(x)$ in two-stage optimization with recourse where $\mathrm{dom}(Q)$ is claimed to be a closed (convex) set in [142, Proposition 2.7, p.35] and [143, Corollary 37; p.158]. See Lemma 13.29 for a correct statement.

[3] In general when we say a sequence $\{x_k\} \subset \mathbb{R}^n$ *converges* to an x, we mean that the limit point x is in $\mathbb{R}^n$, i.e., x is finite. If $\|x_k\| \to \pm\infty$, the sequence is said to be *unbounded*.

epi(f) with $(x_k, y_k) \to (\bar{x}, \bar{y}) \in X \times \mathbb{R}$, $\bar{y}$ is finite. Then the inequalities above hold generally to show the closedness of f.

Hence for an extended real-valued function f defined on $X \subseteq \mathbb{R}^n$, f can be a closed function, or equivalently epi(f) can be a closed set in $X \times \mathbb{R}$, while dom(f) is not closed in X (even when f is lsc on X). Often it is the closedness of f that is needed, not the closedness of dom(f), e.g., in the Weierstrass Theorem 12.22 and its application in Theorem 13.30 to derive conditions for primal optimality of two-stage nonlinear optimization with recourse. □

12.2.2 Indicator function, support function and polyhedral functions

Indicator function and support function.

Given a set $X \subseteq \mathbb{R}^n$ the *indicator function* of X is the extended real-valued function $\delta_X : \mathbb{R}^n \to (-\infty, \infty]$ defined by:

$$\delta_X(x) := \begin{cases} 0 & \text{if} \quad x \in X \\ \infty & \text{if} \quad x \notin X \end{cases} \tag{12.23a}$$

It is proper if and only if the set X is nonempty. It is a convex function if and only if X is a convex set.

The *support function* of X is $\sigma_X : \mathbb{R}^n \to (-\infty, \infty]$ defined by:

$$\sigma_X(x) := \sup_{y \in X} y^\mathsf{T} x \tag{12.23b}$$

It is proper if and only if X is nonempty and $\sup_{y \in X} y^\mathsf{T} x < \infty$ for at least one x. The sets X, cl(X), conv(X), cl(conv(X)), conv(cl(X)) all have the same support function (Exercise 12.9): for all $x \in \mathbb{R}^n$,

$$\sigma_X(x) = \sigma_{\mathrm{cl}(X)}(x) = \sigma_{\mathrm{conv}(X)}(x) = \sigma_{\mathrm{cl}(\mathrm{conv}(X))}(x) = \sigma_{\mathrm{conv}(\mathrm{cl}(X))}(x) \tag{12.24}$$

See Exercise 12.12 for relation between δ_X and σ_X (as well as their subdifferentials).

Theory of convexity, optimality and duality can be developed based either on real-valued functions or on extended real-valued functions. An advantage of extended real-valued functions is that they allow us to represent the minimization of a real-valued function $f : \mathbb{R}^n \to \mathbb{R}$ over X as an unconstrained optimization of the extended real-valued function (12.21):

$$\min_{x \in \mathbb{R}^n} \; f_X(x) = f(x) \; + \; \delta_X(x) \tag{12.25}$$

A unified theory can then be developed for unconstrained optimization as we will see in the following sections.

Example 12.9. Derive $\delta_X(x)$ and $\sigma_X(x)$ for:

1 $X := (0,1) \subseteq \mathbb{R}$.
2 $X := (-1,1) \subseteq \mathbb{R}$.
3 $X := \{x \in \mathbb{R}^n : x_i \in (-1,1)\}$.

Solution. For $X := (0,1)$ and $X := (-1,1)$

$$\delta_{(0,1)}(x) := \begin{cases} 0 & x \in (0,1) \\ \infty & x \notin (0,1) \end{cases} \qquad \sigma_{(0,1)}(x) := \sup_{y \in (0,1)} yx = \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases}$$

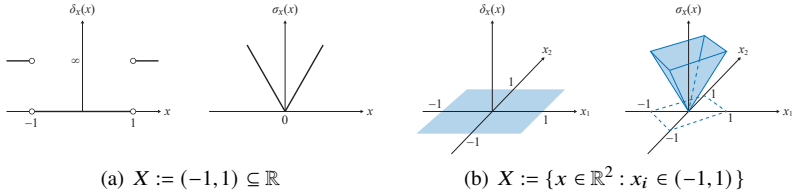
$$\delta_{(-1,1)}(x) := \begin{cases} 0 & x \in (-1,1) \\ \infty & x \notin (-1,1) \end{cases} \qquad \sigma_{(-1,1)}(x) := \sup_{y \in (-1,1)} yx = |x|$$

For $X := \{x \in \mathbb{R}^n : x_i \in (-1,1)\}$

$$\delta_X(x) := \begin{cases} 0 & x_i \in (-1,1) \text{ for all } i \\ \infty & x_i \notin (-1,1) \text{ for some } i \end{cases}$$

$$\sigma_X(x) := \sum_i \sup_{y_i \in (-1,1)} y_i x_i = \sum_i |x_i| = \|x\|_1$$

They are illustrated in Figure 12.12. □



(a) $X := (-1,1) \subseteq \mathbb{R}$ (b) $X := \{x \in \mathbb{R}^2 : x_i \in (-1,1)\}$

Figure 12.12 Example 12.9.

Polyhedral set and polyhedral function.

Recall that a polyhedral set, or a polyhedron, is a set $X := \{x \in \mathbb{R}^n : Ax \leq b\}$ specified by a finite number of affine inequalities. We often assume, sometimes implicitly, that X is nonempty to avoid triviality. Such a set is then nonempty closed and convex by definition. See Appendix A.2 for more discussions on polyhedral sets and extreme points.

We say that a proper function $f : \mathbb{R}^n \to (-\infty, \infty]$ is a *polyhedral function* if if its $\operatorname{epi}(f)$ is a polyhedral set in $\mathbb{R}^{n+1}$. Since a polyhedral set is closed nonempty convex, a polyhedral function is closed proper convex. It can be represented as the pointwise maximum of affine functions e.g. [54, Proposition 2.3.5, p.109].

Lemma 12.12. Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be a convex function. Then f is a polyhedral function if and only if $\operatorname{dom}(f)$ is a polyhedral set and

$$f(x) = \max_{i \in \{1,\ldots,m\}} \left(a_i^\mathsf{T} x + b_i \right), \qquad \forall x \in \operatorname{dom}(f)$$

for some $a_i \in \mathbb{R}^n$, $b_i \in \mathbb{R}$, and integer $m > 0$. □

In particular an affine function is polyhedral.

12.3 Gradient and subgradient

For smooth convex optimization the first-order stationarity condition takes the form $-\nabla f(x^*) = \nabla g(x^*)\lambda^* + \nabla h(x^*)\mu^*$ in terms of the gradients $\nabla f, \nabla g, \nabla h$ of the cost and constraint functions. In this section we generalize gradients of differentiable functions to subgradients of convex but possibly non-differentiable functions and develop conditions for subdifferential calculus. We use these tools in Chapter 12.5 to generalize the KKT Theorem 8.15 of Chapter 8.3.2 to the convex nonsmooth setting.

12.3.1 Derivative, directional derivative and partial derivative

The notion of derivative, directional derivative and partial derivative defined in Chapter 8.1.3 for real-valued functions extend directly to extended real-valued functions. Consider a proper function $f : X \to (-\infty, \infty]$ where $X \subseteq \mathbb{R}^n$ is an open set. The function f is said to be *differentiable at* $x \in X$ if there exists a vector $m \in \mathbb{R}^n$ such that

$$\lim_{\substack{h \in \mathbb{R}^n \\ h \to 0}} \frac{f(x+h) - f(x) - m^\mathsf{T} h}{\|h\|} = 0$$

When this holds, the column vector m is called the *gradient or derivative of* f *at* $x \in X$ and denoted by $\nabla f(x)$. If f is differentiable at every $x \in X$ then f is called *differentiable* on X.

At each $x \in X$ and for each $v \in \mathbb{R}^n$ the one-sided *directional derivative of* f *at* x *in the direction* v is defined as

$$df(x;v) := \lim_{\substack{t \in \mathbb{R} \\ t \downarrow 0}} \frac{f(x+tv) - f(x)}{t}$$

provided the limit exists, possibly $\pm\infty$. For $x \in \mathrm{dom}(f)$, $df(x;v)$ can take finite values or $\pm\infty$, but for $x \in \mathrm{ri}(\mathrm{dom}(f))$, $df(x;v)$ if exists is always finite for any $v \in \mathbb{R}^n$. It can be shown that f is differentiable at $x \in X$ if (i) directional derivatives $df(x;v)$ exist at x for all directions $v \in \mathbb{R}^n$, and (ii) $df(x;v)$ is a linear function of v.

At each $x \in X$ and for the unit vector $e_j \in \{0,1\}^n$, if the directional derivatives $df(x;e_j)$ and $df(x;-e_j)$ exist in both directions and are equal, then they are called the *partial derivative of* f *at* $x \in X$ with respect to x_j and denoted by $\frac{\partial f}{\partial x_j}(x)$:

$$\frac{\partial f}{\partial x_j}(x) := \lim_{\substack{t \in \mathbb{R} \\ t \to 0}} \frac{f(x+te_j) - f(x)}{t}$$

In this case f is called *partially differentiable at* $x \in X$ with respect to x_j. The row vector of partial derivatives of f at $x \in X$ is

$$\frac{\partial f}{\partial x}(x) := \left[\frac{\partial f}{\partial x_1}(x) \quad \cdots \quad \frac{\partial f}{\partial x_n}(x)\right]$$

If f is partially differentiable at all $x \in X$ then it is called *partially differentiable on* X. The partial derivative $\frac{\partial f}{\partial x}(x)$ describes the behavior of f at x only along the coordinate axes whereas the derivative $\nabla f(x)$ describes its behavior in all directions. If f is differentiable then it is partially differentiable, but the converse does not generally hold. If f is not only partially differentiable but $\frac{\partial f}{\partial x}(x)$ is also continuous at x, then the converse holds at $x \in X$. Such an f is called *continuously differentiable at* x. If f is continuously differentiable at all $x \in X$ then it is *continuously differentiable on* X.

As Example 8.3 in Chapter 8.1.3 shows, a partially differentiable function may not be differentiable when the partial derivative $\frac{\partial f}{\partial x}(x)$ is discontinuous at x. Indeed a partially differentiable function may not even be continuous at all $x \in X$. A continuously differentiable function is always continuous. Moreover Lemma 8.1 extends directly to a proper extended real-valued function $f : X \to (-\infty, \infty]$, i.e., if f is differentiable then it is partially differentiable and $\nabla f(x) = \left[\frac{\partial f}{\partial x}(x)\right]^\mathsf{T}$. Conversely, f is differentiable if it is continuously differentiable. Hence f is differentiable at $x \in X$ if and only if $df(x; v) = v^\mathsf{T} \nabla f(x) = \frac{\partial f}{\partial x}(x)\, v$ for all $v \in \mathbb{R}^n$. This is generalized in (12.28) below to proper convex functions that may not be differentiable (but are always subdifferentiable). Moreover the directional derivative of a proper convex function $f : X \to (-\infty, \infty]$ always exists because $(f(x+tv) - f(x))/t$ is increasing in $t > 0$ and hence the limit always exists, possibly $\pm\infty$. The limit $df(x; v)$ may be $-\infty$ or ∞ at the relative boundary of $\operatorname{dom}(f)$ but is always a finite value at an $x \in \operatorname{ri}(\operatorname{dom}(f))$.

12.3.2 Subgradient

Recall that, for the purpose of minimization, a function $f : X \to (-\infty, \infty]$ with $X \subseteq \mathbb{R}^n$ can always be represented as an extended real-valued function $f : \mathbb{R}^n \to (-\infty, \infty]$ by defining $f(x) := \infty$ for $x \notin X$ so that its effective domain $\operatorname{dom}(f) \subseteq X$.

Subgradient.

Consider a proper *convex* function $f : \mathbb{R}^n \to (-\infty, \infty]$. A vector $y \in \mathbb{R}^n$ is a *subgradient of* f at $\bar{x} \in \operatorname{dom}(f)$ if

$$f(x) \geq f(\bar{x}) + y^\mathsf{T}(x - \bar{x}), \qquad \forall x \in \mathbb{R}^n \tag{12.26a}$$

The inequality must hold for all real x, not just for $x \in \operatorname{dom}(f)$, i.e., the affine function on the right-hand side is a lower approximation of f over $\mathbb{R}^n$ and coincides with f at $x = \bar{x}$. The set of all subgradients of a convex function f at $\bar{x}$ is the *subdifferential*

$\partial f(\bar{x})$ of f at $\bar{x}$. By convention $\partial f(\bar{x}) = \emptyset$ if $\bar{x} \notin \mathrm{dom}(f)$. An equivalent definition to (12.26a) is: $y \in \mathbb{R}^n$ is a subgradient of f at $\bar{x} \in \mathrm{dom}(f)$ if

$$f(\bar{x}) - y^{\mathsf{T}}\bar{x} \;=\; \min_{x \in \mathbb{R}^n} \Big(f(x) - y^{\mathsf{T}} x \Big) \tag{12.26b}$$

i.e., $\bar{x} \in \mathrm{dom}(f)$ attains the minimum on the right-hand side.

The definition (12.26) of subgradient immediately implies the following first-order optimality condition for nonsmooth convex optimization. It is used in Chapter 12.5 to derive a general optimality condition which leads to various KKT conditions in subsequent subsections.

Corollary 12.13 (Optimality condition)**.** Consider the unconstrained convex optimization $\inf_{x \in \mathbb{R}^n} f(x)$ where $f : \mathbb{R}^n \to (-\infty, \infty]$ is a proper convex function. Then $x^* \in \mathbb{R}^n$ is optimal if and only if

$$0 \;\in\; \partial f(x^*)$$

If f is differentiable this reduces to $\nabla f(x^*) = 0$.

Proof It is obvious that $f(x^*) = \min_{x \in \mathbb{R}^n} f(x)$ if and only if $y = 0$ in (12.26b), i.e., if and only if $0 \in \partial f(x^*)$. □

Remark 12.4 (Subgradient as certificate of optimality)**.** 1 For unconstrained convex optimization, $0 \in \partial f(x^*)$ is necessary and sufficient for x^* to be an optimal. The fact that there may be subgradients $y \in \partial f(x^*)$ with $y^{\mathsf{T}}(x - x^*) \neq 0$ has no bearing on the optimality of x^*. The zero vector $0 \in \partial f(x^*)$ is a certificate for the optimality of x^*.

2 For constrained convex optimization, $x^* \in X$ is optimal if there exists a subgradient $y^* \in \partial f(x^*)$ such that $y^{*\mathsf{T}}(x - x^*) \geq 0$ for all feasible x (i.e., $-y \in N_X(x^*)$) because (12.26a) then implies $f(x) \geq f(x^*)$ for all feasible x. Such a subgradient y^* is a certificate for the optimality of x^*. A precise statement is Theorem 12.21 below. Again the fact that there may be subgradients $y \in \partial f(x^*)$ with $y^{\mathsf{T}}(x - x^*) < 0$ has no bearing on the optimality of x^*. □

A proper convex function is subdifferentiable at any interior point $\bar{x}$ of its effective domain. The supporting hyperplane of $\mathrm{epi}(f)$ at such a point $(\bar{x}, f(\bar{x}))$ is not vertical and this is the origin of the Slater condition in convex optimality (e.g. see Theorem 12.27 on strong duality and dual optimality).

Lemma 12.14 (Subdifferentiability of convex function at $x \in \mathrm{int}(\mathrm{dom}(f))$)**.** A proper convex function $f : \mathbb{R}^n \to (-\infty, \infty]$ always has a subgradient at any interior $x \in \mathrm{int}(\mathrm{dom}(f))$.

Proof The proof uses the Separating Hyperplane Theorem 8.10. Convexity of f means its epigraph $\mathrm{epi}(f) := \{(x, y) : y \geq f(x), x \in \mathbb{R}^n, y \in \mathbb{R}\}$ is a convex set in $\mathbb{R}^{n+1}$ (Definition 12.3). It is nonempty because f is proper. Fix an $\bar{x} \in \mathrm{int}(\mathrm{dom}(f))$. The

point $(\bar{x}, f(\bar{x}))$ is in $\text{epi}(f) \setminus \text{int}(\text{epi}(f))$. Theorem 8.10 then implies that there exists nonzero $(a,b) \in \mathbb{R}^n \times \mathbb{R}$ such that

$$a^\mathsf{T}(x-\bar{x}) + b(y - f(\bar{x})) \;\le\; 0, \qquad \forall (x,y) \in \text{epi}(f) \tag{12.27a}$$

This implies $b \le 0$ (substitute $(\bar{x}, y) \in \text{epi}(f)$ with $y > f(\bar{x})$ into (12.27a)). If $b = 0$ then $a^\mathsf{T}(x-\bar{x}) \le 0$ for all $x \in \text{dom}(f)$. Since $\bar{x} \in \text{int}(\text{dom}(f))$, we can take $x := \bar{x} \pm e_j$ to show that $a = 0$, contradicting $(a,b) \neq 0$. Hence $b < 0$, i.e., the supporting hyperplane of $\text{epi}(f)$ at $(\bar{x}, f(\bar{x}))$ is not vertical if $\bar{x} \in \text{int}(\text{dom}(f))$ is an interior point. We can therefore divide by b on both sides of (12.27a) to obtain (setting $y := f(x)$)

$$f(x) \;\ge\; f(\bar{x}) - \frac{a^\mathsf{T}}{b}(x-\bar{x}) \tag{12.27b}$$

Since this holds for all $x \in \mathbb{R}^n$, $-(a/b)$ is a subgradient of f at $\bar{x}$.[4] □

Remark 12.5 (Separating hyperplane argument)**.** The separating hyperplane argument that derives (12.3) relies on the fact that $Y(\bar{x})$ is a cone and hence $y \in Y(\bar{x}) \Rightarrow ty \in Y(\bar{x})$ for all $t > 0$ (see Remark 12.2). The same separating hyperplane argument that proves Lemma 12.14 relies on the fact that $(\bar{x}, f(\bar{x})) \in \text{epi}(f) \Rightarrow (\bar{x}, y) \in \text{epi}(f)$ for all $y \ge f(\bar{x})$. □

Lemma 12.14 establishes the existence of subgradient at an interior point. The next result, taken from [54, Propositions 5.4.1 and 5.4.2, pp. 184], presents additional properties. It generalizes Lemma 8.4 for real-valued convex functions to extended real-valued convex functions.

Lemma 12.15 (Subgradient and Lipschitz continuity)**.** Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be a proper convex function.

1 For $x \in \text{ri}(\text{dom}(f))$, $f(x)$ is continuous at x.
2 For $x \in \text{int}(\text{dom}(f))$, $\partial f(x)$ is a nonempty convex compact set.
3 If $X \subseteq \text{int}(\text{dom}(f))$ is nonempty and compact, then $\partial_X f := \cup_{x \in X} \partial f(x)$ is nonempty and bounded. Moreover f is Lipschitz continuous over X with Lipschitz constant $L := \sup_{\xi \in \partial_X f} \|\xi\|_2$.

If f is proper convex, even though it is continuous on $\text{ri}(\text{dom}(f))$, it is not necessarily lsc over $\mathbb{R}^n$ because $f(x)$ can be ∞ on the boundary of $\text{dom}(f)$. Hence convexity of f does not imply closedness. If f is a real-valued convex function, then $\partial f(x)$ is always a nonempty convex compact set. If f is extended real-valued convex, then $\partial f(x)$ can be unbounded or empty on the boundary of or outside $\text{dom}(f)$.

By the definition of subgradient we have, for all $t \in \mathbb{R}$, $f(x+tv) - f(x) \ge t\, y^\mathsf{T} v$ for all subgradients $y \in \partial f(x)$. Hence

$$df(x;v) \ge y^\mathsf{T} v, \qquad \forall y \in \partial f(x),\ x \in \text{dom}(f),\ v \in \mathbb{R}^n$$

[4] The assumption that $\bar{x} \in \text{int}(\text{dom}(f))$ is needed to show that $b \neq 0$. If $\bar{x} \in \text{ri}(\text{dom}(f))$, then the contradiction argument breaks down, but subgradient may still exist at such a $\bar{x}$. See Exercise 12.10.

For any $x \in \mathrm{ri}(\mathrm{dom}(f))$ the function $df(x;\cdot)$ is closed and is the support function of $\partial f(x)$, i.e.,

$$df(x;v) = \sup_{y \in \partial f(x)} y^\mathsf{T} v, \qquad \forall x \in \mathrm{ri}(\mathrm{dom}(f)),\ v \in \mathbb{R}^n \tag{12.28}$$

Hence $df(x;v) > \sup_{y\in\partial f(x)} y^\mathsf{T} v$ can only hold at a boundary point x of $\mathrm{dom}(f)$ where $df(x;\cdot)$ is not a closed function. In particular, if $f : \mathbb{R}^n \to \mathbb{R}$ is a real-valued function then $\mathrm{dom}(f) = \mathbb{R}^n$ and $df(x;v) = \sup_{y\in\partial f(x)} y^\mathsf{T} v$ for all $x, v \in \mathbb{R}^n$.[5]

Conjugate function.

Consider a convex function $f : \mathbb{R}^n \to (-\infty, \infty]$. Fix an $\bar{x} \in \mathrm{dom}(f)$. By definition (12.26), $\bar{y} \in \partial f(\bar{x})$ if and only if $f(x) \geq f(\bar{x}) + \bar{y}^\mathsf{T}(x - \bar{x})$ for all $x \in \mathbb{R}^n$ with equality at $x = \bar{x}$. Hence

$$\bar{y} \in \partial f(\bar{x}) \iff \bar{y}^\mathsf{T}\bar{x} - f(\bar{x}) = \sup_{x\in\mathbb{R}^n} \left(\bar{y}^\mathsf{T} x - f(x)\right) \tag{12.29a}$$

This motivates the definition of the *conjugate function* $f^* : \mathbb{R}^n \to [-\infty, \infty]$ of f defined by:

$$f^*(y) := \sup_{x\in\mathbb{R}^n} \left(x^\mathsf{T} y - f(x)\right), \qquad y \in \mathbb{R}^n$$

Conjugate function is defined for any function f, not only convex functions. Since f^* is the pointwise supremum of affine functions of y it is closed and convex for any f. Then (12.29a) says:

$$\bar{y} \in \partial f(\bar{x}) \iff \bar{y}^\mathsf{T}\bar{x} = f(\bar{x}) + f^*(\bar{y}) \tag{12.29b}$$

i.e., $\bar{y}$ is a subgradient of f at $\bar{x}$ if and only if $\bar{x}$ attains the maximization in $f^*(\bar{y})$. When f is CPC, $f^{**} = f$ and the property becomes symmetric. We summarize important properties of conjugate functions and subgradients in the following result taken from [54, Propositions 1.6.1, 5.4.3 and 5.4.4].

Lemma 12.16 (Conjugate function and subgradient)**.** Let $f : \mathbb{R}^n \to (-\infty, \infty]$.

1 Its conjugate f^* is closed and convex.
2 If f is convex then the properness of any one of f, f^*, f^{**} implies the properness of the other two. In particular if f is proper convex then f^* is CPC (closed proper convex).
3 If f is CPC then $f(x) = f^{**}(x)$ for $x \in \mathbb{R}^n$.
4 *Envelop theorem*: If f is CPC then, for any $\bar{x} \in \mathrm{dom}(f)$, $\bar{y} \in \mathrm{dom}(f^*)$,

$$\bar{x}^\mathsf{T}\bar{y} = f(\bar{x}) + f^*(\bar{y}) \iff \bar{y} \in \partial f(\bar{x}) \iff \bar{x} \in \partial f^*(\bar{y})$$

5 *Dual differentiability and optimality*: If f is CPC then

[5] The right-hand side of (12.28) is attained (i.e., $\exists \bar{y}$ with $\bar{y}^\mathsf{T} v = \sup_{y\in\partial f(x)} y^\mathsf{T} v$) if $x \in \mathrm{int}(\mathrm{dom}(f))$, not just $x \in \mathrm{ri}(\mathrm{dom}(f))$, according to Lemma 12.15.

1 $f^*(y)$ is differentiable at $\bar{y} \in \operatorname{int}(\operatorname{dom}(f^*))$ if and only if $f^*(\bar{y}) := \sup_{x\in\mathbb{R}^n} \left(x^\mathsf{T}\bar{y} - f(x)\right)$ is attained at a unique $\bar{x} \in \mathbb{R}^n$.
2 The set $\arg\min_{x\in\mathbb{R}^n} f(x)$ of unconstrained minima of f is equal to $\partial f^*(0)$.
3 Hence x^* is an unconstrained minimizer if and only if $x^* \in \partial f^*(0)$ if and only if $0 \in \partial f(x^*)$.

Lemma 12.16.4 is a form of envelop theorem for CPC functions: it says that, if $f^*(\bar{y}) = \sup_x(x^\mathsf{T}\bar{y} - f(x)) = \bar{x}^\mathsf{T}\bar{y} - f(\bar{x})$, then $\bar{x}$ is a subgradient of f^* at $\bar{y}$. An implication of Lemma 12.16.5 is that the dual function of a convex program is differentiable if the minimum of the Lagrangian over the primal variable is uniquely attained.

Example 12.10 (Differentiable functions)**.** Consider the real-valued convex and differentiable function $f : \mathbb{R}^n \to (-\infty, \infty)$. The subdifferential of f at $\bar{x}$ is $\partial f(\bar{x}) = \{\nabla f(\bar{x})\}$. Then (12.29b) reduces to

$$\nabla^\mathsf{T} f(\bar{x})\bar{x} \;=\; f(\bar{x}) + f^*(\nabla f(\bar{x})) \;=\; f(\bar{x}) + \sup_{x\in\mathbb{R}^n} \Big(\nabla^\mathsf{T} f(\bar{x})x - f(x)\Big)$$

which says that the supremum on the right-hand side is attained at $\bar{x}$ when f is convex, or re-arranging,

$$f(x) \;\geq\; f(\bar{x}) + \nabla^\mathsf{T} f(\bar{x})(x - \bar{x}), \qquad x \in \mathbb{R}^n$$

which is a property of convexity (or definition of subgradient).

Suppose further that, for all $\bar{y} \in \mathbb{R}^n$, the supremum in $f^*(\bar{y}) := \sup_{x\in\mathbb{R}^n}(\bar{y}^\mathsf{T}x - f(x))$ is attained at a unique $\bar{x}$ so that f^* is differentiable on $\mathbb{R}^n$. Then the envelop theorem in Lemma 12.16 reduces to $\bar{y} = \nabla f(\bar{x})$ if and only if $\bar{x} = \nabla f^*(\bar{y})$. This says that the derivative of the conjugate function at $\bar{y}$,

$$f^*(\bar{y}) \;:=\; \sup_{x\in\mathbb{R}^n} \Big(x^\mathsf{T}\bar{y} - f(x)\Big) \;=\; \bar{x}^\mathsf{T}\bar{y} - f(\bar{x})$$

is the unique maximizer $\bar{x}$. Moreover the unconstrained supremum of the concave function $\bar{y}^\mathsf{T}x - f(x)$ of x is attained at $\bar{x}$ that satisfies $\nabla f(\bar{x}) = \bar{y}$. □

Indicator δ_X and support functions σ_X.

It is shown in Exercise 12.12 that for any nonempty set $X \subseteq \mathbb{R}^n$, the conjugate of the indicator function δ_X is the support function σ_X. Since δ_X is proper, Lemma 12.16 implies that σ_X is CPC (closed proper convex) as long as X is nonempty. This however does not in itself imply that δ_X is itself CPC nor $\delta_X = \sigma_X^*$. Indeed δ_X is CPC if and only if X is a closed nonempty convex set, in which case the conjugate σ_X^* of the support function is indeed δ_X. The results in Exercise 12.12 are summarized in Table 12.2.

For a closed nonempty convex set X we can interpret $\partial\sigma_X(x) = \{y \in \mathbb{R}^n : x^\mathsf{T}y = \sigma_X(x)\}$ as a form of envelop theorem for the function $\sigma_X(x) := \sup_{y\in X} y^\mathsf{T}x$. We can also interpret it as a supporting hyperplane. Indeed fix any $\bar{x} \in X$. Then $\xi := \sigma_X(\bar{x})$ is a constant and hence $\partial\sigma_X(\bar{x}) = \{y \in \mathbb{R}^n : \bar{x}^\mathsf{T}y = \xi\}$ is a hyperplane in $\mathbb{R}^n$. Since $\bar{x}^\mathsf{T}y \leq \xi$

function f	conjugate f^*	subdifferential $\partial f(x)$	condition
$\delta_X(x)$	$\sigma_X(x)$	$N_X(x)$	if X is nonempty convex
$\delta_X(x)$	$\delta_{X^\circ}(y)$	$N_X(x)$	if X is a nonempty convex cone
$\sigma_X(x)$	$\delta_X(x)$	$\{y \in \mathbb{R}^n : x^\mathsf{T} y = \sigma_X(x)\}$	if X is closed nonempty convex

Table 12.2 Indicator function $\delta_X(x) := 0$ if $x \in X$ and ∞ otherwise, support function $\sigma_X(x) := \sup_{y \in X} y^\mathsf{T} x$, their conjugates and subdifferentials ($N_X(x)$ is normal cone of X at x).

for all $y \in X$, the hyperplane $\partial\sigma_X(\bar{x})$ contains X in its "lower" halfspace. If there is a finite $\bar{y} \in X$ that attains the supremum in $\sigma_X(\bar{x}) := \sup_{y \in X} \bar{x}^\mathsf{T} y$, then $\partial\sigma_X(\bar{x})$ is a supporting hyperplane of X at $\bar{y}$. See Figure 12.13.

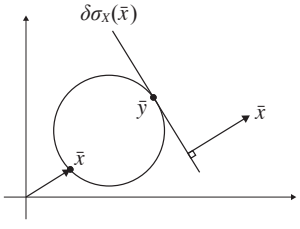


Figure 12.13 For a nonempty closed convex X, $\partial\sigma_X(\bar{x})$ is a supporting hyperplane of X at $\bar{y}$.

12.3.3 Subdifferential calculus

The subdifferential of functions is fundamental. In particularly the result on the sum of functions in Theorem 12.18 is used to derive an exact optimality condition for nonsmooth convex optimization in Chapter 12.5 that underlies the KKT condition. The proof of Theorem 12.18 makes use of the following result on the existence of a dual optimal solution that attains strong duality (even if the primal optimal value is not attained).

Consider the convex optimization

$$f^* := \min_{x \in \mathbb{R}^n} f(x) \text{ s.t. } x \in X', \ Ax = b \tag{12.30a}$$

where the nonempty convex set $X' \subseteq \mathbb{R}^n$ is the intersection of a polyhedral set P and a convex set C:

$$X' := P \cap C$$

$A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, and $f : \mathbb{R}^n \to (-\infty, \infty]$ is an extended real-valued proper convex function. Let the Lagrangian function be

$$L(x, \lambda) := f(x) + \lambda^\mathsf{T}(Ax - b), \qquad x \in \mathbb{R}^n, \ \lambda \in \mathbb{R}^m$$

the dual function be $d(\lambda) := \inf_{x \in X'} L(x, \lambda)$ and the dual problem be

$$d^* := \sup_{\lambda \in \mathbb{R}^m} \; d(\lambda) \tag{12.30b}$$

The problem (12.30) is a special case of (12.41) studied in detail in Chapter 12.7.1 when there is no explicit inequality constraint $h(x) \leq 0$. The following result is a special case of Theorem 12.27 there (whose proof does not require Theorem 12.18 so there is no circular argument). It is presented here because it is needed to prove Theorem 12.18 on subdifferential calculus.

Theorem 12.17 (Slater Theorem)**.** Consider the optimization problem (12.30) with a mixture of polyhedral constraints. Suppose the following conditions hold:

- *Finite primal value*: $f^* > -\infty$.
- *Convexity*: f is proper convex; P is a nonempty polyhedral set and C is a nonempty convex set.
- *Slater condition*: There exists $\bar{x} \in \mathrm{ri}(\mathrm{dom}(f)) \cap P \cap \mathrm{ri}(C)$ such that $A\bar{x} = b$.

Then

1 $f^* = d^*$.
2 The set of dual optimal solutions λ^* with $d(\lambda^*) = d^*$ is nonempty and convex.

Theorem 12.18 is taken from [54, Propositions 5.4.5–5.4.6, p.192]. Its proof makes use of Theorem 12.17 and leads to the requirement of constraint qualifications. They take the form that the intersection of the effective domains of various polyhedral functions is nonempty (if some of the functions are not polyhedral, their effective domains are replaced by their relative interiors).

Theorem 12.18 ([54])**.** 1 *Sum of functions.* Let $f_i : \mathbb{R}^n \to (-\infty, \infty]$, $i = 1, \ldots, m$, be convex functions. Suppose $F(x) := \sum_i f_i(x)$ is proper. If, for some $\bar{m}$ with $1 \leq \bar{m} \leq m$, the functions f_i, $i = 1, \ldots, \bar{m}$, are polyhedral and

$$\left(\cap_{i=1}^{\bar{m}} \mathrm{dom}(f_i)\right) \bigcap \left(\cap_{i=\bar{m}+1}^{m} \mathrm{ri}(\mathrm{dom}(f_i))\right) \neq \emptyset$$

then F is convex and

$$\partial F(x) = \sum_i \partial f_i(x), \qquad \forall x \in \bigcap_{i=1}^{m} \mathrm{dom}(f_i)$$

When f_i are differentiable this reduces to $\nabla F(x) = \sum_i \nabla f_i(x)$.

2 *Chain rule.* Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be a convex function and $A \in \mathbb{R}^{m \times n}$. Suppose $F(x) := f(Ax)$ is proper. If

- either f is polyhedral, or
- there exists an $\tilde{x} \in \mathbb{R}^n$ such that $A\tilde{x} \in \mathrm{ri}(\mathrm{dom}(f))$

then F is convex and $\partial F(x) = A^\mathsf{T} \partial f(Ax)$ for all $x \in \mathbb{R}^n$. When f is differentiable this reduces to $\nabla F(x) = A^\mathsf{T} \nabla f(Ax)$.

Proof *Sum of functions.* Fix an $\bar{x} \in \bigcap_{i=1}^m \operatorname{dom}(f_i)$. Then $\bar{x} \in \operatorname{dom}(F)$. By Lemma 12.15, $\partial f_i(\bar{x})$ and $\partial F(\bar{x})$ are nonempty convex and compact. The proof of $\partial F(\bar{x}) \supseteq \sum_i \partial f_i(\bar{x})$ needs no assumption; its converse does. For any $\bar{y}_i \in \partial f_i(\bar{x})$ we have

$$f_i(x) \;\geq\; f_i(\bar{x}) + \bar{y}_i^{\mathsf{T}}(x-\bar{x}), \qquad x \in \mathbb{R}^n,\; i = 1,\ldots,m$$

Hence

$$F(x) \;:=\; \sum_i f_i(x) \;\geq\; F(\bar{x}) + \left(\sum_i \bar{y}_i\right)^{\mathsf{T}} (x-\bar{x}), \qquad x \in \mathbb{R}^n$$

i.e., $\sum_i \bar{y}_i \in \partial F(\bar{x})$.

For the converse, suppose $\bar{y} \in \partial F(\bar{x})$. Then

$$\min_{x \in \mathbb{R}^n} F(x) - \bar{y}^{\mathsf{T}} x \;\geq\; F(\bar{x}) - \bar{y}^{\mathsf{T}}\bar{x} \;\in\; \mathbb{R} \tag{12.31}$$

i.e., the finite minimum on the left-hand side is attained at $\bar{x}$. To apply Theorem 12.17, we write $F(x) = \sum_i f(x_i)$ with the constraints $x_i = x \in \mathbb{R}^n$. Then $(\bar{x}, x_i = \bar{x}, i = 1,\ldots,m)$ is a minimizer of the following convex optimization:

$$f^* \;=\; \min_{x, x_i \in \mathbb{R}^n} \sum_i f_i(x_i) - \bar{y}^{\mathsf{T}} x \quad \text{s.t.} \quad x_i \in \operatorname{dom}(f_i),\; x_i = x,\; i = 1,\ldots,m \tag{12.32a}$$

Its dual objective function is

$$d(\lambda) \;:=\; \min_{x \in \mathbb{R}^n,\, x_i \in \operatorname{dom}(f_i)} \sum_i f_i(x_i) - \bar{y}^{\mathsf{T}} x - \sum_i \lambda_i^{\mathsf{T}} (x_i - x) \tag{12.32b}$$

where $\lambda := (\lambda_1,\ldots,\lambda_m) \in \mathbb{R}^{mn}$. The application of Theorem 12.17 to (12.32) implies that strong duality holds and that any optimal dual variable $\bar{\lambda}_i$ yields a subgradient in $\partial f_i(\bar{x})$ at $\bar{x}$.

Specifically X' in (12.30) corresponds to the convex constraint

$$X' \;:=\; P \cap C \;:=\; \left(\cap_{i=1}^{\bar{m}} \operatorname{dom}(f_i)\right) \bigcap \left(\cap_{i=\bar{m}+1}^{m} \operatorname{dom}(f_i)\right)$$

Clearly a (finite) primal optimal is attained at $x_i = x = \bar{x}$ due to (12.31). The condition in the theorem guarantees a point $x_i := \tilde{x} \in P \cap \operatorname{ri}(C)$ such that $x_i = x := \tilde{x}$. Theorem 12.17, then implies that strong duality holds for (12.32) and there is a dual optimal solution $\bar{\lambda} := (\bar{\lambda}_1,\ldots,\bar{\lambda}_m) \in \mathbb{R}^{mn}$. Therefore, from (12.32), we have

$$d(\bar{\lambda}) \;:=\; \min_{x \in \mathbb{R}^n,\, x_i \in \operatorname{dom}(f_i)} \sum_i \left(f_i(x_i) - \bar{\lambda}_i^{\mathsf{T}} x_i\right) - \left(\bar{y} - \sum_i \bar{\lambda}_i\right)^{\mathsf{T}} x$$

For the dual problem $\max_\lambda d(\lambda)$, we must have $\bar{y} = \sum_i \bar{\lambda}_i$ since the minimization in $d(\lambda)$ over x is unconstrained. Strong duality then implies

$$d(\bar{\lambda}) \;=\; f^* \;=\; \sum_i \left(f_i(\bar{x}) - \bar{\lambda}_i^{\mathsf{T}} \bar{x}\right)$$

where the last equality follows because $\bar{y} = \sum_i \bar{\lambda}_i$ and $(\bar{x}, x_i = \bar{x}, i = 1,\ldots,m)$ is a

minimizer of (12.32a). Since we can extend the minimization in $d(\lambda)$ over x_i to $\mathbb{R}^n$, this implies (substituting again $\bar{y} = \sum_i \bar{\lambda}_i$)

$$d(\bar{\lambda}) \;=\; \min_{x_i \in \mathbb{R}^n} \sum_i \left(f_i(x_i) - \bar{\lambda}_i^\mathsf{T} x_i \right) \;=\; \sum_i \min_{x_i \in \mathbb{R}^n} \left(f_i(x_i) - \bar{\lambda}_i^\mathsf{T} x_i \right) \;=\; \sum_i \left(f_i(\bar{x}) - \bar{\lambda}_i^\mathsf{T} \bar{x} \right)$$

The last equality means that, for every i, $f_i(\bar{x}) - \bar{\lambda}_i^\mathsf{T} \bar{x} = \min_{x_i \in \mathbb{R}^n} \left(f_i(x_i) - \bar{\lambda}_i^\mathsf{T} x_i \right)$, i.e., $\bar{\lambda}_i \in \partial f_i(\bar{x})$ according to (12.26b). This complete the proof of part 1.

Chain rule. The proof follows a similar argument as that for part 1. Clearly F is convex since f is. Fix an $\bar{x} \in \mathbb{R}^n$. If $A\bar{x} \notin \mathrm{dom}(f)$ then $\bar{x} \notin \mathrm{dom}(F)$ and hence $\partial F(\bar{x}) = \partial f(A\bar{x}) = \emptyset$ by definition. Suppose then $A\bar{x} \in \mathrm{dom}(f)$. The proof of $\partial F(x) \supseteq A^\mathsf{T} \partial f(Ax)$ needs no assumptions; its converse does.

Let $\bar{\xi} \in \partial f(A\bar{x}) \subseteq \mathbb{R}^m$ be any subgradient of f at $A\bar{x}$. Then

$$F(x) - F(\bar{x}) = f(Ax) - f(A\bar{x}) \geq \bar{\xi}^\mathsf{T}(Ax - A\bar{x}) = \left(\bar{\xi}^\mathsf{T} A \right)(x - \bar{x}), \quad x \in \mathbb{R}^n \tag{12.33}$$

i.e., $\bar{y} := A^\mathsf{T} \bar{\xi} \in \mathbb{R}^n$ is in $\partial F(\bar{x})$. This shows $A^\mathsf{T} \partial f(A\bar{x}) \subseteq \partial F(\bar{x})$.

For the converse (under the assumption in the theorem), suppose $\bar{y} \in \partial F(\bar{x})$. We will show that there exists an $\bar{\lambda} \in \mathbb{R}^m$ such that $\bar{\lambda} \in \partial f(A\bar{x})$ and $\bar{y} = A^\mathsf{T} \bar{\lambda}$. From the definition (12.26b) of subgradient we have

$$F(\bar{x}) - \bar{y}^\mathsf{T} \bar{x} \;=\; \min_{x \in \mathbb{R}^n} \; F(x) - \bar{y}^\mathsf{T} x \;\in\; \mathbb{R}$$

i.e., the finite minimum of the right-hand side is attained at $\bar{x}$. Hence $(\bar{x}, A\bar{x})$ is a minimizer of the following constrained convex optimization:

$$\min_{(x,z) \in \mathbb{R}^{n+m}} \; f(z) - \bar{y}^\mathsf{T} x \qquad \text{s.t.} \qquad z \in X' := \mathrm{dom}(f),\; z = Ax \tag{12.34}$$

If f is polyhedral, then $X' := \mathrm{dom}(f) =: P$ is a polyhedral set. Otherwise $X' =: C$ is a convex set since f is a convex function. In the former case the assumption that F is proper means that there exists $\tilde{x} \in \mathbb{R}^n$ such that $\tilde{z} := A\tilde{x} \in X'$. In the latter case the assumption in the theorem means that there exists $\tilde{x} \in \mathbb{R}^n$ such that $\tilde{z} := A\tilde{x} \in \mathrm{ri}(X')$. In both cases Theorem 12.17 implies that strong duality holds and there exists an optimal dual variable $\bar{\lambda} \in \mathbb{R}^m$ that attains the dual optimal value:

$$\min_{x \in \mathbb{R}^n,\, z \in \mathrm{dom}(f)} \left(f(z) - \bar{\lambda}^\mathsf{T} z - (\bar{y} - A^\mathsf{T} \bar{\lambda})^\mathsf{T} x \right) \;=\; f(A\bar{x}) - \bar{y}^\mathsf{T} \bar{x}$$

where the left-hand side is the dual function of (12.34) evaluated at the dual optimal point $\bar{\lambda}$ and the right-hand side is the primal optimal value attained at $(\bar{x}, A\bar{x})$. Since the minimization over x is unconstrained we must have $\bar{y} = A^\mathsf{T} \bar{\lambda}$. Clearly we can extend the minimization over z to $\mathbb{R}^m$ and hence we have

$$\min_{z \in \mathbb{R}^m} f(z) - \bar{\lambda}^\mathsf{T} z \;=\; f(A\bar{x}) - \bar{y}^\mathsf{T} \bar{x} \;=\; f(A\bar{x}) - \bar{\lambda}^\mathsf{T}(A\bar{x})$$

i.e., $\bar{\lambda} \in \partial f(A\bar{x})$ by definition (12.26b). This completes the proof that $\partial F(x) = A^\mathsf{T} \partial f(Ax)$. □

Example 12.11 ($N_{H\cap K}(x) = N_H(x) + N_K(x)$)**.** Consider the linear program:

$$f^* \quad := \quad \min_{x\in\mathbb{R}^n} c^\mathsf{T} x \qquad \text{s.t.} \qquad x \in H \cap K$$

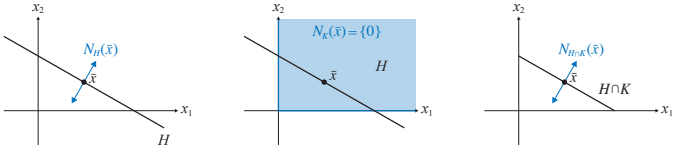
where $H := \{x : Ax = b\}$ is a polyhedron and $K := \{x : x \geq 0\}$ is a convex cone. If f^* (or the dual objective value) is finite, then the effective domain $\operatorname{dom}(c^\mathsf{T} x) = H \cap K$ is nonempty. Theorem 12.18 then implies that

$$\partial\left(\delta_H(x^*) + \delta_K(x^*)\right) \;=\; \partial\delta_H(x^*) + \partial\delta_K(x^*)$$

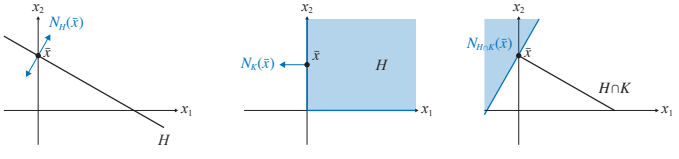
i.e.,

$$N_{H\cap K}(x) \;=\; N_H(x) + N_K(x)$$

This is illustrated in Figure 12.14. □



(a) At a relative interior point $\bar{x}$ of $H \cap K$.



(b) At a relative boundary point $\bar{x}$ of $H \cap K$

Figure 12.14 Example 12.11: Normal cones in Theorem 12.3 satisfy $N_{H\cap K}(\bar{x}) = N_H(\bar{x}) + N_K(\bar{x})$ at all points $\bar{x} \in H \cap K$.

Theorem 12.19. 1 *Finite max.* Let $F(x) := \max\{f_1(x), \ldots, f_m(x)\}$ where $f_i : \mathbb{R}^n \to \mathbb{R}$ are real-valued (and hence proper) convex functions. For any $x \in \mathbb{R}^n$ let

$$I(x) := \{i : f_i(x) = F(x)\}$$

Then

$$dF(x;v) = \max_{i\in I(x)} df_i(x;v), \qquad \forall x, v \in \mathbb{R}^n$$
$$\partial F(x) = \operatorname{conv}\left(\partial f_i(x) : i \in I(x)\right), \qquad \forall x \in \mathbb{R}^n$$

2 *Arbitrary max.* Let $F(x) := \max_{y\in Y} f(x,y)$ where $f : \mathbb{R}^n \times Y \to \mathbb{R}$ is a real-valued function and $Y \subseteq R^m$. Suppose for each $y \in Y$, $f(\cdot, y)$ is convex and hence continuous on $\mathbb{R}^n$. Fix an $\bar{x}$ and suppose there exists a neighborhood $U(\bar{x})$ of $\bar{x}$ such that for each $x \in U(\bar{x})$, $f(x,\cdot)$ is upper semicontinuous on Y.

Let $Y(x) := \{y : f(x,y) = F(x)\}$. Then

$$dF(\bar{x};v) = \sup_{y \in Y(\bar{x})} d_x f(\bar{x},y;v), \qquad \forall v \in \mathbb{R}^n$$
$$\partial F(\bar{x}) = \mathrm{cl}\,(\mathrm{conv}\,(\partial_x f(\bar{x},y) : y \in Y(\bar{x})))$$

where $d_x f(x,y;v)$ and $\partial_x f(x,y)$ are respectively the directional derivative and subdifferential of f with respect to x.

Remark 12.6. Theorem 12.19 is used in Exercise 13.12 to derive the subdifferentials of dual functions defined through minimization over primal variables.

1 Theorem 12.19.1 generalizes Theorem 8.21 from the case where f is real-valued and jointly continuous in (x,y) and Y is compact to the case where f may not be continuous in x and Y may not be compact. It is proved in e.g. [54, Example 5.4.5, p.199]. Since f_i are real-valued convex and hence proper and continuous on $\mathrm{dom}(f_i) = \mathbb{R}^n$, F is also a real-valued convex continuous function. Since $\partial f_i(x)$ is nonempty convex compact by Lemma 12.15, so is $\partial F(x)$.
2 Theorem 12.19.2 is taken from [141, Proposition 4.5.2, p.76].

□

Remark 12.7. Consider a real-valued function $f : \mathbb{R}^n \times Y \to \mathbb{R}$ and

$$F(x) := \sup_{y \in Y} f(x,y), \qquad G(x) := \inf_{y \in Y} f(x,y)$$

where Y is an arbitrary subset of $\mathbb{R}^m$.

1 *Taking supremum.* If f is convex in x for every $y \in Y$ then $F(x)$ is convex in x as Theorem 8.21 shows. Moreover if $f(\cdot,y)$ is closed for each $y \in Y$ then $F(\cdot)$ is closed as well ([54, Proposition 1.1.6, p.13]). This is the situation e.g. when f is the Lagrangian function of a constrained optimization.
2 *Taking infimum.* If $f(x,y)$ is *jointly* convex in (x,y) instead (this is not the case with Lagrangian functions) then $G(x)$ is convex ([54, Proposition 3.3.1, p.122]). Moreover the epigraph $\mathrm{epi}(G(x)) := \{(x,z) : z \geq G(x), x \in \mathbb{R}^n\}$ is essentially the projection of $\mathrm{epi}(f) := \{(x,y,z) : z \geq f(x,y), x \in \mathbb{R}^n, y \in Y\}$ on the space of (x,z), except possibly for some boundary points x when the infimum over $y \in Y$ is not attained in which case $(x,G(x))$ are missing. Precisely

$$P(\mathrm{epi}(f)) \subseteq \mathrm{epi}(G) \subseteq \mathrm{cl}\,P(\mathrm{epi}(f))$$

where the projection P is defined by $P(S) := \{(x,z) : (x,y,z) \in S\}$ for any subset $S \subseteq \mathbb{R}^n \times Y \times \mathbb{R}$. □

We next use the tools developed in Chapters 12.3.1, 12.3.2 and 12.3.3 to derive optimality conditions for general convex optimization, following the structure of Chapter 8.3.

12.4 Characterization: saddle point = p-d optimality + strong duality

In this section we present a primal-dual characterization of an optimal solution when some or all of the constraints are specified explicitly and can be dualized. In smooth optimization the Saddle Point Theorem 8.14 states that a saddle point attains primal-dual optimality and strong duality. We show that this characterization extends directly to the nonsmooth setting, without the need for the machinery in Chapters 12.3.1, 12.3.2 and 12.3.3 for nonsmooth analysis.

Consider the optimization problem where the feasible set is partially specified by constraint functions:

$$f^* \;:=\; \min_{x\in\mathbb{R}^n} f(x) \;\text{ s.t. }\; x\in X',\;\; g(x)=0,\;\; h(x)\le 0 \tag{12.35}$$

where $X'\subseteq\mathbb{R}^n$ is a nonempty set and $f:\mathbb{R}^n\to(-\infty,\infty]$, $g:\mathbb{R}^n\to(-\infty,\infty]^m$ and $h:\mathbb{R}^n\to(-\infty,\infty]^l$ are extended real-valued functions. As for the smooth case in Chapter 8.3.1, we do not assume X' to be a convex set or f,g,h be convex functions. Therefore (12.35) is generally a nonconvex problem.

Let the Lagrangian function be

$$L(x,\lambda,\mu) := f(x) \;+\; \lambda^{\mathsf{T}} g(x) \;+\; \mu^{\mathsf{T}} h(x), \qquad x\in\mathbb{R}^n,\; \lambda\in\mathbb{R}^m,\; \mu\in\mathbb{R}^l \tag{12.36a}$$

the dual function be

$$d(\lambda,\mu) := \inf_{x\in X'} L(x,\lambda,\mu) \tag{12.36b}$$

and the dual problem be

$$d^* := \sup_{\lambda,\mu\ge 0} \; d(\lambda,\mu) \tag{12.36c}$$

Let $X := \{x\in\mathbb{R}^n : x\in X', g(x)=0, h(x)\le 0\}$ denote the primal feasible set and $Y := \{(\lambda,\mu)\in\mathbb{R}^{m+l} : \mu\ge 0\}$ the dual feasible set. The primal problem (12.35) is the same as (8.25) in Chapter 8.3.1 except the cost and constraint functions are allowed to be nonsmooth and extended real-valued (see Remark 8.4 for the case where $X'\subseteq\mathbb{R}^n$ in (8.25)). The Saddle Point Theorem 8.14 applies directly in the nonsmooth setting here. For simplicity, we require a saddle point to attain a finite value of the Lagrangian L by definition.

Definition 12.4 (Saddle point for extended real-value functions)**.** A point $(x^*,\lambda^*,\mu^*)\in X'\times Y$ is called a *saddle point* of the Lagrangian L if it satisfies

$$\max_{(\lambda,\mu)\in Y} L(x^*,\lambda,\mu) \;=\; L(x^*,\lambda^*,\mu^*) \;=\; \min_{x\in X'} L(x,\lambda^*,\mu^*) \;\in\; \mathbb{R} \tag{12.37}$$

In particular this common value $L(x^*,\lambda^*,\mu^*)$ is finite.

With this finiteness requirement, Definition 12.4 is equivalent to Definition 8.8 for real-valued functions f,g,h, and Theorem 8.14 on primal-duality optimality and strong duality extends directly to the nonsmooth setting.

Theorem 12.20 (Saddle-point Theorem 8.14)**.** Consider the primal problem (12.35) and its dual (12.36). A point $(x^*, \lambda^*, \mu^*) \in X' \times Y$ is a saddle point that satisfies (12.37) if and only if

1 It is optimal-dual optimal, i.e., x^* is optimal for (12.35) and (λ^*, μ^*) is optimal for (12.36).
2 The duality gap is zero at (x^*, λ^*, μ^*), i.e.,

$$d(\lambda^*, \mu^*) = d^* = f^* = f(x^*) \tag{12.38}$$

In particular a saddle point (x^*, λ^*, μ^*), if it exists, attains both the primal and dual objective values (f^*, d^*).

Proof The proof of Theorem 8.14 does not use any smoothness properties of the cost and constraint functions f, g, h, except that they are real-valued. In particular, when $(x^*, \lambda^*, \mu^*) \in X' \times Y$ is a saddle point, the proof there uses Remark 8.3 to deduce that $x^* \in X$ is primal feasible. This conclusion still holds here due to the finiteness requirement in Definition 12.4. Since the weak duality lemma 8.13 applies to extended real-valued functions, it can be checked that the argument in the proof of Theorem 8.14 goes through in the nonsmooth setting. □

12.5 Characterization: generalized KKT condition

Consider the convex optimization

$$\min_{x \in \mathbb{R}^n} f(x) \text{ s.t. } x \in P \cap C \tag{12.39}$$

where $P \subseteq \mathbb{R}^n$ is a nonempty polyhedral set, $C \subseteq \mathbb{R}^n$ is a nonempty convex set, and $f : \mathbb{R}^n \to (-\infty, \infty]$ is a proper convex extended real-valued function. In particular f may not be differentiable, though subgradients always exist since f is convex. We now derive an exact characterization of primal optimal solutions when they exist. When the feasible set $P \cap C$ is specified explicitly by equality and inequality constraints, the characterization reduces to the KKT condition for nonsmooth convex problems. This is studied in Chapter 12.8.

Corollary 12.13 in Chapter 12.3.2 says that a vector x^* is an unconstrained minimizer of an extended real-valued convex function f if and only if $0 \in \partial f(x^*)$. For constrained minimization (12.39) this condition is generalized to the existence of a subgradient $y^* \in \partial f(x^*)$ such that $-y^*$ is in the normal cone $N_X(x^*)$ of the feasible set $P \cap C$ at x^*. Constrained optimization also requires a constraint qualification which is a kind of feasibility condition, e.g., $\text{dom}(f) \cap P \cap \text{ri}(C)$ is nonempty if f is polyhedral. If f is not polyhedral then $\text{dom}(f)$ is replaced by $\text{ri}(\text{dom}(f))$.

Theorem 12.21 (Generalized KKT condition)**.** Consider the convex optimization (12.39) with a nonempty polyhedral set P, a nonempty convex set C, and a proper convex function f. Suppose one of the following constraint qualifications holds, depending on whether f is polyhedral:

1 $\operatorname{ri}(\operatorname{dom}(f)) \cap P \cap \operatorname{ri}(C) \neq \emptyset$;
2 f is polyhedral and $\operatorname{dom}(f) \cap P \cap \operatorname{ri}(C) \neq \emptyset$;

Then $x^* \in P \cap C$ is optimal for (12.39) if and only if

$$0 \in \partial f(x^*) + N_P(x^*) + N_C(x^*) \tag{12.40a}$$

i.e., there exists a subgradient $y^* \in \partial f(x^*)$ such that $-y^* \in N_P(x^*) + N_C(x^*)$, or equivalently

$$y^{*\mathsf{T}}(x - x^*) \geq 0, \qquad \forall x \in P \cap C \tag{12.40b}$$

Proof The proof is from [54, Proposition 5.4.7, p.195]. The problem (12.39) is equivalent to the unconstrained minimization:

$$\min_{x \in \mathbb{R}^n} f(x) + \delta_P(x) + \delta_C(x)$$

where the indicator function $\delta_{X'}(x) = 0$ if $x \in X'$ and ∞ if $x \notin X'$. Corollary 12.13 in Chapter 12.3.2 says that $x^* \in P \cap C$ is optimal if and only if $0 \in \partial\left(f(x^*) + \delta_P(x^*) + \delta_C(x^*)\right)$. The stated constraint qualifications allow us to apply the result on the sum of functions in Theorem 12.18 to conclude that $x^* \in P \cap C$ is optimal if and only if

$$0 \in \partial f(x^*) + \partial \delta_P(x^*) + \partial \delta_C(x^*) = \partial f(x^*) + N_P(x^*) + N_C(x^*)$$

where the second equality follows from Table 12.2. □

Theorem 12.21 characterizes an optimal solution x^* but does not guarantee its existence. See Examples 8.9 and 8.10 in Chapter 8 for cases where primal optimal solutions do not exist even though the constraint qualifications in Theorem 12.21 are satisfied. In both examples the feasible set is not compact, but the primal optimal objective values are finite, strong duality holds, and dual optimal solutions exist. As discussed in Remark 12.4 we only need one subgradient $y^* \in \partial f(x^*)$ to certify the optimality of x^* and does not require $y^{\mathsf{T}}(x - x^*) \geq 0$ to hold for all $y \in \partial f(x^*)$. The theorem is proved by reducing the constrained minimization (12.39) to an unconstrained minimization using the indicator function δ_X. It illustrates the simplicity of argument based on the set theoretic concepts of nonsmooth optimization introduced in Chapter 12.1 and the concept of subdifferentials introduced in Chapters 12.3.2 and 12.3.3.

Remark 12.8 (Real-valued f)**.** 1 When $f : \mathbb{R}^n \to \mathbb{R}$ is real-valued then $\text{ri}(\text{dom}(f)) = \text{dom}(f) = \mathbb{R}^n$ and the constraint qualifications in Theorem 12.21 reduce to

$$P \cap \text{ri}(C) \neq \emptyset$$

whether or not f is polyhedral.

2 If the cost function f is differentiable then y^* and $\partial f(x^*)$ in the optimality condition in (12.40) can be replaced by $\nabla f(x^*)$.

Similarly for other duality and optimality conditions. □

When the feasible set $X := P \cap C$ is a general convex set X, Theorem 12.21 on the characterization of (primal) optimal solutions and Theorem 12.26 on its existence are almost all that we can say without more knowledge about X. When X is at least partially specified by affine equalities and convex inequalities, we characterize saddle points and strong duality in Theorem 12.20 of Chapter 12.4 and the existence of dual optimal solutions in the Slater Theorem 12.27 of Chapter 12.7.1. When the feasible set X is fully specified, all constraints can be dualized. When the normal cones $N_P(x^*)$ and $N_C(x^*)$ can be explicitly derived, such as those in Theorems 12.3, 12.4, 12.10 and Corollary 12.11, the exact optimality condition (12.40) reduces to KKT conditions; see Chapter 12.8.

12.6 Existence: primal optimal solutions

Theorem 12.21 of Chapter 12.5 provides an exact characterization of primal optimal solutions and the Saddle Point Theorem 12.20 of Chapter 12.4 characterizes saddle points as primal-dual optimal solutions that close the duality gap. They do not ensure that primal or dual optimal solutions exist. For smooth optimization Theorem 8.16 states that the primal optimal value is attained if the cost function is continuous and the feasible set is compact. It is a consequence of the Weierstrass theorem. In this section we extend this result to a nonsmooth setting where the continuity of the cost function is replaced by the closedness of f (recall that a function $f : \mathbb{R}^n \to [-\infty, \infty]$ is closed if and only if f is lsc on $\mathbb{R}^n$; see Remark 12.3).

A function $f : \mathbb{R}^n \to (-\infty, \infty]$ is called *radially unbounded* if $\lim_k f(x_k) = \infty$ for every sequence $\{x_k\}$ with $\|x_k\| \to \infty$. All nonempty level sets of a radially unbounded function are bounded. The next result from [54, Proposition 3.2.1, p.119] provides sufficient conditions for the existence of optimal solutions $x^* \in \mathbb{R}^n$ for unconstrained optimization.

Theorem 12.22 (Weierstrass Theorem)**.** Consider

$$\min_{x \in \mathbb{R}^n} \; f(x)$$

where $f : \mathbb{R}^n \to (-\infty, \infty]$ is closed and proper. If any of the following conditions holds:

1 $\mathrm{dom}(f)$ is bounded; or
2 There exists $\gamma \in \mathbb{R}$ such that the level set $V_\gamma := \{x : f(x) \le \gamma\}$ is nonempty and bounded; or
3 f is radially unbounded;

then the set $X^* \subseteq \mathbb{R}^n$ of unconstrained minima of f is nonempty and compact. □

A constrained optimization of f over a nonempty closed subset $X \subseteq \mathbb{R}^n$ can be turned into an unconstrained optimization of the extended real-valued function $f_X(x) : \mathbb{R}^n \to [-\infty, \infty]$ defined in (12.25). An optimality condition then follows immediately from Theorem 12.22 and the fact that f_X is closed if $\mathrm{dom}(f)$ is closed and f is lower semicontinuous on $\mathrm{dom}(f)$ (Exercise 12.15). It is a generalization of Theorem 8.16 to the nonsmooth setting.

Corollary 12.23 (Sufficient optimality condition)**.** Consider

$$\min_{x \in \mathbb{R}^n} \quad f(x) \quad \text{s.t.} \quad x \in X$$

where $X \subseteq \mathbb{R}^n$, $f : X \to (-\infty, \infty]$ and $X \cap \mathrm{dom}(f) \neq \emptyset$. If X is closed, f is lower semicontinuous at every $x \in X$, and one of the following holds:

1 X is bounded; or
2 There exists $\gamma \in \mathbb{R}$ such that the level set $V_\gamma := \{x : f(x) \le \gamma\}$ is nonempty and bounded; or
3 f is radially unbounded;

then the set $X^* \subseteq X$ of minima of f over X is nonempty and compact. □

CPC function f.

Theorem 12.22 and Corollary 12.23 guarantee that the minimum of f is attained (at a finite point in $\mathbb{R}^n$) when there is a nonempty level set that is bounded. When level sets are not bounded, the set X^* of constrained minima can be exactly characterized if f is not only closed and proper but also convex and X is closed and convex. The key idea is that x cannot wander to infinity within a level set V_γ while staying within its feasible set X. We next make this intuition precise.

Definition 12.5 (Recession cone)**.** Let $X \subseteq \mathbb{R}^n$ be a nonempty convex set.

1 A vector $d \in \mathbb{R}^n$ is a *direction of recession of* X if $x + \alpha d \in X$ for all $x \in X$ and all $\alpha \ge 0$.
2 The *recession cone of* X, denoted by $\mathrm{rc}(X)$, is the set of all directions of recession of X. □

Lemma 12.24. [54, Proposition 1.4.1; p.43] Let $X \subseteq \mathbb{R}^n$ be a nonempty closed convex set. Then

1 $\mathrm{rc}(X)$ is closed and convex.
2 $d \in \mathrm{rc}(X)$ as long as there exists one $x \in X$ such that $x + \alpha d \in X$ for all $\alpha \geq 0$.
3 $\mathrm{rc}(X)$ contains a nonzero direction if and only if X is unbounded. □

The next result allows us to define the direction of recession for a closed proper convex (CPC) function f in terms of its level set.

Lemma 12.25. [54, Proposition 1.4.5; p.51] Consider a closed proper convex function $f : \mathbb{R}^n \to (-\infty, \infty]$ and its level sets

$$V_\gamma := \{\, x \;:\; f(x) \leq \gamma \,\}, \qquad \gamma \in \mathbb{R}$$

Then:

1 All nonempty level sets V_γ have the same recession cone $\mathrm{rc}(V_\gamma) = \{d : (d, 0) \in \mathrm{rc}(\mathrm{epi}(f))\}$.
2 If one nonempty level set V_γ is compact, then all level sets are compact. □

In view of the lemma we can define, for a CPC function $f : \mathbb{R}^n \to (-\infty, \infty]$, the *recession cone of* f as $\mathrm{rc}(f) := \mathrm{rc}(V_\gamma)$ for any nonempty level set V_γ. A vector $d \in \mathrm{rc}(f)$ is called a *direction of recession of* f. A vector d is called a *common direction of recession of* f *and* X if $d \in \mathrm{rc}(f) \cap \mathrm{rc}(X)$. The next result from [54, Proposition 3.2.2; p.120] characterizes exactly the set X^* of minima of a constrained optimization.

Theorem 12.26. [54, Proposition 3.2.2; p.120] Consider

$$\min_{x \in \mathbb{R}^n} \; f(x) \quad \text{s.t.} \quad x \in X$$

where $X \subseteq \mathbb{R}^n$ is nonempty closed and convex, $f : \mathbb{R}^n \to (-\infty, \infty]$ is closed proper convex, and $X \cap \mathrm{dom}(f) \neq \emptyset$. The set $X^* \subseteq X$ of minima of f over X is nonempty, convex and compact if and only if X and f have no common nonzero direction of recession. □

Theorem 12.26 is used in Exercise 13.12 to derive the subdifferentials of dual functions defined through minimization over primal variables. If X and f in the theorem do have a common nonzero direction of recession, then either the optimal solution set is empty (infeasible problem) or else it is nonempty and unbounded (optimal value may be finite or infinite and may or may not be attained). This is because for any common nonzero direction d of recession in $\mathrm{rc}(X) \cap \mathrm{rc}(f)$, there is a feasible point $x \in X$ such that $x + \alpha d$ remains in X and in the level set V_γ as $\alpha \to \infty$. Moreover this holds for all nonempty level sets V_γ by Lemma 12.25. Therefore either $\lim_{\gamma \to -\infty} V_\gamma \neq \emptyset$ (limit exists because V_γ are nested) or $V_\gamma = \emptyset$ for small enough γ. In the former case there is a $d \in \mathrm{rc}(X) \cap \mathrm{rc}\left(\lim_{\gamma \to -\infty} V_\gamma\right)$ and the primal solution is not attained, e.g., $X = \mathbb{R}$,

$f(x) = x$ and $d = -1$. Otherwise there is a smallest γ_0 for which $V_{\gamma_0} \neq \emptyset$ and the primal optimal solution set is nonempty and unbounded since the intersection of $\mathrm{rc}(X)$ and $\mathrm{rc}(V_{\gamma_0})$ is nonempty (Exercise 12.16), e.g., $X = \mathbb{R}$, $f(x) = \max\{0, x\}$ and $d = -1$.

Example 12.12 (Linear program)**.** Consider the linear program (8.56a) reproduced here:

$$f^* := \min_{x \in \mathbb{R}^n} c^\mathsf{T} x \qquad \text{s.t.} \qquad Ax \geq b$$

where $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. If the feasible set X is bounded or if there is a $\gamma \in \mathbb{R}$ such that the level set V_γ is nonempty and bounded, then Corollary 12.23 implies that the set $X^* \subseteq X$ of optimal solutions is nonempty and compact. Consider then the case where X is unbounded and every nonempty level set $V_\gamma := \{x \in \mathbb{R}^n : c^\mathsf{T} x \leq \gamma\}$ is unbounded. This means that both $\mathrm{rc}(X)$ and $\mathrm{rc}(f)$ contain nonzero directions of recession (Lemma 12.24). Suppose f^* is finite.

Suppose $d \in \mathrm{rc}(X)$ and $d \neq 0$. Then there are two mutually exclusive cases:

1 $d \notin \mathrm{rc}(f)$ and $c^\mathsf{T} d > 0$: In this case Theorem 12.26 implies the existence of an optimal solution x^*; moreover the set X^* of optimal solutions is compact.
2 $d \in \mathrm{rc}(f)$ and $c^\mathsf{T} d = 0$: In this case Lemma 8.22 shows that an optimal solution x^* exists but X^* may not be compact.

To show that these are the only two possible cases when f^* is finite, suppose $d \in \mathrm{rc}(X) \cap \mathrm{rc}(f)$ and $d \neq 0$, i.e., for all $x \in X \cap V_\gamma$ and all $\alpha \geq 0$, $x + \alpha d \in X \cap V_\gamma$. This means $A(x + \alpha d) \geq b$ and $c^\mathsf{T} x + \alpha c^\mathsf{T} d \leq \gamma$ for all $\gamma \geq 0$. This is possible if only if

$$Ad \geq 0, \qquad c^\mathsf{T} d \leq 0$$

If $c^\mathsf{T} d < 0$, then letting $\alpha \to \infty$ the cost $c^\mathsf{T}(x + \alpha d) \to -\infty$, contradicting $f^* > -\infty$. Therefore if $d \in \mathrm{rc}(X)$ (i.e., $Ad \geq 0$), then either $d \in \mathrm{rc}(f)$ and $c^\mathsf{T} d = 0$, or $d \notin \mathrm{rc}(f)$ and $c^\mathsf{T} d > 0$. □

12.7 Existence: dual optimal solutions and strong duality

In Chapter 12.6 we study the existence of primal optimal solutions (Corollary 12.23 and Theorem 12.26). In this section we study dual optimality. In smooth optimization the Slater Theorem 8.17 states that a dual optimal solution exists and strong duality holds if the optimal primal value is finite (even if it is not attained) and the Slater condition is satisfied. We extend this assertion to the nonsmooth setting in Chapter 12.7.1 and provide a detailed proof in 12.7.2 and 12.7.3 (which also proves Theorem 8.17). These results are mostly adapted from [54, Chapters 4 and 5].

12.7.1 Slater Theorem

Consider the convex optimization (12.35) where the feasible set is specialized to be the intersection of a polyhedral set and a convex set and the equality constraint $g(x) = 0$ is polyhedral:

$$f^* := \min_{x \in \mathbb{R}^n} f(x) \text{ s.t. } x \in X', \ Ax = b, \ h(x) \leq 0 \tag{12.41a}$$

Here the nonempty convex set $X' \subseteq \mathbb{R}^n$ is the intersection of a polyhedral set P and a convex set C:

$$X' := P \cap C$$

$A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, and $f : \mathbb{R}^n \to (-\infty, \infty]$ and $h : \mathbb{R}^n \to (-\infty, \infty]^l$ are extended real-valued proper convex functions. [6] Suppose, for some $\bar{l}$ with $0 \leq \bar{l} \leq l$, $h_i, i = 1, \ldots, \bar{l}$, are polyhedral functions. In contrast to (12.35) the polyhedral equality constraint $Ax = b$ ensures that the feasible set of (12.41a) is convex.

Let the Lagrangian function be

$$L(x, \lambda, \mu) := f(x) + \lambda^\mathsf{T}(Ax - b) + \mu^\mathsf{T} h(x), \qquad x \in \mathbb{R}^n, \ \lambda \in \mathbb{R}^m, \ \mu \in \mathbb{R}^l$$

the dual function be

$$d(\lambda, \mu) := \inf_{x \in X'} L(x, \lambda, \mu), \qquad \lambda \in \mathbb{R}^m, \ \mu \in \mathbb{R}^l$$

and the dual problem be

$$d^* := \sup_{\lambda, \mu \geq 0} d(\lambda, \mu) \tag{12.41b}$$

The following result from [54, Proposition 5.3.6, p.175] extends the Slater Theorem 8.17 to the nonsmooth setting.

Theorem 12.27 (Slater Theorem)**.** Consider the optimization problem (12.41) with a mixture of polyhedral and nonpolyhedral constraints. Suppose the following conditions hold:

- *Finite primal value*: $f^* > -\infty$.
- *Convexity*: f, h are proper convex functions; P is a nonempty polyhedral set and C is a nonempty convex set.
- *Slater condition*: There exists $\bar{x} \in \text{ri}(\text{dom}(f)) \cap P \cap \text{ri}(C)$ such that $A\bar{x} = b$, $h_i(\bar{x}) \leq 0$, $i = 1, \ldots, \bar{l}$, and $h_i(\bar{x}) < 0$ for $i = \bar{l} + 1, \ldots, l$.

Then

1. $f^* = d^*$.
2. The set of dual optimal solutions (λ^*, μ^*) with $d(\lambda^*, \mu^*) = d^*$ is nonempty, convex and closed.

6

Remark 12.9 (Real-valued functions)**.** When f and h are real-valued the constraint qualification for strong duality in Theorem 12.27 can be slightly weakened to [54, Proposition 5.3.6, p.175]:

1 There exists $\tilde{x} \in P \cap \mathrm{ri}(C)$ such that $A\tilde{x} = b$ and $h_i(\tilde{x}) \leq 0$, $i = 1, \ldots, \bar{l}$; and
2 There exists $\bar{x} \in P \cap C$ such that $A\bar{x} = b$, $h_i(\bar{x}) \leq 0$, $i = 1, \ldots, \bar{l}$, and $h_i(\bar{x}) < 0$ for $i = \bar{l}+1, \ldots, l$.

□

Instead of the problem (12.41) where the constraints are explicitly decomposed into polyhedral constraints $x \in P$ and $Ax = b$ and (possibly nonpolyhedral) convex constraints $x \in C$ and $h(x) \leq 0$, we will prove Theorem 12.27 in the following equivalent but simpler form:

$$f^* := \min_{x \in \mathbb{R}^n} f(x) \text{ s.t. } x \in X', \ h(x) \leq 0 \tag{12.42a}$$

where $X' \subseteq \mathbb{R}^n$ is a nonempty convex set, and $f : \mathbb{R}^n \to (-\infty, \infty]$ and $h : \mathbb{R}^n \to (-\infty, \infty]^l$ are proper convex extended real-valued functions. Let the Lagrangian function be

$$L(x, \mu) := f(x) + \mu^{\mathsf{T}} h(x), \qquad x \in \mathbb{R}^n, \ \mu \in \mathbb{R}^l$$

the dual function be

$$d(\mu) := \inf_{x \in X'} L(x, \mu), \qquad \mu \in \mathbb{R}^l$$

and the dual problem be

$$d^* := \sup_{\mu \geq 0} d(\mu) \tag{12.42b}$$

This problem is equivalent to (12.41) since X' can take the form $X' = P \cap C$ for a convex set C and $Ax = b$ is equivalent to $Ax \leq 0, Ax \geq 0$. For simplicity, however, we will prove the following version where the Slater condition is less refined than that in Theorem 12.27. Define the set of all dual optimal solutions μ^* that attain strong duality.

$$Q^* := \left\{ \mu^* \geq 0 : d(\mu^*) = \inf_{x \in X'} f(x) + \mu^{*\mathsf{T}} h(x) = f^* \right\} \subseteq \mathbb{R}^l \tag{12.43}$$

Due to weak duality, Q^* can be equivalently defined to be $Q^* := \{\mu^* \geq 0 : d(\mu^*) \geq f^*\}$.

Theorem 12.28 (Slater Theorem)**.** Consider the convex optimization problem and its dual (12.42). Suppose the following conditions hold:

- *Finite primal value*: $f^* > -\infty$.
- *Convexity*: f, h are proper convex functions; X' is a nonempty convex set.
- *Slater condition*: one of the following constraint qualifications holds:

CQ1: There exists $\bar{x} \in \mathrm{dom}(f) \cap X'$ such that $h(\bar{x}) < 0$;[7] or
CQ2 The functions h_i, $i = 1, \ldots, l$, are polyhedral, i.e., $h(x) = Ax + b$ for some $A \in \mathbb{R}^{l \times n}$ and $b \in \mathbb{R}^l$, and there exists $\bar{x} \in \mathrm{ri}(\mathrm{dom}(f)) \cap \mathrm{ri}(X')$ such that $A\bar{x} + b \leq 0$.

Then

1 $f^* = d^*$.
2 If CQ1 holds then Q^* in (12.43) is nonempty, convex and compact.
3 If CQ2 holds then Q^* is nonempty, convex and closed.

Due to weak duality $d^* \leq f^*$, finite f^* means that the dual problem is either finite feasible or infeasible. The constraint qualification CQ1 or CQ2 in the theorem ensures strong duality and the existence of dual optimal solutions. The proof of Theorem 12.28 illustrates the typical argument in this type of results. In particular it shows how constraint qualifications ensures that a *nonvertical* separating hyperplane exists between two disjoint convex sets. The normal vector of the hyperplane defines a dual optimal solution. The closedness of the dual optimal set Q^* is due to the property that the dual function $d(\mu)$ is concave, closed (i.e., $\mathrm{epi}(d)$ is a closed set in $\mathbb{R}^{l+1}$) and upper semicontinuous (see Lemma 12.29). If a *strictly* feasible $\bar{x}$ exists (CQ1), then Q^* is compact, not just closed (this corresponds to $0 \in \mathrm{int}\left(D_{\overline{M}}\right)$ in Lemma 12.30, not just $0 \in \mathrm{ri}\left(D_{\overline{M}}\right)$).

We next develop over Chapters 12.7.2 and 12.7.3 the proof of Theorem 12.28, adapted from [54, Chapters 4 and 5].

12.7.2 MC/MC problems

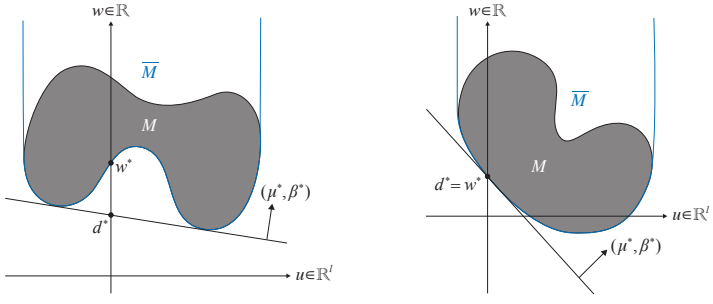
The proof of strong duality relies on the following geometric idea. Let $M \subseteq \mathbb{R}^{l+1}$ be a nonempty set and let (u, w) with $u \in \mathbb{R}^l$ and $w \in \mathbb{R}$ denote a variable in $\mathbb{R}^{l+1}$. Define the primal problem:

$$\text{Primal (minimum common)}: \qquad w^* := \inf_{(0,w) \in M} w \tag{12.44a}$$

where $w^* := \infty$ if $(0, w) \notin M$ for any $w \in \mathbb{R}$. As we will see below duality expresses the situation where there exists a nonvertical hyperplane that contains the set M in its "upper" closed halfspace; see Figure 12.15. The normal to the hyperplane defines a dual optimal solution. To describe this, recall that a hyperplane in the (u, w)-space specified by a normal $(\mu, 1) \in \mathbb{R}^{l+1}$ and an w-intercept $\xi \in \mathbb{R}$ is given by

$$\{(u, w) \in \mathbb{R}^{l+1} \ : \ \mu^{\mathsf{T}} u + w \ = \ \xi\}$$

[7] CQ1 is customarily called the Slater condition.



(a) Nonconvex $\overline{M}$ (b) Convex $\overline{M}$

Figure 12.15 The primal and dual problems (12.44) defined by the nonempty set M. Their optimal values are (w^*, d^*) respectively. The normal $(\mu^*, \beta^* := 1)$ of the nonvertical hyperplane attains the dual optimal solution μ^*, i.e., $d(\mu^*) = d^*$. (a) Nonzero duality gap $d^* < w^*$ when $\overline{M}$ is not convex. (b) Zero duality gap $d^* = w^*$ when $\overline{M}$ is convex even though M is nonconvex. In both cases, $0 \in \text{ri}(D_{\overline{M}})$ which ensures that $\beta^* > 0$ (nonvertical hyperplane).

We desire $\mu^\mathsf{T} u + w \geq \xi$ for all $(u, w) \in M$, corresponding to containing M in the "upper" halfspace. Hence define

$$d(\mu) := \inf_{(u,w)\in M} \mu^\mathsf{T} u + w$$

and the dual problem:

$$\text{Dual (maximum crossing)}: \qquad d^* := \sup_{\mu\in\mathbb{R}^l} d(\mu) \tag{12.44b}$$

Given μ, $d(\mu)$ is the smallest w-intercept of the hyperplane with normal $(\mu, 1)$ that touches (supports) the set M. The dual problem is to find a normal $(\mu^*, 1)$ such that this smallest w-intercept $d(\mu^*)$ is the maximum over $\mu \in \mathbb{R}^l$. If the normal to the hyperplane is $(\mu, 0)$, i.e., $\beta^* = 0$ in Figure 12.15, then the hyperplane is vertical and there is no finite maximum crossing d^*.

It is straightforward to show weak duality: $d^* \leq w^*$ (Exercise 12.17). The following useful property of the dual function $d(\mu)$ is derived in the proof of Lemma 12.30.

Lemma 12.29 (Dual function)**.** Consider the function $d(\mu) := \inf_{(u,w)\in M} \mu^\mathsf{T} u + w$ where $M \subseteq \mathbb{R}^{l+1}$ is nonempty. Then $d(\mu)$ is a concave, closed (i.e., epi(d) is a closed set in $\mathbb{R}^{l+1}$) and upper semicontinuous function.

It is easier to work with the positive extension $\overline{M}$ of M defined by:

$$\overline{M} \;:=\; M + \{(0, w) : w \geq 0\} \;=\; \{(u, w) \in \mathbb{R}^{l+1} : w \geq \bar{w} \text{ for some } (u, \bar{w}) \in M\} \tag{12.45}$$

because $\overline{M}$ ignores nonconvexity in the "upper" part of M which does not affect the

minimization in (12.44a). We can define (12.44) equivalently by replacing M with $\overline{M}$:

$$\text{Primal (minimum common)}: \qquad w^* := \inf_{(0,w)\in\overline{M}} w \tag{12.46a}$$

$$\text{Dual (maximum crossing)}: \qquad d^* := \sup_{\mu\in\mathbb{R}^l} d(\mu) \tag{12.46b}$$

where $d(\mu) := \inf_{(u,w)\in\overline{M}} \mu^{\mathsf{T}}u + w$.

The starting point for our proof is the following condition from [54, Propositions 4.4.1 and 4.4.2, p.150] for $d^* = w^*$ and the existence of a dual optimal solution μ^*. Let the set of all dual optimal solutions μ^* that attain strong duality be

$$Q^* = \left\{ \mu^* \in \mathbb{R}^l \,:\, d(\mu^*) := \inf_{(u,w)\in\overline{M}} \mu^{*\mathsf{T}}u + w = w^* \right\} \tag{12.47}$$

Every dual optimal $\mu^* \in Q^*$ defines a supporting hyperplane $H := \{(u,w) \in \mathbb{R}^{l+1} : \mu^{*\mathsf{T}}u + w = w^*\}$ at $(0,w^*) \in \mathrm{cl}(\overline{M})$, with $\mathrm{cl}(\overline{M})$ in the "upper" halfspace of H. See Figure 12.16.

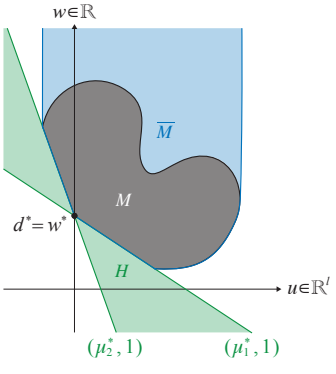


Figure 12.16 Every dual optimal $\mu^* \in Q^*$ defines a hyperplane H that passes through $(0,w^*)$ and separates it from $\mathrm{cl}(\overline{M})$. The shaded region labeled H in the figure shows all the hyperplanes defined by Q^*.

Define $D_{\overline{M}}$ to be the projection of $\overline{M}$ onto the u-space:

$$D_{\overline{M}} := \{u \in \mathbb{R}^l : (u,w) \in \overline{M} \text{ for some } w \in \mathbb{R}\} \tag{12.48}$$

We may write D for $D_{\overline{M}}$ if $\overline{M}$ is understood from the context. Then the relative interior of $\overline{M}$ and that of $D_{\overline{M}}$ are related as:

$$\mathrm{ri}(\overline{M}) = \{(u,w) \in \mathbb{R}^{l+1} \,:\, u \in \mathrm{ri}(D_{\overline{M}}),\ w > \bar{w} \text{ for some } (u,\bar{w}) \in \overline{M}\}$$

Lemma 12.30 (MC/MC strong duality)**.** Suppose

- *Finite primal value*: $w^* > -\infty$.
- *Convexity*: $\overline{M}$ is convex.

- *Constraint qualification*: $0 \in \text{ri}\left(D_{\overline{M}}\right)$.

Then

1 $d^* = w^*$ in (12.46).
2 the set Q^* in (12.47) of dual optimal solutions is nonempty, convex and closed. In particular dual optimality is attained, i.e., $d^* = d(\mu^*)$, $\mu^* \in Q^*$.
3 If $0 \in \text{int}\left(D_{\overline{M}}\right)$ then Q^* is nonempty, convex and compact. □

Note that the lemma only requires $\overline{M}$ to be convex, even if M is not. It guarantees that the dual optimal value d^* is attained at some $\mu^* \in \mathbb{R}^l$, but does not guarantee that the primal optimal value w^* is attained even though w^* is finite, i.e., $(0, w^*)$ may be in $\text{cl}(\overline{M})$ but not in $\overline{M}$. The lemma is proved by constructing a nonvertical proper separating hyperplane defined by its normal $(\mu^*, 1)$ that establishes the existence of an optimal dual vector μ^* (the hyperplane is called proper if it does not fully contain the convex set $\overline{M}$). The requirement $0 \in \text{ri}(D_{\overline{M}})$ ensures that the hyperplane is nonvertical so that the maximum crossing point is finite. The proof below that Q^* is closed also proves Lemma 12.29 on dual function $d(\mu)$. If $0 \in \text{int}\left(D_{\overline{M}}\right)$ (not just $0 \in \text{ri}\left(D_{\overline{M}}\right)$) then Q^* is compact (not just closed).

Proof We first prove parts 1 and 2 of the lemma, in five steps.

Step 1: $(0, w^*) \notin ri(\overline{M})$. We claim that w^* is finite, i.e., $-\infty < w^* < \infty$, and $(0, w^*) \notin \text{ri}(\overline{M})$. The first inequality follows from the first assumption of the lemma. The constraint qualification says that there exists $\bar{w}$ such that $(0, \bar{w}) \in \overline{M}$, and hence $w^* := \inf_{(0,w)\in\overline{M}} w \le \bar{w} < \infty$. This confirms that w^* is finite. We claim that $(0, w^*) \notin \text{ri}(\overline{M})$ because otherwise, (12.48) implies that $w^* > \bar{w}$ for some $(0, \bar{w}) \in \overline{M}$, a contradiction.

Step 2: H separating $(0, w^*)$ *from and not containing* $\overline{M}$. The Separating Hyperplane Theorem 8.10 then implies that there exists a hyperplane that passes through $(0, w^*)$ and separates $(0, w^*)$ from $\overline{M}$ (Theorem 8.10 extends easily to the case where $\text{int}(X)$ is replaced by $\text{ri}(X)$). Specifically there exists $(\mu, \beta) \in \mathbb{R}^{l+1}$ such that

$$\beta w^* \le \mu^{\mathsf{T}} u + \beta w, \qquad \forall (u, w) \in \overline{M}$$

Moreover, $(0, w^*) \notin \text{ri}(\overline{M})$ implies that the separating hyperplane $H := \{(u, w) \in \mathbb{R}^{l+1} : \mu^{\mathsf{T}} u + \beta w = \beta w^*\}$ does not fully contain the convex set $\overline{M}$ (see [54, Proposition 1.5.5, p.74]). This means that

$$\beta w^* \le \inf_{(u,w)\in\overline{M}} \mu^{\mathsf{T}} u + \beta w < \sup_{(u,w)\in\overline{M}} \mu^{\mathsf{T}} u + \beta w \tag{12.49}$$

Step 3: $\beta > 0$. We claim that $\beta > 0$. Clearly β cannot be negative because otherwise, since there exists $(0, \bar{w}) \in \overline{M}$ (constraint qualification in the lemma), the definition (12.45) of $\overline{M}$ implies that $(0, \bar{w} + w') \in \overline{M}$ as $w' \to \infty$. Hence $\inf_{(u,w)\in\overline{M}} \left(\mu^{\mathsf{T}} u + \beta w\right) \le$

$\beta(\bar{w}+w') \to -\infty$, contradicting (12.49). Suppose for the sake of contradiction that $\beta = 0$. Then (12.49) implies

$$0 \;\le\; \inf_{(u,w)\in\overline{M}} \mu^{\mathsf{T}} u \;=\; \inf_{u\in D_{\overline{M}}} \mu^{\mathsf{T}} u$$

Since $0 \in D_{\overline{M}}$ from the constraint qualification, this infimum is attained at the origin $u = 0$ over the convex set $D_{\overline{M}}$ ($D_{\overline{M}}$ is convex since it is a projection of the convex set $\overline{M}$). But $0 \in \mathrm{ri}(D_{\overline{M}})$, which is possible only if $\mu^{\mathsf{T}} u$ is constant (and equal to 0) over $D_{\overline{M}}$, for otherwise the minimum will be attained at a relative boundary point of the convex set $D_{\overline{M}}$. This contradicts the strict inequality in (12.49) with $\beta = 0$, i.e., it contradicts the fact that the separating hyperplane H does not fully contain the convex set $\overline{M}$. Hence $\beta > 0$.

Step 4: strong duality and dual optimality. Since $\beta > 0$, we can renormalize to define the hyperplane by $\mu^* := \mu/\beta$ and $\beta^* = 1$. Substitute $\beta^* = 1$ into (12.49) to get

$$w^* \;\le\; \inf_{(u,w)\in\overline{M}} \mu^{*\mathsf{T}} u + w \;=:\; d(\mu^*) \;\le\; d^*$$

where the last inequality follows from the definition (12.46b) of d^*. Weak duality $w^* \ge d^*$ then implies that $w^* = d^*$. This also shows $d(\mu^*) = d^*$, i.e., the dual optimal is attained at μ^*.

Step 5: $d(\mu)$ is concave, closed and upper semicontinuous, and Q^* is closed. For each (u,w), define the affine function $g_{u,w}(\mu) := -(\mu^{\mathsf{T}} u + w)$. Then

$$-d(\mu) \;=\; \sup_{(u,w)\in\overline{M}} g_{u,w}(\mu)$$

Hence $-d$ is convex, i.e., $\mathrm{epi}(-d)$ is a convex set in $\mathbb{R}^{l+1}$. Since $\mathrm{epi}(g_{u,w})$ is a closed set for each (u,w), $\mathrm{epi}(-d) = \cap_{(u,w)\in\overline{M}} \mathrm{epi}(g_{u,w})$ is closed in $\mathbb{R}^{l+1}$. On $\mathbb{R}^l$, $-d$ is closed if and only if $-d$ is lower semicontinuous; see Remark 12.3. Hence d is concave, closed and upper semicontinuous.

Finally Q^* is a convex set because d is a concave function and $\overline{M}$ is a convex set. Since d is upper semicontinuous on $\mathbb{R}^l$, Q^* is a closed set because, if $\{\mu_k\} \subseteq Q^*$ with $\mu_k \to \mu^* \in \mathbb{R}^l$, then $d(\mu^*) \ge \lim_k d(\mu_k) = d^*$, i.e., $\mu^* \in Q^*$. This completes the proof of parts 1 and 2 of the lemma.

For part 3, we only have to show that Q^* is bounded when $0 \in \mathrm{int}(D_{\overline{M}})$. Suppose Q^* is unbounded, i.e., there exists a sequence $\{\mu_k\} \subseteq Q^*$ such that $\|\mu_k\|_1 := \sum_i |[\mu_k]_i| \ge k$ for each integer $k > 0$. Consider the finite set $U := \{u \in \mathbb{R}^l : u_i = \pm 1\}$, i.e., U consists of 2^l vectors u whose entries u_i are 1 or -1. Since $0 \in \mathrm{int}(D_{\overline{M}})$, we can find a small enough $r > 0$ and, for each integer $k > 0$, vectors $u_k \in U$ (with $[u_k]_i = \mathrm{sign}([\mu_k]_i)$) and scalars $w_k \in \mathbb{R}$ such that $(ru_k, w_k) \in \overline{M}$ and $\mu_k^{\mathsf{T}}(ru_k) = -r\|\mu_k\|_1$. Since $\mu_k \in Q^*$ we have

$$w^* \;=\; d(\mu_k) \;\le\; \mu_k^{\mathsf{T}}(ru_k) + w_k \;=\; -r\|\mu_k\|_1 + w_k, \qquad k = 1,2,\ldots,$$

Since $\{u_k\}_k$ take values in the finite set U, there must exist an infinite subsequence

$\{(ru_{k_i}, w_{k_i})\}_i$ such that $(ru_{k_i}, w_{k_i}) = (r\hat{u}, \hat{w})$ for all i. We therefore have

$$w^* \le -r\|\mu_{k_i}\|_1 + \hat{w} \quad \text{with} \quad \|\mu_{k_i}\|_1 \ge k_i, \qquad i = 1, 2, \ldots,$$

Taking $i \to \infty$ gives $w^* = -\infty$, contradicting the assumption that $w^* > -\infty$. This shows that Q^* is bounded and hence compact. □

Lemma 12.30 applies to an arbitrary nonempty set $M \subseteq \mathbb{R}^l$. The formulation of the primal and dual problems (12.46) is very general. In the following we will use the lemma to prove Theorem 12.28 under CQ1, by specifying M in terms of the cost and constraint functions f, h. The theorem under CQ2 may not satisfy the condition $0 \in \mathrm{ri}(D_{\overline{M}})$ in the lemma, but we will modify the proof of Lemma 12.30 to prove CQ2 directly.

12.7.3 Slater Theorem 12.28: proof

We now prove Theorem 12.28. Let $X' \subseteq \mathbb{R}^n$ be a nonempty convex set and $f : \mathbb{R}^n \to (-\infty, \infty]$ and $h : \mathbb{R}^n \to (-\infty, \infty]^l$ be proper convex extended real-valued functions. Consider the convex optimization problem (12.42), reproduced here:

$$\text{Primal:} \qquad f^* := \inf_{x \in \mathbb{R}^n} f(x) \quad \text{s.t.} \quad x \in X',\ h(x) \le 0 \tag{12.50a}$$

$$\text{Dual:} \qquad d^* := \sup_{\mu \ge 0} d(\mu) \tag{12.50b}$$

where $d(\mu) := \inf_{x \in X'} L(x, \mu)$ for $\mu \in \mathbb{R}^l_+$ and $L(x, \mu) := f(x) + \mu^\mathsf{T} h(x)$, $x \in \mathbb{R}^n$, $\mu \in \mathbb{R}^l$, is the Lagrangian. We can treat the dual function $d : \mathbb{R}^l \to [-\infty, \infty]$ as an extended real-valued function defined as

$$d(\mu) := \begin{cases} \inf_{x \in X'} f(x) + \mu^\mathsf{T} h(x), & \mu \ge 0 \\ -\infty, & \text{otherwise} \end{cases} \tag{12.50c}$$

The feasible set is $X := \{x \in X' : h(x) \le 0\} \subseteq \mathbb{R}^n$.

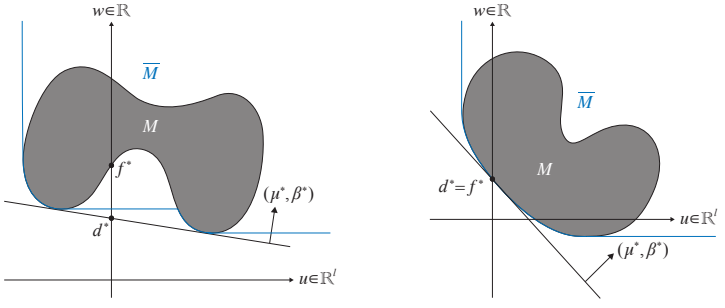
To apply Lemma 12.30 let $M := \{(h(x), f(x)) \in \mathbb{R}^{l+1} : x \in \mathrm{dom}(f) \cap X'\}$. Let its positive extension be

$$\overline{M} := \{(u, w) \in \mathbb{R}^{l+1} : u \ge h(x),\ w \ge f(x) \text{ for some } x \in \mathrm{dom}(f) \cap X'\} \tag{12.51a}$$

and the projection onto the u-space be

$$D_{\overline{M}} = \{u \in \mathbb{R}^l : u \ge h(x) \text{ for some } x \in \mathrm{dom}(f) \cap X'\} \tag{12.51b}$$

Note that since x that underlies $D_{\overline{M}}$ lies in $\mathrm{dom}(f)$, there always exists $w > f(x)$ so that $u \in \mathrm{ri}(D_{\overline{M}})$ if and only if $(u, w) \in \mathrm{ri}(\overline{M})$ for some $w > f(x)$. The extended set $\overline{M}$ defined by X' differs slightly from $\overline{M}$ in Figure 12.15 in that $u \in \mathbb{R}^l$ extends to the "right" indefinitely; see Figure 12.17. In the result below constraint qualifications imply that the primal problem (12.50a) is feasible so that $\overline{M}$ is nonempty. Indeed if $\bar{x}$



(a) Nonconvex $\overline{M}$ (b) Convex $\overline{M}$

Figure 12.17 The (u,w) space: $M := (h(X'), f(X')) := \{(h(x), f(x) : x \in \mathrm{dom}(f) \cap X'\}$ and its positive extension $\overline{M}$.

is a feasible point for (12.50a) then $(0, f(\bar{x})) \in \overline{M}$. Moreover $\overline{M}$ is convex since X' is a convex set and f, h are convex functions.

The primal and dual problems (12.46) in terms of $\overline{M}$ are then

$$\text{Primal:} \qquad f^* := \inf_{(0,w)\in\overline{M}} w \tag{12.52a}$$

$$\text{Dual:} \qquad d^* := \sup_{\mu\in\mathbb{R}^l} d(\mu) \tag{12.52b}$$

where the dual function $d(\mu) := \inf_{(u,w)\in\overline{M}} \mu^{\mathsf{T}}u + w$. The dual problem (12.52b) with the dual function $d(\mu)$ in terms of $\overline{M}$ is equivalent to the dual problem (12.50) with the dual function $d(\mu) := \inf_{x\in X'} \mu^{\mathsf{T}}h(x) + f(x)$, in the sense that μ^* is optimal for one dual problem if and only if it is optimal for the other dual problem with the same optimal value, as long as the problem is feasible (Exercise 12.18). This implies that Q^* in (12.43) is the same as Q^* in (12.47). Both are the set of all dual optimal solutions $\mu^* \geq 0$ that attain strong duality.

We first use Lemma 12.30 to prove Theorem 12.28 under CQ1, by verifying the three conditions in the lemma. Under CQ2 for an polyhedral function h, the requirement $0 \in \mathrm{ri}(D_{\overline{M}})$ in Lemma 12.30 may not hold and we will modify the proof of the lemma to prove CQ2 directly.

Proof of CQ1: $\exists \bar{x} \in dom(f) \cap X'$ *such that* $h(\bar{x}) < 0$. We verify the three conditions in Lemma 12.30, in particular $0 \in \mathrm{int}(D_{\overline{M}})$:

1 $f^* > -\infty$: This holds by assumption. Indeed CQ1 implies that f^* is finite ($-\infty < f^* < \infty$).
2 Convex $\overline{M}$: Let $(u_1, w_1), (u_2, w_2) \in \overline{M}$. Then there exist $x_1, x_2 \in \mathrm{dom}(f) \cap X'$ such that

$$u_i \geq h(x_i), \quad w_i \geq f(x_i) \qquad i = 1, 2$$

The convexity of h implies that for any $\alpha \in [0,1]$

$$\alpha u_1 + (1-\alpha)u_2 \geq \alpha h(x_1) + (1-\alpha)h(x_2) \quad \geq \quad h(\alpha x_1 + (1-\alpha)x_2)$$

Similarly the convexity of f implies $\alpha w_1 + (1-\alpha)w_2 \geq f(\alpha x_1 + (1-\alpha)x_2)$. This means $\alpha(u_1, w_1) + (1-\alpha)(u_2, w_2)$ is in $\overline{M}$, proving the convexity of $\overline{M}$.

3 $0 \in \text{int}(D_{\overline{M}})$: CQ1 gives $\bar{x} \in \text{dom}(f) \cap X'$ with $h(\bar{x}) < 0$. Therefore $0 \in D_{\overline{M}}$, where $D_{\overline{M}}$ is defined in (12.51b). Moreover there is an $\epsilon > 0$ such that $u \in D_{\overline{M}}$ for any u with $\|u\| \leq \epsilon$ and a $w > f(\bar{x})$ such that $(u, w) \in \text{ri}(\overline{M})$. This implies that $0 \in \text{int}\left(D_{\overline{M}}\right)$.

Lemma 12.30 then implies that

$$d^* \;=\; f^*, \qquad \exists \mu^* \in \mathbb{R}^m \text{ s.t. } d^* \;=\; d(\mu^*) \;=\; \inf_{(u,w)\in\overline{M}} \mu^{*\mathsf{T}} u + w \tag{12.53}$$

Moreover the set Q^* of dual optimal solutions is convex and compact. This completes the proof of Theorem 12.28 under CQ1. □

Proof of CQ2: $\exists \bar{x} \in ri(dom(f)) \cap ri(X')$ *such that* $h(\bar{x}) := A\bar{x} + b \leq 0$. In this case, the condition $0 \in \text{ri}\left(D_{\overline{M}}\right) := \text{ri}\left(\{u : u \geq Ax + b \text{ for some } x \in \text{dom}(f) \cap X'\}\right)$ in Lemma 12.30 may not hold, but we will modify the 5 steps in the proof of Lemma 12.30 to establish (12.53) and properties of Q^* directly (the key difference being Step 2).

Step 1: $f^* > -\infty$. This holds by assumption. Indeed CQ2 implies that f^* is finite ($-\infty < f^* < \infty$).

Step 2: Separating hyperplane. Substitute $h(x) = Ax - b$ into the definition (12.51a) of $\overline{M}$:

$$\overline{M} \;:=\; \{(u,w) \in \mathbb{R}^{l+1} : u \geq Ax - b,\; w \geq f(x) \text{ for some } x \in \text{dom}(f) \cap X'\}$$

The key to the proof is a clever decomposition of $\overline{M}$ as a Minkowski sum of a convex set $C \subseteq \mathbb{R}^{l+1}$ defined by the convex function f and a polyhedral set $P \subseteq \mathbb{R}^{l+1}$ defined by the affine functions h, as follows. With the view of a slack variable $v := u - (Ax - b) \geq 0$, we can write $\overline{M} = C + P$ where

$$C \;:=\; \{(Ax - b, w) : w \geq f(x) \text{ for some } x \in \text{dom}(f) \cap X'\}, \quad P \;:=\; \{(v, 0) : v \geq 0\}$$

$\overline{M} = C + P$ because $(u, w) \in \overline{M}$ if and only if $u = Ax - b + v$ for some $v \geq 0$ and $w \geq f(x)$.

Guided by the sets C and P (see Step 4 below), we define the convex set $\tilde{C} \subseteq \mathbb{R}^{l+1}$ and the polyhedral set $\tilde{P} \subseteq \mathbb{R}^{l+1}$ (since f^* is finite):

$$\tilde{C} \;:=\; \{(Ax - b, w) \;:\; w > f(x) \;\text{ for some } x \in \text{dom}(f) \cap X'\}, \quad \tilde{P} \;:=\; \{(v, f^*) \;:\; v \leq 0\}$$

(When X' is open, $\tilde{C} = \text{ri}(C)$. More generally, when restricted to $x \in \text{ri}(X')$, $\hat{C} := \{(Ax - b, w) : w > f(x) \text{ for some } x \in \text{ri}(X')\}$ is $\text{ri}(C)$.) We claim that $\tilde{C} \cap \tilde{P} = \emptyset$ because otherwise if $(\tilde{v}, f^*) \in \tilde{C} \cap \tilde{P}$ then there exists an $\tilde{x} \in X'$ such that

$$\tilde{v} \;=\; A\tilde{x} - b \;\leq\; 0, \qquad\qquad f^* > f(\tilde{x})$$

contradicting that f is uniformly lower bounded by f^* on its feasible set.

The separating hyperplane Theorem 8.11 then implies that there exists a hyperplane that separates $\tilde{C}$ and $\tilde{P}$, i.e., $\exists(\mu,\beta) \in \mathbb{R}^{l+1}$ such that

$$\sup_{(v,f^*)\in\tilde{P}} \mu^\mathsf{T} v + \beta f^* \;\le\; \inf_{(u,w)\in\tilde{C}} \mu^\mathsf{T} u + \beta w$$

Moreover the separating hyperplane does not fully contain the convex set $\tilde{C}$ (follows from [54, Proposition 1.5.7, p.77] since $\mathrm{ri}(\tilde{C}) \cap \tilde{P} = \emptyset$). This means that

$$\sup_{(v,f^*)\in\tilde{P}} \mu^\mathsf{T} v + \beta f^* \;\le\; \inf_{(u,w)\in\tilde{C}} \mu^\mathsf{T} u + \beta w \;<\; \sup_{(u,w)\in\tilde{C}} \mu^\mathsf{T} u + \beta w \tag{12.54}$$

This corresponds to (12.49) in the proof of Lemma 12.30. The remaining Steps 3 and 4 follow the same idea there, working with $\tilde{C}$, $\tilde{P}$ and the decomposition of $\overline{M} = C + P$ here instead of $\overline{M}$ directly in Lemma 12.30.

Step 3: $\beta > 0$. We claim that $\beta > 0$. Clearly β cannot be negative because otherwise, since $(0, f(\bar{x})) \in \overline{M}$ (where $\bar{x}$ is the point in CQ2), the definition (12.51a) of $\overline{M}$ implies that $(0, f(\bar{x}) + w') \in \overline{M}$ as $w' \to \infty$. Hence $\inf_{(u,w)\in\overline{M}} \left(\mu^\mathsf{T} u + \beta w\right) \le \beta(f(\bar{x}) + w') \to -\infty$, contradicting (12.54). Suppose for the sake of contradiction that $\beta = 0$. Then (12.54) implies

$$\sup_{(v,f^*)\in\tilde{P}} \mu^\mathsf{T} v \;\le\; \inf_{(u,w)\in\tilde{C}} \mu^\mathsf{T} u \;\le\; \mu^\mathsf{T} \bar{v}$$

where $\bar{v} := A\bar{x} - b$ with $\bar{x}$ being the point in CQ2. Here the last inequality follows because the point $(\bar{v}, f(\bar{x}))$ is in $\tilde{C}$. But $\bar{v} \le 0$ and hence $(\bar{v}, f^*) \in \tilde{P}$. Therefore

$$\mu^\mathsf{T} \bar{v} \;\le\; \sup_{(v,f^*)\in\tilde{P}} \mu^\mathsf{T} v \;\le\; \inf_{(u,w)\in\tilde{C}} \mu^\mathsf{T} u \;\le\; \mu^\mathsf{T} \bar{v}$$

i.e., all inequalities above must hold with equality. Therefore $\bar{v} := A\bar{x} - b$ attains the minimization of $\mu^\mathsf{T} u$ over the projection $\tilde{D} := \{u = Ax - b : (u,w) \in \tilde{C}\}$ of $\tilde{C}$ onto the u-space. Since CQ2 says that $\bar{x} \in \mathrm{ri}(\mathrm{dom}(f)) \cap \mathrm{ri}(X')$, $\bar{v} := A\bar{x} - b$ is in $\mathrm{ri}(\tilde{D})$. This is possible only if $\mu^\mathsf{T} u$ is constant (and equal to $\mu^\mathsf{T} \bar{v}$) over $\tilde{D}$, for otherwise the infimum will be attained at a relative boundary point of the convex set $\tilde{D}$. This contradicts the strict inequality in (12.54), i.e., it contradicts the fact that the separating hyperplane does not fully contain the convex set $\tilde{C}$.

Step 4: strong duality and dual optimality. Since $\beta > 0$, we can renormalize to define the hyperplane by $\mu^* := \mu/\beta$ and $\beta^* = 1$. Substitute $\beta^* = 1$ into (12.54) to get

$$\begin{aligned}
\sup_{v\le 0} \mu^{*\mathsf{T}} v + f^* \;&\le\; \inf_{(u,w)\in\tilde{C}} \mu^{*\mathsf{T}} u + w \\
f^* \;&\le\; \inf_{(u,w)\in\tilde{C}} \inf_{v\le 0} \mu^{*\mathsf{T}}(u - v) \;+\; w \\
&=\; \inf_{(u,w)\in C} \inf_{(v,0)\in P} \mu^{*\mathsf{T}}(u + v) \;+\; w \\
&=\; \inf_{(u,w)\overline{M}} \mu^{*\mathsf{T}} u + w \;=:\; d(\mu^*) \;\le\; d^*
\end{aligned}$$

where the first equality uses the fact that the infimum of $\mu^{*\mathsf{T}}u + w$ over $\tilde{C}$ or C is the same. Weak duality $f^* \geq d^*$ then implies that $f^* = d^*$. This also shows $d(\mu^*) = d^*$, i.e., the dual optimal is attained at μ^*. This establishes (12.53), i.e., the set Q^* of dual optimal solutions is nonempty.

Step 5: Q^* is convex and closed. This step is the same as Step 5 in the proof of Lemma 12.30. As shown there, the dual function $d(\mu)$ is concave, closed and upper semicontinuous. Therefore the set Q^* of dual optimal solutions is a convex set ($\overline{M}$ is convex as shown above for the case CQ1). Since d is upper semicontinuous on $\mathbb{R}^n$, Q^* is a closed set because, if $\{\mu_k\} \subseteq Q^*$ with $\mu_k \to \mu^* \in \mathbb{R}^n$, then $d(\mu^*) \geq \lim_k d(\mu_k) = d^*$, i.e., $\mu^* \in Q^*$. □

12.8 Special convex programs

In this section we apply the general theory developed in Chapters 12.4–12.7 to special classes convex optimization problems widely used in applications. In particular we apply the Slater Theorem 12.27 and the generalized KKT Theorem 12.21 to derive conditions for strong duality, dual optimality and the KKT conditions for some of the problem classes in Figure 8.14 of Chapter 8.4.1 (specifically linear program, second-order cone program, conic program, and convex program specified by a general convex inequality). It extends some of the results of Chapter 8.4 for differentiable problems to a nonsmooth setting.

12.8.1 Summary: general method

Consider the convex problem:

$$f^* := \min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t.} \quad Ax = b,\ x \in X \subseteq \mathbb{R}^n \tag{12.55}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function, $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ and X is a nonempty closed convex set that may be specified explicitly as $h(x) \leq 0$ for a convex function $h : \mathbb{R}^n \to \mathbb{R}^l$. The problems studied in this section is summarized in Figure 8.14 and the conclusions are summarize in Table 8.3 of Chapter 8.4.1. A general method to derive these conclusions is also described in Chapter 8.4.1 for smooth problems. Here we summarize how to adapt that method to the nonsmooth setting using concepts of subgradients, normal cones and dual cones. The key difference is the approach to derive the KKT condition without differentiability and for abstract specifications of the feasible set X.

1 *Dual problem.* Given the primal problem (12.55), if X is explicitly specified, e.g., by a convex inequality $h(x) \leq 0$, then the Lagrangian function L and the dual problem are defined by (8.55a) (8.55b) in Chapter 8.4.1. Otherwise if $X \subseteq \mathbb{R}^n$ is

specified by $Bx+d \in K$ for a closed convex cone $K \subseteq \mathbb{R}^l$ then the Lagrangian can be defined in terms of its dual cone K^*:

$$L(x,\lambda,\mu) \;:=\; f(x) - \lambda^\mathsf{T}(Ax-b) + \mu(Bx+d), \qquad x \in \mathbb{R}^n,\ \lambda \in \mathbb{R}^m,\ \mu \in K^* \subseteq \mathbb{R}^l$$

The dual function is $d(\lambda,\mu) := \min_{x\in\mathbb{R}^n} L(x,\lambda,\mu)$ and the dual problem is

$$d^* \;:=\; \max_{(\lambda,\mu)\in\mathbb{R}^{m+l}} d(\lambda,\mu) \qquad \text{s.t.} \qquad \mu \in K^*$$

This is derived in Chapter 12.8.4.

2 *Strong duality and dual optimality*. This does not require differentiability and the results hold almost verbatim in the nonsmooth setting using Theorem 12.27 (except substituting subgradients for gradients).

3 *KKT condition and primal optimality*. Suppose $X \subseteq \mathbb{R}^n$ is specified by $Bx+d \in K$ for a closed convex cone $K \subseteq \mathbb{R}^l$. Without differentiability the KKT condition cannot be derived simply from $\nabla_x L(x^*,\lambda^*,\mu^*) = 0$ as done in (8.55c) of Chapter 8.4.1. Instead we convert (12.55) into an unconstraint problem

$$f^* \;:=\; \min_{x\in\mathbb{R}^n} \; f(x) + \delta_H(x) + \delta_K(Bx+d)$$

where $H := \{x \in \mathbb{R}^n : Ax = b\}$. Recall that (i) f is a convex function. Suppose (ii) the Slater condition is satisfied, i.e., there exists $\bar{x} \in \mathrm{ri}(\mathrm{dom}(f)) \cap \mathrm{ri}(K)$ with $A\bar{x} = b$ ($\mathrm{dom}(f) = \mathbb{R}^n$ if we assume f is real-valued). Then the generalized KKT Theorem 12.21 implies that x^* is optimal if and only if there exists a subgradient $\xi^* \in \partial f(x^*)$, $\lambda^* \in \mathbb{R}^m$ and $\mu^* \in \mathbb{R}^l$ such that (from Corollary 12.9):

$$\xi^* \;\in\; -N_H(x^*) - B^\mathsf{T} N_K(Bx^*+d)$$

Using Theorem 12.3 on normal cones the KKT condition is equivalent to

$$\xi^* \;=\; A^\mathsf{T}\lambda^* + B^\mathsf{T}\mu^*, \qquad \mu^{*\mathsf{T}}(Bx^*+d) \;=\; 0, \qquad \mu^* \in K^*$$

Indeed the conditions $\mu^* \in K^*$ and $\mu^{*\mathsf{T}}(Bx^*+d) = 0$ define a vector μ^* in $-N_K(Bx^*+d)$ according to Theorem 12.3 for a general convex cone K. When K is specified explicitly, e.g., K is the second-order cone, these conditions define the vector μ^* more specifically based on the primal optimal x^*.

In the rest of this section we apply this general method to common convex programs. The results are summarized in Table 8.3.

12.8.2 Linear program (LP)

Consider the linear program:

$$f^* \;:=\; \min_{x\in\mathbb{R}^n} \; c^\mathsf{T}x \qquad \text{s.t.} \qquad Ax = b,\ x \geq 0 \tag{12.56a}$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ and $c \in \mathbb{R}^n$. Let $H := \{x \in \mathbb{R}^n : Ax = b\}$ and $K := \{x \in \mathbb{R}^n : x \geq 0\}$. Theorem 8.23 and Example 8.13 in Chapter 8.4.2 show that if either the

optimal primal or the optimal dual value is finite then both primal and dual optimality is attained, strong duality holds, and a primal and dual feasible solution is optimal if and only if it satisfies complementary slackness. In this subsection we derive the same result using Theorem 12.21 to illustrate the simplicity of the set-theoretic approach in the nonsmooth setting.

For strong duality and the existence of primal and dual optimal solutions, the dual problem of (12.56a) is derived in Example 8.13 to be:

$$d^* := \max_{\lambda,\mu\geq 0} b^\mathsf{T}\mu \qquad \text{s.t.} \qquad A^\mathsf{T}\lambda+\mu = c \tag{12.56b}$$

where $\lambda \in \mathbb{R}^m$, $\mu \in \mathbb{R}^n$. Let $X := \{x \in \mathbb{R}^n : Ax = b, x \geq 0\}$ and $Y := \{(\lambda,\mu) \in \mathbb{R}^{m+n} : A^\mathsf{T}\lambda+\mu = c, \mu \geq 0\}$ be the feasible sets. If either f^* or d^* is finite then the Slater condition of Theorem 12.27 (or Slater Theorem 8.17) is satisfied. The exact same proof for part 1 of Theorem 8.23 shows that there exists a primal-dual optimal solution $(x^*,\lambda^*,\mu^*) \in X \times Y$ that closes the duality gap, i.e.,

$$c^\mathsf{T}x^* \;=\; f^* \;=\; d^* \;=\; b^\mathsf{T}\lambda^*$$

For KKT characterization, rewrite (12.56a) as an unconstrained optimization of an extended real-valued function:

$$\min_{x\in\mathbb{R}^n} \; c^\mathsf{T}x \;+\; \delta_H(x) \;+\; \delta_K(x) \tag{12.56c}$$

Since the objective function $f(x) := c^\mathsf{T}x$ is real-valued and polyhedral, $\mathrm{dom}(f) = \mathbb{R}^n$. Application of Theorem 12.21 then says that $x^* \in \mathbb{R}^n$ is optimal if and only if

$$-c \;\in\; \partial\left(\delta_H(x^*)+\delta_K(x^*)\right) \;=\; \partial\delta_H(x^*)+\partial\delta_K(x^*)$$

where the equality follows from Theorem 12.18, provided (12.56) is feasible ($H \cap K \neq \emptyset$). Since $\partial\delta_X(x) = N_X(x)$ from Table 12.2, x^* is optimal if and only if

$$-c \;\in\; N_H(x^*)+N_K(x^*)$$

From Theorem 12.3 in Chapter 12.1.3,

$$\begin{aligned} N_H(x^*) \;&=\; \{A^\mathsf{T}\lambda \in \mathbb{R}^n : \lambda \in \mathbb{R}^m\} \\ N_K(x^*) \;&=\; \{y \in \mathbb{R}^n : y \leq 0,\; y^\mathsf{T}x^* = 0\} \end{aligned}$$

Substituting these normal cones into the condition $c \in -N_H(x^*) - N_K(x^*)$ leads to KKT condition for linear program: a feasible x^* is optimal if and only if there exists a $(\lambda^*,\mu^*) \in \mathbb{R}^{m+n}$ such that

$$A^\mathsf{T}\lambda^*+\mu^* \;=\; c, \qquad \mu^{*\mathsf{T}}x^* \;=\; 0, \qquad \mu^* \;\geq\; 0 \tag{12.57}$$

Such a point (x^*,λ^*,μ^*) is a saddle point and a KKT point and is hence primal-dual optimal with $c^\mathsf{T}x^* = b^\mathsf{T}\mu^*$. Since the constraint qualification in Theorem 12.21 reduces to feasibility for a linear program, the KKT characterization (12.57) requires only feasibility of the linear program (12.56). Strong duality and the existence of primal and dual optimal solutions requires, in addition, $f^* > -\infty$ (or $-\infty < d^* < \infty$).

12.8.3 Second-order cone program (SOCP)

Second-order cone.

Recall the second-order cone program (SOCP):

$$f^* \;:=\; \min_{x\in\mathbb{R}^n} f(x) \;\text{ s.t. }\; Ax=b,\; x\in K_{\text{soc}} \tag{12.58a}$$

where $f:\mathbb{R}^n\to\mathbb{R}$ is a real-valued convex function (not necessarily differentiable), $A\in\mathbb{R}^{m\times n}$, $b\in\mathbb{R}^m$, and $K_{\text{soc}}\subseteq\mathbb{R}^n$ is the standard second-order cone defined in (8.16), reproduced here ($x^k := (x_1,\cdots,x_k)$ denotes the vector consisting of the first k entries of x),

$$K_{\text{soc}} \;:=\; \{x\in\mathbb{R}^n : \|x^{n-1}\|_2 \;\le\; x_n\} \tag{12.58b}$$

and studied in Theorem 12.10. The Lagrangian $L:\mathbb{R}^{n+m+1}\to\mathbb{R}$ of (12.58a)(12.58b) is

$$L(x,\lambda,\mu) \;:=\; f(x)-\lambda^{\mathsf{T}}(Ax-b)+\mu\left(\|x^{n-1}\|_2-x_n\right), \qquad x\in\mathbb{R}^n,\; \lambda\in\mathbb{R}^m, \mu\in\mathbb{R}$$

the dual function is $d(\lambda,\mu) := \min_{x\in\mathbb{R}^n} L(x,\lambda,\mu)$ and the dual problem is

$$d^* \;:=\; \max_{\lambda,\mu\ge 0} d(\lambda,\mu) \tag{12.58c}$$

We now show that Theorem 8.26 on strong duality, dual optimality and the KKT condition for SOCP in Chapter 8.4.4 for smooth convex optimization holds almost verbatim in the nonsmoonth setting, except that Theorem 8.26 only covers the case where $[x^*]^{n-1}\neq 0$ so that the constraint function $h(x) := \|x^{n-1}\|_2 - x_n$ is differentiable whereas the derivation below covers the case where $[x^*]^{n-1}=0$ as well.

Indeed, strong duality and dual optimality follow from the Slater Theorem 12.27. To derive the KKT condition, we rewrite the primal problem of SOCP (12.58) as an unconstrained optimization of an extended real-valued function. It illustrates both how nonsmooth analysis handles points of nondifferentiability and the simplicity of the set-theoretic approach here. Specifically rewrite (12.58a)(12.58b) as:

$$\min_{x\in\mathbb{R}^n} \; f(x) \;+\; \delta_H(x) \;+\; \delta_K(x)$$

where $H := \{x\in\mathbb{R}^n : Ax=b\}$ where $K := K_{\text{soc}}$ is the second-order cone. Since f is real-valued, $\text{ri}(\text{dom}(f))=\mathbb{R}^n$ and hence the constraint qualifications in Theorem 12.21 reduces to the Slater condition $H\cap\text{ri}(K)\neq\emptyset$ (Remark 12.8). Under this condition Theorem 12.21 says that $x^*\in H\cap K$ is optimal if and only if there exists a $\xi^*\in\partial f(x^*)$ such that

$$-\xi^* \;\in\; \partial\left(\delta_H(x^*)+\delta_K(x^*)\right) \;=\; \partial\delta_H(x^*)+\partial\delta_K(x^*) \;=\; N_H(x^*)+N_K(x^*) \tag{12.59}$$

(The first equality follows from Theorem 12.18 under the Slater condition $H\cap\text{ri}(K)\neq\emptyset$ and the second equality follows from $\partial\delta_X(x)=N_X(x)$ in Table 12.2.) Theorems 12.3

and 12.10 in Chapter 12.1 then give

$$N_H(x^*) = \{A^\mathsf{T}\lambda \in \mathbb{R}^n : \lambda \in \mathbb{R}^m\}$$

$$N_K(x^*) = \begin{cases} \{\xi \in \mathbb{R}^n : \|\xi^{n-1}\|_2 \le -\xi_n\} & \text{if } x^* = 0 \\ \{0 \in \mathbb{R}^n\} & \text{if } \|[x^*]^{n-1}\|_2 < x_n^* \\ \{\mu([x^*]^{n-1}, -x_n^*) \in \mathbb{R}^n : \mu \ge 0\} & \text{if } \|[x^*]^{n-1}\|_2 = x_n^* > 0 \end{cases} \tag{12.60a}$$

Substituting these normal cones into (12.59) leads to the following KKT condition. Suppose the SOCP (12.58) satisfies the Slater condition $H \cap \mathrm{ri}(K) \neq \emptyset$. We separate three cases according to N_K in (12.60a): A feasible $x^* \in H \cap K$ is optimal if and only if there exist $\xi^* \in \partial f(x^*)$, $\lambda^* \in \mathbb{R}^m$ and

1 Case $x_n^* > \|[x^*]^{n-1}\|_2 \ge 0$: such that

$$\xi^* = A^\mathsf{T}\lambda^* \tag{12.61a}$$

which is the same as the KKT condition in Theorem 8.26. This includes the case not covered in Theorem 8.26 in which $[x^*]^{n-1} = 0$ where the constraint function $h(x) := \|x^{n-1}\|_2 - x_n$ is nondifferentiable.

2 Case $x_n^* = \|[x^*]^{n-1}\|_2 > 0$: there exists $\mu^* \in \mathbb{R}_+$ such that

$$\xi^* = A^\mathsf{T}\lambda^* + \mu^* \begin{bmatrix} -[x^*]^{n-1} \\ x_n^* \end{bmatrix} \tag{12.61b}$$

which is the same as the KKT condition in Theorem 8.26. Note that $\mu^*(-[x^*]^{n-1}, x_n^*)$ is a vector in K_{soc} as in the next case.

3 Case $x_n^* = \|[x^*]^{n-1}\|_2 = 0$: there exists $\tilde{\eta} \in K_{\mathrm{soc}}^\circ := \{\eta \in \mathbb{R}^n : \|\eta^{n-1}\|_2 \le -\eta_n\}$ such that $-\xi^* = A^\mathsf{T}(-\lambda^*) + \tilde{\eta}$. This is equivalent to: $x^* = 0$ is optimal if and only if there exist $\xi^* \in \partial f(0)$, $\lambda^* \in \mathbb{R}^m$ and $\eta^* \in K_{\mathrm{soc}}$ such that

$$\xi^* = A^\mathsf{T}\lambda^* + \eta^* \tag{12.61c}$$

Note that $b = Ax^* = 0$. As in case 1, the constraint function $h(x) := \|x^{n-1}\|_2 - x_n$ is nondifferentiable at $x^* = 0$, the case not covered in Theorem 8.26.

Here we assume the Slater condition and the conclusion is slightly stronger than that in Theorem 8.26 (see Remark 8.11).

Remark 12.10 ($\eta^* \in K_{\mathrm{soc}}$ for SOCP)**.** Note that all the KKT conditions in (12.61) are of the form $\xi^* = A^\mathsf{T}\lambda + \eta^*$ for some $\eta^* \in K_{\mathrm{soc}}$. This is due to (12.59) that requires $\xi^* \in -N_H(x^*) - N_K(x^*)$ and Theorem 12.3 that says that $N_K(x^*) \subseteq K_{\mathrm{soc}}^\circ$, the polar cone of K_{soc}. Hence η^* is in the dual cone $K_{\mathrm{soc}}^* = -K_{\mathrm{soc}}^\circ = K_{\mathrm{soc}}$ since the second-order cone is self-dual. Indeed the conditions in (12.61) specialize the description $\eta^* \in K^*$ and $\eta^{*\mathsf{T}}x^* = 0$ in Theorem 12.3 for a general convex cone K to the case of second-order cone based on x^*. □

SOC constraint.

Recall the second-order cone program (SOCP):

$$f^* \;:=\; \min_{x\in\mathbb{R}^n} f(x) \;\text{ s.t. }\; Ax=b,\; \|Bx+d\|_2 \;\le\; \beta^{\mathsf{T}}x+\delta \tag{12.62}$$

where $f:\mathbb{R}^n\to\mathbb{R}$ is a real-valued convex function (not necessarily differentiable), $A\in\mathbb{R}^{m\times n}$ and $b\in\mathbb{R}^m$, $B\in\mathbb{R}^{(l-1)\times n}$, $d\in\mathbb{R}^{l-1}$, $\beta\in\mathbb{R}^n$ and $\delta\in\mathbb{R}$. The constraint $\|Bx+d\|_2\le\beta^{\mathsf{T}}x+\delta$ is the second-order cone constraint studied in Chapter 8.2.1. It is a convex constraint but does not necessarily defines a cone. We now show that Theorem 8.27 in Chapter 8.4.4 on strong duality, dual optimality and the KKT condition holds almost verbatim in the nonsmoonth setting, except that Theorem 8.27 only covers the case where $Bx^*+d\neq 0$ so that the constraint function $h(x):=\|Bx+d\|_2-(\beta^{\mathsf{T}}x+\delta)$ is differentiable whereas the derivation below allows $\|Bx^*+d\|_2=0$.

As for the SOCP (12.58), strong duality and dual optimality follow from the Slater Theorem 12.27. To derive the KKT condition in Theorem 8.27, we will use Theorem 12.21 to handle points of nondifferentiability. First we reduce the SOC constraint in (12.62) to the conic constraint in (12.58) with an auxiliary variables z and an additional linear equality constraint:

$$z^{l-1} \;=\; Bx+d, \qquad\qquad z_l \;=\; \beta^{\mathsf{T}}x+\delta, \qquad\qquad \|z^{l-1}\|_2 \;\le\; z_l$$

Then we rewrite SOCP (12.62) as an unconstrained optimization: let

$$\tilde{B} \;:=\; \begin{bmatrix} B \\ \beta^{\mathsf{T}} \end{bmatrix}, \qquad\qquad \tilde{d} \;:=\; \begin{bmatrix} d \\ \delta \end{bmatrix}$$

and

$$\begin{aligned}
\tilde{H}_1 \;&:=\; \{(x,z)\in\mathbb{R}^{n+l} : Ax=b\} \;=:\; H_1\times\mathbb{R}^l, && H_1 \;:=\; \{x\in\mathbb{R}^n : Ax=b\}\\
\tilde{K} \;&:=\; \{(x,z)\in\mathbb{R}^{n+l} : \|z^{l-1}\|_2 \;\le\; z_l\} \;=:\; \mathbb{R}^n\times K, && K \;:=\; \{z\in\mathbb{R}^l : \|z^{l-1}\|_2 \;\le\; z_l\}\\
H_2 \;&:=\; \{(x,z)\in\mathbb{R}^{n+l} : z=\tilde{B}x+\tilde{d}\}
\end{aligned}$$

with normal cones $N_{\tilde{H}_1}(x,z)=N_{H_1}(x)\times\{0\in\mathbb{R}^l\}$ and $N_{\tilde{K}}(x,z)=\{0\in\mathbb{R}^n\}\times N_K(z)$. Rewrite SOCP (12.62) as:

$$\min_{(x,z)\in\mathbb{R}^{n+l}} \; f(x) \;+\; \delta_{\tilde{H}_1}(x,z) \;+\; \delta_{\tilde{K}}(x,z) \;+\; \delta_{H_2}(x,z)$$

The constraint qualification in Theorem 12.21 reduces to the Slater condition $\tilde{H}_1\cap \mathrm{ri}(\tilde{K})\cap H_2\neq\emptyset$ (Remark 12.8). Under this condition Theorem 12.21 says that $(x^*,z^*)\in \tilde{H}_1\cap\tilde{K}\cap H_2$ is optimal if and only if there exists a $\xi^*\in\partial f(x^*)$ such that

$$-\begin{bmatrix} \xi^* \\ 0 \end{bmatrix} \;\in\; N_{\tilde{H}_1}(x^*,z^*)+N_{\tilde{K}}(x^*,z^*)+N_{H_2}(x^*,z^*) \;=\; \begin{bmatrix} N_{H_1}(x^*) \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ N_K(z^*) \end{bmatrix} + N_{H_2}(x^*,z^*) \tag{12.63}$$

Theorems 12.3 and 12.10 in Chapter 12.1 give

$$N_{H_1}(x^*) = \{A^\mathsf{T}\lambda \in \mathbb{R}^n : \lambda \in \mathbb{R}^m\}$$

$$N_K(z^*) = \begin{cases} \{\eta \in \mathbb{R}^l : \|\eta^{l-1}\|_2 \le -\eta_l\} & \text{if } z^* = 0 \\ \{0 \in \mathbb{R}^l\} & \text{if } \|[z^*]^{l-1}\|_2 < z_l^* \\ \{\mu([z^*]^{l-1}, -z_l^*) \in \mathbb{R}^l : \mu \ge 0\} & \text{if } \|[z^*]^{l-1}\|_2 = z_l^* > 0 \end{cases}$$

Now $N_{H_2}(x^*, z^*) = \{(\xi, \eta) \in \mathbb{R}^{n+l} : \xi = \tilde{B}^\mathsf{T}\gamma, \eta = -\gamma, \gamma \in \mathbb{R}^l\}$ and hence

$$N_{H_2}(x^*, z^*) = \{(\tilde{B}^\mathsf{T}\gamma, -\gamma) \in \mathbb{R}^{n+l} : \gamma \in \mathbb{R}^l\}$$

Substituting these normal cones into (12.63) leads to the following KKT condition. Suppose the SOCP (12.62) satisfies the Slater condition that there exists $\bar{x}$ such that $A\bar{x} = b$ and $\|B\bar{x} + d\|_2 < \beta^\mathsf{T}\bar{x} + \delta$. We separate three cases according to N_K: A feasible x^* is optimal if and only if there exists $\xi^* \in \partial f(x^*)$, $\lambda^* \in \mathbb{R}^m$, and

1 Case $\beta^\mathsf{T}x^* + \delta > \|Bx^* + d\|_2 \ge 0$: such that ($\gamma^* = 0$ in this case)

$$\xi^* = A^\mathsf{T}\lambda^* \tag{12.64a}$$

which is the same as the KKT condition in Theorem 8.27. This includes the case not covered in Theorem 8.27 in which $Bx^* + d = 0$ where the constraint function $h(x) := \|Bx + d\|_2 - (\beta^\mathsf{T}x^* + \delta)$ is nondifferentiable.

2 Case $\beta^\mathsf{T}x^* + \delta = \|Bx^* + d\|_2 > 0$: there exist $\gamma^* \in \mathbb{R}^l$ and $\mu^* \in \mathbb{R}_+$ such that $-\xi^* = A^\mathsf{T}\lambda^* + \tilde{B}^\mathsf{T}\gamma^*$ and $\gamma^* = \mu^*([z^*]^{k-1}, -z_k^*)$ where $z^* = \tilde{B}x^* + \tilde{d}$. Eliminating γ^* and z^* yields: A feasible x^* is optimal if and only if there exists $\xi^* \in \partial f(x^*)$, $\lambda^* \in \mathbb{R}^m$ and $\mu^* \in \mathbb{R}_+$ such that

$$\xi^* = A^\mathsf{T}\lambda^* + \mu^*\left(-B^\mathsf{T}(Bx^* + d) + \beta(\beta^\mathsf{T}x^* + \delta)\right) \tag{12.64b}$$

This is the same as the KKT condition in Theorem 8.27.

3 Case $\beta^\mathsf{T}x^* + \delta = \|Bx^* + d\|_2 = 0$: there exist $\gamma^* \in \mathbb{R}^l$ and $\tilde{\eta} \in K^\circ := \{\tilde{\eta} \in \mathbb{R}^l : \|\tilde{\eta}^{l-1}\|_2 \le -\tilde{\eta}_l\}$ such that $-\xi^* = A^\mathsf{T}(-\lambda^*) + \tilde{B}^\mathsf{T}\gamma^*$ and $\gamma^* = \tilde{\eta}$. Eliminating γ^* yields: x^* with $0 = \|Bx^* + d\|_2 = \beta^\mathsf{T}x^* + \delta$ is optimal if and only there exist $\xi^* \in \partial f(x^*)$, $\lambda^* \in \mathbb{R}^m$ and $\eta^* \in K$ such that

$$\xi^* = A^\mathsf{T}\lambda^* + \tilde{B}^\mathsf{T}\eta^* \tag{12.64c}$$

As in case 1, the constraint function $h(x) := \|Bx + d\|_2 - (\beta^\mathsf{T}x^* + \delta)$ is nondifferentiable at x^* where $0 = \|Bx^* + d\|_2$, the case not covered in Theorem 8.27.

12.8.4 Conic program and convex inequality

In this subsection we derive conditions for strong duality and dual optimality and the KKT condition for conic programs and for convex programs specified by a general convex inequality.

Conic feasible set.

A generalization of SOCP (12.58) is the following convex optimization

$$f^* \;:=\; \min_{x\in\mathbb{R}^n} f(x) \;\text{ s.t. }\; Ax=b,\; x\in K \tag{12.65}$$

where $f:\mathbb{R}^n\to\mathbb{R}$ is a real-valued convex function, $A\in\mathbb{R}^{m\times n}$, $b\in\mathbb{R}^m$, and $K\subseteq\mathbb{R}^n$ is a closed convex cone. Even though $K\subseteq\mathbb{R}^n$ in (12.65) is not explicitly specified by convex inequalities, but because K is a convex cone, we can formulate the Lagrangian dual problem using the dual cone of K. Recall the polar cone K° and the dual cone K^* of K in Definition 12.1:

$$K^\circ \;:=\; \{\xi\in\mathbb{R}^n : \xi^{\mathsf{T}}x\le 0 \;\; \forall x\in K\} \tag{12.66a}$$

$$K^* \;:=\; -K^\circ \;:=\; \{\xi\in\mathbb{R}^n : \xi^{\mathsf{T}}x\ge 0 \;\; \forall x\in K\} \tag{12.66b}$$

Let the dual variables be $\lambda\in\mathbb{R}^m$ and $\mu\in K^*$. Define the Lagrangian function:

$$L(x,\lambda,\mu) := f(x) \;-\; \lambda^{\mathsf{T}}(Ax-b) \;-\; \mu^{\mathsf{T}}x, \qquad x\in\mathbb{R}^n,\; \lambda\in\mathbb{R}^m,\; \mu\in K^*\subseteq\mathbb{R}^n$$

The dual function is

$$d(\lambda,\mu) \;:=\; \min_{x\in\mathbb{R}^n} L(x,\lambda,\mu) \;=\; \lambda^{\mathsf{T}}b + d_0(\lambda,\mu),\;\; \lambda\in\mathbb{R}^m,\; \mu\in K^*\subseteq R^n \tag{12.67a}$$

where

$$d_0(\lambda,\mu) \;:=\; \min_{x\in\mathbb{R}^n} \Big(f(x)-(A^{\mathsf{T}}\lambda+\mu)^{\mathsf{T}}x\Big) \tag{12.67b}$$

The dual problem is:

$$d^* \;:=\; \max_{\lambda\in\mathbb{R}^m,\,\mu\in K^*} \lambda^{\mathsf{T}}b \;+\; d_0(\lambda,\mu) \tag{12.67c}$$

For a linear program where $f(x)=c^{\mathsf{T}}x$, $d_0(\lambda,\mu)=0$ if $c=A^{\mathsf{T}}\lambda+\mu$ and $-\infty$ otherwise in which case the dual problem becomes:

$$d^* \;:=\; \max_{\lambda\in\mathbb{R}^m,\,\mu\in K^*} \lambda^{\mathsf{T}}b \quad \text{s.t.} \quad c=A^{\mathsf{T}}\lambda+\mu$$

For strong duality and dual optimality, we can extend the Slater Theorem 12.27 to the more general formulation of dual problem (12.67).

For KKT characterization, we again let $H:=\{x\in\mathbb{R}^n : Ax=b\}$ and rewrite the primal problem (12.65) as an unconstrained convex optimization:

$$\min_{x\in\mathbb{R}^n} \; f(x) \;+\; \delta_H(x) \;+\; \delta_K(x)$$

The constraint qualification in Theorem 12.21 reduces to the Slater condition $H\cap \mathrm{ri}(K)\neq\emptyset$. Under this condition Theorem 12.21 says that $x^*\in\mathbb{R}^n$ is optimal if and only if there exists $\xi^*\in\partial f(x^*)$ such that

$$-\xi^* \;\in\; \partial\left(\delta_H(x^*)+\delta_K(x^*)\right) \;=\; N_H(x^*)+N_K(x^*) \tag{12.68a}$$

where we have used Theorem 12.18 and Table 12.2. From Theorem 12.3 in Chapter 12.1.2,

$$N_H(x^*) = \{A^\mathsf{T}\lambda \in \mathbb{R}^n : \lambda \in \mathbb{R}^m\} \tag{12.68b}$$

$$N_K(x^*) = \{\tilde{\mu} \in K^\circ \subseteq \mathbb{R}^n : \tilde{\mu}^\mathsf{T} x^* = 0\} \tag{12.68c}$$

where K° is the polar cone of K in (12.66a). Substituting these normal cones into (12.68a) leads to the KKT condition for conic program (12.65) in terms of the dual cone K^* of K in (12.66b).[8]

Theorem 12.31 (Strong duality and KKT for conic program)**.** Consider the conic program (12.65) and its dual (12.67). Suppose there exists $\bar{x} \in \mathrm{ri}(K)$ such that $A\bar{x} = b$. Then

1 *Strong duality and dual optimality*. If f^* is finite then there exists a dual optimal solution $(\lambda^*, \mu^*) \in \mathbb{R}^m \times K^*$ that closes the duality gap, i.e., $f^* = d^* = d(\lambda^*, \mu^*)$.
2 *KKT characterization*. A feasible x^* is optimal if and only if there exist a subgradient $\xi^* \in \partial f(x^*)$, a dual feasible $(\lambda^*, \mu^*) \in \mathbb{R}^m \times K^*$ such that

$$\xi^* = A^\mathsf{T}\lambda^* + \mu^*, \qquad \mu^{*\mathsf{T}} x^* = 0$$

In this case (x^*, λ^*, μ^*) is a saddle point that closes the duality gap and is primal-dual optimal. □

Conic constraint.

A generalization of SOCP (12.62) is the following convex optimization

$$f^* := \min_{x \in \mathbb{R}^n} f(x) \text{ s.t. } Ax = b,\ Bx + d \in K \tag{12.69a}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a real-valued convex function, $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $B \in \mathbb{R}^{l \times n}$, $d \in \mathbb{R}^l$ and $K \subseteq \mathbb{R}^l$ is a closed convex cone. The feasible set may not be a cone but (12.69) is still called a conic program because an affine transformation of x is in a closed convex cone. The dual problem can be shown to be (Exercise 12.20):

$$d^* := \max_{(\lambda,\mu) \in \mathbb{R}^{m+l}} d(\lambda, \mu) := \left(b^\mathsf{T}\lambda - d^\mathsf{T}\mu\right) + d_0(\lambda, \mu) \quad \text{s.t.} \quad \mu \in K^* \subseteq \mathbb{R}^l \tag{12.69b}$$

where $d_0(\lambda, \mu) := \min_{x \in \mathbb{R}^n} f(x) - (A^\mathsf{T}\lambda + B^\mathsf{T}\mu)^\mathsf{T} x$. It reduces to (12.67b)(12.67c) when $B = \mathbb{I}_n$ the identity matrix of size n and $d = 0$. When $f(x) = c^\mathsf{T} x$, $d_0(\lambda, \mu) = 0$ if $c = A^\mathsf{T}\lambda + B^\mathsf{T}\mu$ and $-\infty$ otherwise in which case the dual problem becomes:

$$d^* := \max_{\lambda \in \mathbb{R}^m,\ \mu \in K^*} b^\mathsf{T}\lambda - d^\mathsf{T}\mu \quad \text{s.t.} \quad c = A^\mathsf{T}\lambda + B^\mathsf{T}\mu$$

Theorem 12.31 on strong duality, dual optimality and the KKT characterization extends to problem (12.69) (Exercise 12.20). The KKT condition in the next theorem reduces to that in Theorem 12.31 when $B = \mathbb{I}_l$ and $d = 0$.

[8] The definition of the dual problem (12.67) does not require K to be a convex cone, but the normal cone expression (12.68c) holds only if K is a convex cone.

Theorem 12.32 (Strong duality and KKT for conic program)**.** Consider the conic program and its dual (12.69). Suppose the Slater condition is satisfies, i.e., there exists $\bar{x}$ such that $A\bar{x} = b$ and $B\bar{x} + d \in \mathrm{ri}(K)$. Then

1 *Strong duality and dual optimality*. If f^* is finite then there exists a dual optimal solution $(\lambda^*, \mu^*) \in \mathbb{R}^m \times K^*$ that closes the duality gap, i.e., $f^* = d^* = d(\lambda^*, \mu^*)$.
2 *KKT characterization*. A feasible x^* is optimal if and only if there exist a subgradient $\xi^* \in \partial f(x^*)$, a dual feasible $(\lambda^*, \mu^*) \in \mathbb{R}^m \times K^*$ such that

$$\xi^* = A^\mathsf{T}\lambda^* + B^\mathsf{T}\mu^*, \qquad \mu^{*\mathsf{T}}(Bx^* + d) = 0$$

In this case (x^*, λ^*, μ^*) is a saddle point that closes the duality gap and is primal-dual optimal. □

Convex inequality constraint.

A generalization of the conic programs (12.65) and (12.69) is the general convex program whose feasible set is convex but not necessarily of the form $Bx + d \in K$:

$$\min_{x \in \mathbb{R}^n} f(x) \text{ s.t. } Ax = b,\ h(x) \le 0$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function, $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ and $h : \mathbb{R}^n \to \mathbb{R}^l$ is a convex function. If f and h are continuously differentiable then the KKT condition is given by the KKT Theorem 8.15. Otherwise the KKT condition can be derived using the nonsmooth method of this chapter (cf. Exercise 12.21).

12.9 Bibliographical notes

12.10 Problems

Chapter 12.1.

Exercise 12.1 (Feasible direction cones)**.** Let $H := \{x \in \mathbb{R}^n : Ax = b\}$ where $A \in \mathbb{R}^{m \times n}$ and $C \subseteq \mathbb{R}^n$ be a convex cone. Show that the feasible direction cone $D_X(\bar{x}) := \mathrm{cone}(X - \bar{x})$ at an $\bar{x} \in X$ are respectively:

1 $D_H(\bar{x}) = \{y \in \mathbb{R}^n : Ay = 0\}$.
2 $D_C(\bar{x}) = \{y = x - \gamma\bar{x} : x \in C, \gamma \ge 0\}$.

Exercise 12.2 (Normal cone and tangent cone)**.** Let $X \subseteq \mathbb{R}^n$ be a nonempty set and $\bar{x} \in X$.

1 Prove Proposition 12.2.
2 Show that $T_X(\bar{x})$ is generally different from the dual cone $(X-\bar{x})^* = \{y \in \mathbb{R}^n : y^\mathsf{T}(x-\bar{x}) \geq 0, \forall x \in X\}$.
3 Derive the normal cone $N_K(\bar{x})$ and the tangent cone $T_K(\bar{x})$ in Figure 12.3. In particular
   - For Figure 12.3(a), show that $N_K(\bar{x})$ is of the form $N_K(\bar{x}) = \{y = \lambda a : \lambda \geq 0\}$ for some vector $a \in \mathbb{R}^2$ and $T_K(\bar{x}) = \{x \in \mathbb{R}^2 : x^\mathsf{T} a \leq 0\}$ is a half-space.
   - For Figure 12.3(d), show that $N_K(\bar{x})$ is of the form $N_K(\bar{x}) = \{y = \lambda a : \lambda \in \mathbb{R}\}$ and $T_K(\bar{x}) = K$.

   (Hint: Use Theorem 12.3 for $N_K(\bar{x})$ and then Propositions 12.2 and 12.1 for $T_K(\bar{x})$.) □

Exercise 12.3 (Normal cone $N_C(x)$)**.** Let $C := \{x \in \mathbb{R}^n : h(x) \leq 0\}$ where $h : \mathbb{R}^2 \to \mathbb{R}^2$ are given by $h_i(x_1,x_2) := \frac{1}{2}\left(x_1^2 + a_i x_2^2\right) - b_i$, $i = 1,2$ with $a_i > 0, b_i > 0$ and $b_1/a_1 < b_2/a_2$; see Figure 12.5. Let $\bar{x} := \left(0, \sqrt{2b_1/a_1}\right)$. This exercise derives the normal cone $N_C(\bar{x})$ without the LICQ assumption in Theorem 12.4.

1 Show directly that the normal cone $N_C(\bar{x}) = \{(0, y_2) \in \mathbb{R}^2 : y_2 \geq 0\}$.
2 Show that $N_C(\bar{x}) = \{\nabla h(\bar{x})\lambda \in \mathbb{R}^2 : \lambda \in \mathbb{R}^2_+, \lambda^\mathsf{T} h(\bar{x}) = 0\} = \{\lambda_1 \nabla h_1(\bar{x}) \in \mathbb{R}^2 : \lambda_1 \geq 0\}$ as Theorem 12.4 indicates.

Exercise 12.4 (Tangent cones)**.** Derive the tangent cones in Table 12.1 of Chapter 12.1.2. Assume h is twice continuously differentiable and satisfies LICQ (12.5) at $\bar{x} \in X$. (Hint: Proposition 12.2 and Theorem 12.3.)

Exercise 12.5 (Image of linear transformation of convex cone)**.** Given a nonempty set $X \subseteq \mathbb{R}^n$ let $Y := AX$ for some matrix $A \in \mathbb{R}^{m \times n}$. From Theorem 12.6, the normal cone of Y at a $\bar{y} = A\bar{x} \in Y$ with $\bar{x} \in X$ is the pre-image of $N_X(\bar{x})$: $N_Y(\bar{y}) = \{y \in \mathbb{R}^m : A^\mathsf{T} y \in N_X(\bar{x})\}$. Show that when X is a convex cone then

$$N_Y(\bar{y}) \;=\; \{y \in \mathbb{R}^m : A^\mathsf{T} y \in X^\circ,\; y^\mathsf{T}\bar{y} = 0\}$$

Exercise 12.6 (Pre-image of linear transformation of convex cone)**.** Let $Y := \{y \in \mathbb{R}^m : y \leq 0\}$ and $X := \{x \in \mathbb{R}^n : Ax \in Y\}$ be its pre-image under $A \in \mathbb{R}^{m \times n}$. Use the Farkas Lemma (Theorem 8.12) to show directly that $X^\circ = A^\mathsf{T} Y^\circ$.

Exercise 12.7 (Pre-image of linear transformation of convex cone)**.** Consider the convex cone Y and its pre-image X under a singular matrix A:

$$Y := \{y \in \mathbb{R}^2 : y_1 \geq y_2 \geq 0\}, \quad A := \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad X := \{x \in \mathbb{R}^2 : Ax \in Y\}$$

1 Derive X°, Y° and compare $A^\mathsf{T} Y^\circ$ and X°.
2 Derive $N_Y(\bar{y})$ and $N_X(\bar{x})$ where $\bar{y} = A\bar{x} \in Y$ for $\bar{x} = (0,0), (1,-1), (1,1)$.

Chapter 12.2.

Exercise 12.8 (Closedness and lsc of f; [54].)**.** 1 For a function $f : \mathbb{R}^n \to [-\infty, \infty]$, show that it is closed if and only if it is lsc on $\mathbb{R}^n$ if and only if its level set $V_\gamma := \{x | f(x) \leq \gamma\}$ is closed for every $\gamma \in \mathbb{R}$.
2 For $f : X \to [-\infty, \infty]$ where $X \subseteq \mathbb{R}^n$, show that it is closed if its effective domain $\mathrm{dom}(f)$ is closed and f is lsc on $\mathrm{dom}(f)$.
3 Consider a real-valued function $f : X \to \mathbb{R}^n$ where $X \subseteq \mathbb{R}^n$ is nonempty. Extend f to the extended real-valued function $f_X(x) : \mathbb{R}^n \to [-\infty, \infty]$ defined by

$$f_X(x) := \begin{cases} f(x) & \text{if } x \in X \\ \infty & \text{if } x \notin X \end{cases}$$

Show that f_X is closed (on $\mathbb{R}^n$) if the effective domain $\mathrm{dom}(f)$ is closed and f is lower semicontinuous on $\mathrm{dom}(f)$.

Exercise 12.9 (Support function $\sigma_X(x)$)**.** Prove (12.24).

Chapter 12.3.

The proof of the existence of subgradient for a proper convex function at $\bar{x}$, using (12.27), requires $\bar{x} \in \mathrm{int}(\mathrm{dom}(f))$. The next exercise shows that, even though the contradiction argument there may break down if $\bar{x} \in \mathrm{ri}(\mathrm{dom}(f))$, a subgradient may still exist at such a $\bar{x}$.

Exercise 12.10 (Existence of subgradient.)**.** Consider the proper extended real-valued function $f : \mathbb{R}^2 \to (-\infty, \infty]$ defined by

$$f(x_1, x_2) = \begin{cases} x_1^2 & \text{if} \quad x_2 = 0 \\ \infty & \text{if} \quad x_2 \neq 0 \end{cases}$$

The effective domain $\mathrm{dom}(f) = \{x \in \mathbb{R}^2 : x_2 = 0\} = \mathrm{ri}(\mathrm{dom}(f))$, $\mathrm{epi}(f)$ is in a vertical plane, and hence $\mathrm{int}(\mathrm{dom}(f)) = \emptyset$. Show that subgradient exists at every point $\bar{x} \in \mathrm{dom}(f)$, even though $\frac{\partial f}{\partial x_2}(x)$ is not well defined.

Exercise 12.11 (Jensen's inequality)**.** Suppose X is a random variable taking value in $\mathbb{R}^n$ with finite expectation EX. Show that if $f : \mathbb{R}^n \to \mathbb{R}$ is convex on $\mathbb{R}^n$ then $E(f(X)) \geq f(EX)$. (Hint: Use subgradient of f.)

Exercise 12.12 (δ_X, σ_X and their subdifferentials)**.** Fix any nonempty subset $X \subseteq \mathbb{R}^n$. Consider the extended real-valued indicator function and support function defined respectively by:

$$\delta_X(x) := \begin{cases} 0 & \text{if} \quad x \in X \\ \infty & \text{if} \quad x \notin X \end{cases}, \qquad \sigma_X(x) := \sup_{y \in X} y^T x$$

Let f^* and ∂f denote respectively the conjugate and subdifferential of f. Show that:

1. $\delta_X^*(y) = \sigma_X(y)$.
2. If X is a cone then $\delta_X^*(y) = \delta_{X^\circ}(y)$, i.e., the support function of a cone is an indicator function of its polar cone.
3. Suppose X is a convex set. Then $\partial \delta_X(x) = N_X(x)$.
4. Suppose X is a nonempty closed convex set.
   1. $\sigma_X^*(x) = \delta_X(x)$.
   2. $\partial \sigma_X(x) = \{y \in \mathbb{R}^n : y^T x = \sigma_X(x)\}$. (Hint: Apply Lemma 12.16 to earlier results.)

Exercise 12.13 (Normal cone of set intersection.)**.**
1. Prove Lemma 12.5. (Hint: Use Theorem 12.18 whose proof does not rely on Lemma 12.5 so there is no circular argument.)
2. As an application of Lemma 12.5 consider $C := \{x \in \mathbb{R}^n : Ax = b, x \in K\}$ where $A \in \mathbb{R}^{m \times n}$ and $K \subseteq \mathbb{R}^n$ is a convex cone. Suppose there is $\bar{x} \in \mathrm{ri}(K)$ with $A\bar{x} = b$. Show that $N_C(\bar{x}) = \{A^\mathsf{T}\lambda + y : \lambda \in \mathbb{R}^m, y \in K^\circ, y^\mathsf{T}\bar{x} = 0\}$ for any $\bar{x} \in C$, where K° denotes the polar cone of K.

Chapter 12.4.

Chapter 12.5.

Exercise 12.14 (Generalized KKT)**.** Consider the second-order cone program:

$$f^* := \min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t.} \quad x \in K := \{x \in \mathbb{R}^n : \|x^{n-1}\|_2 \leq x_n\}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a real-valued convex function (not necessarily differentiable) and K is the standard second-order cone. Suppose $\mathrm{ri}(\mathrm{dom}(f)) \cap \mathrm{int}(K) \neq \emptyset$. Show that $x^* := 0$ is optimal if and only if there exists $y^* \in \partial f(0)$ such that $\|y^{*n-1}\|_2 \leq y_n^*$.

Chapter 12.6.

Exercise 12.15 (Primal optimality.)**.** Prove Corollary 12.23. (Hint: Use Remark 12.3 and the Weierstrass Theorem 12.22.)

Exercise 12.16 (Primal optimal solutions.)**.** Consider X and f in Theorem 12.26 where $X \subseteq \mathbb{R}^n$ is closed and convex, $f : \mathbb{R}^n \to (-\infty, \infty]$ is closed proper convex, and $X \cap \mathrm{dom}(f) \neq \emptyset$. Suppose X and f have a common nonzero direction of recession. Let the level sets be $V_\gamma := \{x \in \mathbb{R}^n : f(x) \leq \gamma\}$ and $X'(\gamma) := X \cap V_\gamma$.

1 Show that $X'(\gamma)$ is unbounded for any $\gamma \in \mathbb{R}^n$.
2 If $V_\gamma = \emptyset$ for small enough γ, show that there is a smallest γ_0 for which $V_{\gamma_0} \neq \emptyset$. Moreover the primal solution set is unbounded.

Chapter 12.7.

Exercise 12.17 (Weak duality)**.** Let $M \subseteq \mathbb{R}^{l+1}$ be a nonempty set, not necessarily convex, and define the following pair of problems:

$$w^* := \inf_{(0,w)\in M} w, \qquad d^* := \sup_{\mu\in\mathbb{R}^l} d(\mu)$$

where $d(\mu) := \inf_{(u,w)\in M} \mu^T u + w$ and $w^* := \infty$ if $(0, w) \notin M$ for any w. Show that $d^* \leq w^*$.

Exercise 12.18 (Equivalent dual problem)**.** Show that the problems in (12.52) are equivalent to those in (12.50), assuming there is a feasible point $\bar{x} \in \mathrm{dom}(f) \cap X' \cap \{x : h(x) \leq 0\}$.

Exercise 12.19 (Dual function and level set)**.** Consider Theorem 12.28 under CQ1 (there exists $\bar{x} \in \mathrm{dom}(f) \cap X'$ such that $h(\bar{x}) < 0$). Recall the dual function defined in (12.50c):

$$d(\mu) := \begin{cases} \inf_{x\in X} f(x) + \mu^{\mathsf{T}} h(x), & \mu \geq 0 \\ -\infty, & \text{otherwise} \end{cases}$$

and define the level set of the dual function d:

$$Q := Q_a := \{\mu \in \mathbb{R}^l : \mu \geq 0,\ f(x) + \mu^{\mathsf{T}} h(x) \geq a, \forall x \in X\}$$

(Since $f^* \geq a$, $Q^* := Q_{f^*} \subseteq Q_a =: Q$.) Show that

1 $-d(\mu)$ is a closed proper convex (CPC) function over $\mathbb{R}^l$.

2 Q is nonempty, convex and compact.

Chapter 12.8.

Exercise 12.20 (Conic program: KKT)**.** Consider the conic program (12.69).

1 Derive its dual problem (12.69b).
2 Prove Theorem 12.32.

Exercise 12.21 (Convex inequality constraints: KKT)**.** Consider the convex optimization

$$\min_{x\in\mathbb{R}^n} f(x) \text{ s.t. } Ax = b,\ h(x) \le 0 \tag{12.70}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function, $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ and $h : \mathbb{R}^n \to \mathbb{R}^l$ are convex functions. Suppose the Slater condition is satisfied, i.e., there exists $\bar{x}$ with $A\bar{x} = b$ and $h(\bar{x}) < 0$, and that the primal optimal value is finite.

1 Suppose h is twice continuously differentiable (but f may not) and a feasible x^* satisfies the LIQC (12.5). Show that x^* is optimal if and only if there exist a subgradient $\xi^* \in \partial f(x^*)$, a $\lambda^* \in \mathbb{R}^m$, and a *unique* $\mu^* \in \mathbb{R}^l_+$ such that

$$\xi^* + A^\mathsf{T}\lambda^* + \nabla h(x^*)\mu^* = 0, \qquad \mu^{*\mathsf{T}} h(x^*) = 0$$

2 Suppose neither f nor h are continuously differentiable. Show that a feasible x^* is optimal if and only if there exist subgradients $\xi^* \in \partial f(x^*)$ and $\theta_i^* \in \partial h_i(x^*)$, and a dual optimal solution $(\lambda^*, \mu^*) \in \mathbb{R}^{m+l}$ such that $\mu^* \ge 0$ and

$$\xi^* + A^\mathsf{T}\lambda^* + \Theta^{*\mathsf{T}}\mu^* = 0, \qquad \mu^{*\mathsf{T}} h(x^*) = 0$$

where the rows of the matrix Θ^* are θ_i^* (provided an appropriate constraint qualification is satisfied).

13 Stochastic OPF

This chapter presents basic methods for stochastic optimization and their application to optimal power flow problems. Optimal power flow problems we have studied in previous chapters take the form

$$\min_{x \in \mathbb{R}^n} \quad f(x) \ \text{ s.t. } \ h(x, \zeta) \leq 0 \tag{13.1}$$

where x is a decision variable and $f : \mathbb{R}^n \to \mathbb{R}$ is a cost function, and $h(x, \zeta) : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^m$ is a constraint function, as a function of both the decision variable x and a parameter $\zeta \in \mathbb{R}^k$. For instance the problem computes an optimal generation schedule x^* to meet a demand ζ subject to power flow equations and operational constraints. Or it computes an optimal setpoint x^* for smart inverters to help stabilize voltages in a distribution system in response to photovoltaic generation ζ. In general the parameter ζ is uncertain. So far we have implicitly assumed that the decision x^* is made either based on a forecast $\hat{\zeta}$ of the random parameter ζ or after ζ has been realized, and therefore the problem is deterministic. In this chapter we study the case where at least some of the decisions must be made before the random ζ is realized and simply substituting the forecast $\hat{\zeta}$ for ζ is inadequate. We study four approaches to making decisions under uncertain ζ.

In the first approach an uncertainty set Z is assumed known in which the uncertain parameter ζ takes value. An optimal x^* is chosen with respect to a worst-case $\zeta \in Z$, i.e., the constraint $h(x^*, \zeta) \leq 0$ must be satisfied for all $\zeta \in Z$. This leads to robust optimization (Chapter 13.1) where the single constraint in (13.1) is replaced by a possibly infinite set of constraints $(h(x^*, \zeta) \leq 0, \forall \zeta \in Z)$. Robust optimization can be too conservative as it demands constraint satisfaction in the worst-case realization of the uncertain parameter $\zeta \in Z$. This motivates the second approach where the uncertain parameter $\zeta := \zeta(\omega) \in Z$ is a random vector on a given probability space with a known probability measure $\mathbb{P}$. An optimal x^* is chosen so that the constraint $h(x^*, \zeta) \leq 0$ is satisfied with high probability, not necessarily for all $\zeta \in Z$ (or with probability 1 under $\mathbb{P}$). This leads to chance constrained optimization (Chapter 13.2) where the constraint $h(x^*, \zeta) \leq 0$ in (13.1) is replaced by $\mathbb{P}(h(x^*, \zeta) \leq 0) \geq 1 - \epsilon$ with a given tolerance ϵ for constraint violation. Chance constrained optimization can be intractable for common $\mathbb{P}$; moreover $\mathbb{P}$ may not be known in many applications even when random samples of ζ under $\mathbb{P}$ are available, e.g., measurements of ζ from a real power system. This motivates the third approach, called scenario optimization (Chapter 13.3), where

the single constraint $h(x^*,\zeta) \le 0$ in (13.1) is replaced by N randomized constraints $\left(h(x^*,\zeta^i) \le 0, i = 1,\ldots,N\right)$ defined by N independent random samples of $\zeta^1,\ldots,\zeta^N$ under $\mathbb{P}$. Unlike the other three approaches where the optimization problem is deterministic, a scenario program is a randomized problem. If N is sufficiently large then the resulting randomized optimal solution x^* will likely satisfy the chance constraint, in expectation or probability. Finally we study two-stage stochastic optimization with recourse where some decisions must be made before the random ζ is realized and other decisions can be made afterwards in response to the observed realization of ζ.

In this chapter we introduce the basic theory for each of these four approaches and apply it to power system problems. Most stochastic optimization problems are nonconvex and computationally hard. Our emphasis is on conditions under which these problems have equivalent finite convex reformulations. Even though these reformulated problems often introduce extended real-valued and nondifferentiable functions, especially in two-stage optimization problems, optimality conditions can be derived using nonsmooth techniques studied in Chapter 12. Moreover computation algorithms studied in Chapter 8 can be adapted to solve these convex but nonsmooth problems with gradients replaced by subgradients.

13.1 Robust optimization

13.1.1 General formulation

A robust optimization problem is of the form:

$$\min_{x\in\mathbb{R}^n} \quad f(x) \quad \text{s.t.} \quad h_i(x,\zeta_i) \le 0,\ \forall \zeta_i \in Z_i(x),\ i = 1,\ldots,m \tag{13.2}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a cost function. For $i = 1,\ldots,m$, $\zeta_i \in \mathbb{R}^{k_i}$ are given parameters, and $h_i : \mathbb{R}^n \times \mathbb{R}^{k_i} \to \mathbb{R}$ are constraint functions. Here ζ_i are *uncertain parameters* that take values in *uncertainty sets* $Z_i(x) \subseteq \mathbb{R}^{k_i}$. It is convenient in applications to allow the uncertainty sets $Z_i(x)$ to depend on x (see Example 13.1) and hence we can regard each $Z_i : \mathbb{R}^n \to 2^{\mathbb{R}^{k_i}}$ as a set-valued map on $\mathbb{R}^n$. The problem seeks an optimal solution x^* that minimizes the cost function $f(x)$ and remains feasible for all possible realizations of the uncertain parameters $\zeta_i \in Z_i(x^*)$, $\forall i$. It is called a *robust program*. If some of the $Z_i(x)$ are continuous sets, then (13.2) is called a *semi-infinite* problem because it contains a finite number of variables but an infinite number of constraints. As a consequence the robust counterpart of a nominal problem (when $Z_i(x)$ are singletons) is generally computationally intractable even if the nominal problem is simple such as a linear program. In Chapters 13.1.2, 13.1.3 and 13.1.4 we present three classes of robust programs that are tractable. Specifically we will derive finite convex reformulation for these problems to which techniques in Chapters 12 and 8 can be applied.

Remark 13.1. The formulation (13.2) makes two assumptions without loss of generality:

1 *Certain and linear cost function*. It assumes that the cost function f is certain. Otherwise, we can introduce an additional variable t and an additional constraint to obtain the following equivalent problem that has uncertainty only in the constraints:

$$\min_{x\in\mathbb{R}^n, t\in\mathbb{R}} \quad t \quad \text{s.t.} \quad f(x,\zeta_0) - t \le 0, \quad h_i(x,\zeta_i) \le 0, \ \forall \zeta_i \in Z_i(x), \ i = 0,\dots,m$$

where $\zeta_0 \in Z_0(x)$ is the uncertain parameter of the cost function f. This also shows that we can assume without loss of generality that the cost is linear.

2 *Direct product of uncertainties*. It assumes that the uncertainty set is a direct product $Z(x) := Z_1(x) \times \cdots \times Z_m(x)$ of individual uncertainty sets $Z_i(x)$. If the uncertainty set $Z(x) \subseteq \mathbb{R}^{\sum_{i=1}^m k_i}$ is not a direct product, the robust optimization problem

$$\min_{x\in\mathbb{R}^n} \quad f(x) \quad \text{s.t.} \quad \sup_{\zeta\in Z(x)} \ h_i(x,\zeta_i) \le 0, \ i = 1,\dots,m$$

can be specified with an equivalent uncertainty set $\hat{Z}(x) := Z_1(x) \times \cdots \times Z_m(x)$ that is a direct product:

$$\min_{x\in\mathbb{R}^n} \quad f(x) \quad \text{s.t.} \quad \sup_{\zeta_i\in Z_i(x)} \ h_i(x,\zeta_i) \le 0, \ i = 1,\dots,m$$

where $Z_i(x) := \{\zeta_i : \zeta := (\zeta_1,\dots,\zeta_m) \in Z(x)\}$ is the projection of $Z(x)$ onto the ith coordinate. This is because h_i depends on ζ_i, not on $\zeta_j, j \ne i$, and therefore given x, $\sup_{\zeta\in Z(x)} h_i(x,\zeta_i) = \sup_{\zeta_i\in Z_i(x)} h_i(x,\zeta_i)$.

3 *Equality constraints without recourse*. The nominal problem for many applications contain equality constraints, resulting in a robust counterpart involving uncertain equality constraints of the form:

$$\min_{x,y} \ f(x) \ \text{ s.t. } \ g_i(x,y_i,\zeta_i) = 0, \ h_i(x,y_i,\zeta_i) \le 0, \ \forall \zeta_i \in Z_i(x), \ i = 1,\dots,m \tag{13.3}$$

An equality constraint such as $y_i = \zeta_i$ where $\zeta_i \in \{0,1\}$ is generally infeasible for robust optimization if y_i is also an optimization variable that must be chosen and fixed before the uncertain ζ_i is realized. There are three common approaches to avoid infeasibility by eliminating equality constraints. The first is to allow slack by replacing equality constraints by inequality constraints on the size of the slack; see (13.133) in Chapter 13.5.2 on robust economic dispatch for an example. The second is to replace the inequality constraints on the slack by penalty terms in the cost function that allow but penalize violation of the equality constraints. The third is to eliminate the equality constraints by substituting dependent variables into the cost function and inequality constraints, as we now explain. We assume the equality constraint $g_i(x,y_i,\zeta_i) = 0$ means that given (the control) x, (the system state) y_i will be determined by x and the realization of the uncertain parameter ζ_i. Given an x let $Y_i(x) := \{y_i : g_i(x,y_i,\zeta_i) = 0, \ \zeta_i \in Z_i(x)\}$ denote the set of y_i

implicitly defined by g_i as ζ_i varies over $Z_i(x)$. Then the constraints in (13.3) are interpreted as

$$h_i(x, y_i, \zeta_i) \le 0, \qquad \forall (y_i, \zeta_i) \in Y_i(x) \times Z_i(x), \quad i = 1, \dots, m$$

which is of the form in (13.2), i.e., (the system state) y_i becomes an uncertain parameter determined by the equality constraint g_i and ζ_i. Note that y_i depends only on ζ_i, but not ζ_j, $j \neq i$, so that the uncertainty set $Y_i(x)$ is separable in i. Hence the equality constrained problem (13.3) should be interpreted as

$$\min_x f(x) \quad \text{s.t.} \quad \sup_{\zeta_i \in Z_i(x)} \sup_{y_i \in Y_i(x)} h_i(x, y_i, \zeta_i) \le 0, \ i = 1, \dots, m$$

See Example 13.1 and Exercise 13.3.

This is different from stochastic optimization with recourse studied in Chapter 13.4 where a first-stage decision is made before the uncertain parameter ζ is realized and a second-stage decision is made after ζ is realized. With recourse, it is possible to satisfy uncertain equality constraints and, indeed, the feasibility condition plays an important role in optimality conditions for two-stage optimization studied in Chapter 13.4.

4 *Closed and convex Z*. We will assume without loss of generality that the uncertainty set Z is closed and convex (Exercise 13.2).

□

Example 13.1 (Robust optimization: voltage control)**.** Consider a solar panel with uncertain real power generation ζ_t at time t that takes value in a set $Z_t \subseteq \mathbb{R}_+$. Suppose its reactive power q_t is controllable within the range $q_t \in [q^{\min}, q^{\max}]$ for all t. The solar panel is connected to a battery through a line with a given series admittance $y := g + \mathbf{i}b \in \mathbb{C}$. The DC discharging power d_t of the battery is controllable within the range $d_t \in [d^{\min}, d^{\max}]$ as long as its state of charge b_t satisfies the energy capacity $b_t \in [0, B]$. Let $v_{1t} := |v_{1t}| e^{\mathbf{i}\theta_{1t}}$ and $v_{2t} := |v_{2t}| e^{\mathbf{i}\phi_{2t}}$ denote the voltage phasors at the solar panel and the battery respectively at time t. Our goal is to schedule the reactive power $q := (q_1, \dots, q_T) \in \mathbb{R}^T$ and discharging power $d := (d_1, \dots, d_T) \in \mathbb{R}^T$ to minimize a certain cost f subject to the constraint that the voltages $v_t := (v_{1t}, v_{2t}) \in \mathbb{C}^2$ satisfy voltage limits $|v_{it}| \in [v^{\min}, v^{\max}]$ for $i = 1, 2$, for all realizations of the solar generation $\zeta_t \in Z_t$, for $t = 1, \dots, T$.

This can be formulated as a robust optimal power flow (OPF) problem.[1] Let $x := (q, d) \in \mathbb{R}^{2T}$ where q, d are defined above. Let $f(x)$ denote the cost function. Let $b := (b_1, \dots, b_T) \in \mathbb{R}^T$ and $v := (v_1, \dots, v_T)$. Suppose the uncertain solar generation $\zeta := (\zeta_1, \dots, \zeta_T) \in \mathbb{R}^{\mathsf{T}}$ takes value in $Z \subseteq \mathbb{R}^{\mathsf{T}}$, independent of x. As explained in Remark 13.1 we can assume without loss of generality that $Z = Z_1 \times \cdots \times Z_T$ with $Z_t := (\zeta_t : z \in Z)$. The robust scheduling problem is

$$\min_x f(x) \ \text{ s.t. } \ g(x, v, b\zeta) = 0, \ h(x, v, b, \zeta) \le 0, \ \forall \zeta_t \in Z_1 \times \cdots \times Z_T \tag{13.4a}$$

[1] We formulate the OPF problem in the complex domain for notational simplicity; it is straightforward to convert it into OPF in the real domain.

where the equality constraint $g(x,v,b,\zeta)=0$ is power flow equations and battery state transition: for $t=1,\ldots,T$,

$$\zeta_t + \mathbf{i}q_t \;=\; y^{\mathsf{H}}\left(|v_{1t}|^2 - v_{1t}v_{2t}^{\mathsf{H}}\right), \qquad d_t + \mathbf{i}0 \;=\; y^{\mathsf{H}}\left(|v_{2t}|^2 - v_{2t}v_{1t}^{\mathsf{H}}\right) \tag{13.4b}$$

$$b_{t+1} \;=\; b_t - d_t \tag{13.4c}$$

and the inequality constraint $h(x,v,b,\zeta)\le 0$ is voltage and battery limits: for $t=1,\ldots,T$,

$$v^{\min} \le |v_{it}| \le v^{\max},\;\; i=1,2, \qquad 0 \le b_t \le B \tag{13.4d}$$

The equality constraint (13.4c) has no uncertainty. The uncertain equality constraint (13.4b) should be interpreted. Both can be eliminated, as follows. In reality we set the values of the reactive power q_t and discharging power d_t, which then, together with the uncertain solar generation ζ_t, determine the voltages $v_t := (v_{1t}, v_{2t})$ according to the power flow equation (13.4b). Let $V_t(x) := \{v_t \in \mathbb{C}^2 : v_t \text{ satisfies (13.4b)},\ \zeta_t \in Z_t\}$ denote the set of power flow solutions as ζ_i varies in Z_t. We can eliminate the uncertain equality constraint (13.4b) using the new uncertainty set $V_t(x)$, and eliminate the (fixed) equality constraint on the battery's state of charge b_t by expanding on the battery state (given initial state b_0):

$$b_t \;=\; b_0 - \sum_{s<t} d_s, \qquad t=1,\ldots,T$$

to obtain the reformulation:

$$\begin{aligned}\min_x\; f(x) \quad \text{s.t.} \quad & v^{\min} \le |v_{it}| \le v^{\max},\; i=1,2,\; \forall v_t \in V_t(x),\; t=1,\ldots,T \\ & 0 \le b_0 - \sum_{s<t} d_s \le B,\; t=1,\ldots,T\end{aligned}$$

which is in the form (13.2). Note that the uncertainty sets Z_t, which are independent of x, have been incorporated into the new uncertainty sets $V_t(x)$ which depend on x. □

The tractability of the robust optimization problem (13.2) depends on the structure of the nominal problem and that of the uncertainty set $Z(x) := Z_1(x)\times\cdots\times Z_m(x) \subseteq \mathbb{R}^{k_1}\times\cdots\times\mathbb{R}^{k_m}$ for each x. If we let the robust feasible set be:

$$X \;:=\; X(Z) := \{\, x \,:\, h_i(x,\zeta_i)\le 0,\; \forall \zeta_i \in Z_i(x),\; i=1,\ldots,m \,\}$$

then the tractability of the robust problem often boils down to whether there is a finite convex representation of $X(Z)$. Since the direct product of $Z_i(x)$ preserves convexity we can assume without loss of generality that $m=1$ in (13.2) and consider the tractability of

$$\min_{x\in\mathbb{R}^n}\; f(x) \quad \text{s.t.} \quad h(x,\zeta)\le 0,\; \forall \zeta \in Z(x) \tag{13.5}$$

where $f:\mathbb{R}^n\to\mathbb{R}$ and $h:\mathbb{R}^n\times\mathbb{R}^k\to\mathbb{R}$ are convex functions and $Z(x)\subseteq\mathbb{R}^k$ is a convex set for every $x\in\mathbb{R}^n$.

The derivation of a tractable reformulation of (13.5) often uses the following concept.

Definition 13.1. A set $X^+ \subseteq \mathbb{R}^n \times \mathbb{R}^m$ is said to *represent* a set $X \subseteq \mathbb{R}^n$ if the projection of X^+ onto the space of x-variable is exactly X, i.e., $X = \{x \ : \ (x,y) \in X^+, \ y \in \mathbb{R}^m\}$.

This simple technique can sometimes be used to greatly reduce the number of constraints. For instance the l_1-norm ball

$$X := \left\{ x \in \mathbb{R}^n \ : \ \|x\|_1 := \sum_i |x_i| \leq 1 \right\}$$

is defined by 2^n linear inequalities, but can be represented by a much simpler set X^+ defined by $2n+1$ linear inequalities in $2n$ variables (Exercise 13.1):

$$X^+ := \left\{ (x,y) \in \mathbb{R}^{2n} \ : \ -y_i \leq x_i \leq y_i, \ i = 1, \ldots, n, \ \sum_i y_i \leq 1 \right\}$$

Note that y in X^+ satisfies $y_i \geq 0$ for all i. Indeed y_i plays the role of $|x_i|$.

More importantly we will use this concept to derive a finite convex representation X^+, which does not depend on the uncertainty set Z, of the possibly semi-infinite feasible set $X(Z)$. Then (13.5) can be reformulated as

$$\min_{x,y} \ f(x) \quad \text{s.t.} \quad (x,y) \in X^+ \subseteq \mathbb{R}^{n+m} \tag{13.6}$$

which is tractable when f is a convex cost function and X^+ is a convex feasible set. We first summarize the general strategy.

Derivation strategy.

The key observation is that (13.5) is equivalent to

$$\min_{x \in \mathbb{R}^n} \ f(x) \quad \text{s.t.} \quad \sup_{\zeta \in Z(x)} \ h(x,\zeta) \leq 0 \tag{13.7}$$

This is called a *bi-level* problem and generally intractable. It often has a tractable reformulation when, for each fixed $x \in \mathbb{R}^n$, the subproblem

$$\bar{h}(x) := \sup_{\zeta \in Z(x)} \ h(x,\zeta) \tag{13.8}$$

is a convex problem and the constraint $\bar{h}(x) \leq 0$ has a finite convex representation. By assumption $f : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}$ are convex functions and $Z(x) \subseteq \mathbb{R}^k$ is a convex set for every $x \in \mathbb{R}^n$.

There are three general strategies to eliminate the uncertain parameter ζ from (13.7) and derive a tractable reformulation:

1 *Solve $\hat{h}(x)$ in closed form*. When the subproblem (13.8) for each $x \in \mathbb{R}^n$ can be solved to obtain $\bar{h}(x)$ in closed form then the semi-infinite problem (13.7) is equivalent to the finite problem

$$\min_{x \in \mathbb{R}^n} \; f(x) \quad \text{s.t.} \quad \bar{h}(x) \leq 0$$

If $Z(x) = Z$ independent of x then, since $h(x, \zeta)$ is convex in x for each ζ, $\bar{h}(x)$ is convex in x. In this case the robust program has a tractable convex representation studied in Chapters 8 and 12. This strategy is used to prove Theorem 13.1 (for linear and SOC uncertainty).

2 *Replace $\hat{h}(x) \leq 0$ by strong duality and KKT condition*. Suppose the subproblem (13.8) is convex for each $x \in \mathbb{R}^n$ but cannot be explicitly solved. Then

   1 Using strong duality we replace $\bar{h}(x) \leq 0$ in (13.7) by $d(y; x) \leq 0$ where, for each x, $d(\cdot; x)$ is the Lagrangian dual function of (13.8) and y is a dual optimal solution.

   2 The dual optimality of y is enforced by dual feasibility and stationarity $\nabla_\zeta L(\zeta, y; x) = 0$ of the KKT condition for (13.8). These conditions do not contain ζ, but only (x, y), because (i) $h(x, \zeta)$ is affine in ζ and hence the stationarity condition $\nabla_\zeta L(\zeta, y; x) = 0$ is independent of ζ; and (ii) strong duality and stationarity imply complementary slackness and hence the complementary slackness condition can be omitted. Feasibility is reformulated as: x is feasible for (13.7) if and only if there exists y such that (x, y) satisfies

$$d(y, x) \;\leq\; 0, \qquad \text{KKT}(x, y) \;\leq\; 0 \tag{13.9}$$

   where $\text{KKT}(x, y) \leq 0$ is dual feasibility and stationarity. If $d(x, y)$ and the KKT function $\text{KKT}(x, y)$ are convex then the semi-infinite problem (13.7) is equivalent to the convex problem $\min_{x,y} f(x)$ s.t. (13.9) which is of the form (13.6).

   This strategy needs the Slater Theorem 8.17 to ensure strong duality and dual optimality. It is used to prove Theorems 13.1 (for conic uncertainty) and 13.2 below.

3 *Replace $\hat{h}(x) \leq 0$ by linear matrix inequalities*. Sometimes the semi-infinite constraint in (13.7) takes the form $h_0(x) + h(x, \zeta) \in K$ for all $\zeta \in Z(x)$ where, for each ζ, $h_0(\cdot)$ and $h(\cdot, \zeta)$ are affine functions of x, for each x, $h(x, \cdot)$ is an affine function of the uncertain parameter ζ, and K is a closed convex cone such as the second-order cone $K_{\text{soc}} \subseteq \mathbb{R}^n$ or the semidefinite cone $K_{\text{sdp}} \subseteq \mathbb{S}^n$. This is the case in Theorems 13.3 and 13.4 where $Z(x)$ is a set of matrices with bounded spectrum norms. For both theorems the constraint can be reformulated as a finite set of linear matrix inequalities using the S-lemma and the resulting problem is a semidefinite program.

As we will see below tractability often requires the uncertainty set $Z(x) = Z$ to be independent of x. For instance a robust linear program with the uncertainty set $Z :=$

$\{\zeta \in \mathbb{R}^L : \|\zeta\|_\infty \le 1\}$ remains a linear program, but may become intractable if $Z(x) := \{\zeta \in \mathbb{R}^L : \|\zeta\|_\infty \le h(x)\}$; see Exercise 13.4.

In the rest of this section we use the general strategy above to derive the convex reformulations of three classes of (h, Z) for which (13.5) is tractable, corresponding to robust counterparts of uncertain linear program, second-order cone program, and semidefinite program.

13.1.2 Robust linear program

Consider (13.5) where f is linear and h is affine in x and ζ separately, giving rise to the following robust counterpart of an uncertain linear program:

$$\min_{x\in\mathbb{R}^n} c^\mathsf{T} x \text{ s.t. } a^\mathsf{T} x \le b, \forall [a^\mathsf{T}\ b] \in \left\{ \left[a_0^\mathsf{T}\ b_0\right] + \sum_{l=1}^{k} \zeta_l \left[a_l^\mathsf{T}\ b_l\right] : \zeta \in Z \subseteq \mathbb{R}^k \right\} \tag{13.10}$$

where $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$ are uncertain parameters. The row vector $\left[a_0^\mathsf{T}\ b_0\right]$ are nominal parameters and $\left[a_l^\mathsf{T}\ b_l\right]$ are basic perturbations modulated by the uncertain ζ in the uncertainty set Z. It does not lose generality to assume that the uncertain vector $[a^\mathsf{T}\ b]$ takes this form because taking $k = n+1$ will allow each entry of a and b to vary independently. We assume without loss of generality that Z is such that the feasible set is nonempty, closed and convex. The uncertainty set Z is independent of x; otherwise (13.10) may not be tractable; see Exercise 13.4.

Write (13.10) as a bi-level problem:

$$\min_{x\in\mathbb{R}^n} c^\mathsf{T} x \quad \text{s.t.} \quad \max_{\zeta\in Z} \sum_{l=1}^{k} \zeta_l (a_l^\mathsf{T} x - b_l) \le -(a_0^\mathsf{T} x - b_0) \tag{13.11}$$

The corresponding constraint function $h(x, \zeta)$ is affine in x for each ζ and affine in ζ for each x. Our goal is to derive a finite convex representation of the semi-infinite feasible set in (13.10), and thus convert the semi-infinite linear program into an explicit convex program. This amounts to replacing the subproblem

$$\bar{h}(x) := \max_{\zeta\in Z} \sum_{l=1}^{k} \zeta_l (a_l^\mathsf{T} x - b_l) \ \le\ -(a_0^\mathsf{T} x - b_0) \tag{13.12}$$

in (13.11) by a finite set of convex constraints involving x and possibly the dual variable y of the subproblem but not the uncertain parameter ζ. The next theorem presents three uncertainty sets Z that lead to tractable reformulations of the problem (13.11).

Theorem 13.1 (Tractable robust LP)**.** Consider the robust linear program (13.11).

1 *Linear uncertainty*. Suppose $Z := \{\zeta \in \mathbb{R}^k : \|\zeta\|_\infty \le 1\}$. Then (13.11) is equivalent

to the LP:

$$\min_{(x,y)\in\mathbb{R}^{n+k}} \; c^\mathsf{T}x \;\text{ s.t. }\; \sum_l y_l \;\le\; -(a_0^\mathsf{T}x - b_0), \;\; -y_l \;\le\; a_l^\mathsf{T}x - b_l \;\le\; y_l, \;\; l = 1,\ldots,k$$

2 *SOC uncertainty*. Suppose $Z := \{\,\zeta \in \mathbb{R}^k \;:\; \|\zeta\|_2 \le r\,\}$. Then (13.11) is equivalent to the SOCP:

$$\min_{x\in\mathbb{R}^n} \; c^\mathsf{T}x \;\text{ s.t. }\; r\sqrt{\sum_l \left(a_l^\mathsf{T}x - b_l\right)^2} \;\le\; -(a_0^\mathsf{T}x - b_0)$$

3 *Conic uncertainty*. Suppose

$$Z := \{\,\zeta \in \mathbb{R}^k \;:\; \exists u \in \mathbb{R}^p \;\text{ s.t. }\; P\zeta + Qu + d \in K\,\}$$

where K is a closed convex pointed cone in $\mathbb{R}^m$ with a nonempty interior, $P \in \mathbb{R}^{m\times k}$ and $Q \in \mathbb{R}^{m\times p}$ are given matrices, and $d \in \mathbb{R}^m$ is a given vector. Suppose Z is nonempty and

- Either K is a polyhedral cone or

$$\exists(\bar\zeta,\bar u) \in \mathbb{R}^{k+p} \quad \text{s.t.} \quad P\bar\zeta + Q\bar u + d \;\in\; \mathrm{ri}(K)$$

- For each $x \in \mathbb{R}^n$, the subproblem $\max_{\zeta\in Z} \sum_l \zeta_l \left(a_l^\mathsf{T}x - b_l\right)$ in (13.11) is finite.

Then X is represented by the set X^+ of $(x,y) \in \mathbb{R}^{n+m}$ defined by the following system of conic inequalities:

$$a_0^\mathsf{T}x + d^\mathsf{T}y \;\le\; b_0 \tag{13.13a}$$

$$y \;\in\; K^*, \quad Q^\mathsf{T}y \;=\; 0, \quad a_l^\mathsf{T}x + \left(P^\mathsf{T}y\right)_l \;=\; b_l, \quad l = 1,\ldots,k \tag{13.13b}$$

where $K^* := \{y \in \mathbb{R}^m \;:\; y^\mathsf{T}z \ge 0 \;\forall z \in K\}$ is the dual cone of K. The robust linear program (13.11) is equivalent to the conic program

$$\min_{(x,y)\in\mathbb{R}^{n+m}} \; c^\mathsf{T}x \;\text{ s.t. }\; (13.13)$$

The form Z for the conic uncertainty is common in applications and says that even though the full uncertain parameter is (ζ,u) (whose affine transformation is in K), only the subvector ζ affects the optimization (13.11). As we will see in the proof, (13.13b) is the feasibility condition for the dual of the subproblem $\max_{\zeta\in Z} \sum_l \zeta_l \left(a_l^\mathsf{T}x - b_l\right)$ in (13.11).

Proof For parts 1 and 2, see Exercise 13.4. For part 3 fix any $x \in \mathbb{R}^n$. Let $s \in \mathbb{R}^k$ be defined by $s_l := s_l(x) := a_l^\mathsf{T}x - b_l$. Then the subproblem (13.12) is the following conic program (12.69) studied in Chapter 12.8.4:

$$p^*(x) \;:=\; \max_{(\zeta,u)\in\mathbb{R}^{k+p}} \; s^\mathsf{T}(x)\zeta \;\text{ s.t. }\; \begin{bmatrix} P & Q \end{bmatrix} \begin{bmatrix} \zeta \\ u \end{bmatrix} + d \in K \tag{13.14a}$$

i.e., x is feasible for (13.11) if $p^*(x) \le b_0 - a_0^\mathsf{T}x$. We will show that this holds if and only if there exists $(x,y) \in \mathbb{R}^{n+m}$ that satisfies (13.13).

The Lagrangian of (13.14a) is

$$L(\zeta,u,y) := s^\mathsf{T}\zeta + y^\mathsf{T}\left(\begin{bmatrix}P & Q\end{bmatrix}\begin{bmatrix}\zeta\\ u\end{bmatrix} + d\right), \qquad (\zeta,u)\in\mathbb{R}^{k+p},\ y\in K^*$$

where K^* is the dual cone of K (see Chapter 12.8.4 for details). Since

$$L(\zeta,u,y) := y^\mathsf{T}d + \left(s^\mathsf{T} + y^\mathsf{T}P\right)\zeta + y^\mathsf{T}Qu$$

the dual function is

$$d(y) := \max_{(\zeta,u)\in\mathbb{R}^{k+p}} L(\zeta,u,y) \;=\; \begin{cases} d^\mathsf{T}y & \text{if } P^\mathsf{T}y = -s,\ Q^\mathsf{T}y = 0\\ \infty & \text{otherwise}\end{cases}$$

and the dual problem is:

$$d^*(x) \;:=\; \min_{y\in K^*}\ d^\mathsf{T}y \quad \text{s.t.}\quad P^\mathsf{T}y = -s(x),\ Q^\mathsf{T}y = 0 \tag{13.14b}$$

where the constraints above correspond to the stationarity condition $\nabla_{(\zeta,u)}L = 0$. For every $x\in\mathbb{R}^n$, since the Slater condition is satisfied and the optimal value $p^*(x)$ of (13.14a) is finite, Theorem 12.32 implies that strong duality holds and there exists $y(x)\in K^*$ that attains dual optimality, i.e., $p^*(x) = d^*(x) = d^\mathsf{T}y(x)$, whether or not primal optimality is attained.

Fix an $x\in\mathbb{R}^n$. Since strong duality holds, $p^*(x)\le b_0 - a_0^\mathsf{T}x$ will be equivalent to (13.13a) if and only if $y = y(x)$ in (13.13a) is dual optimal. We now show that a y is dual optimal if and only if (x,y) satisfies (13.13b). Since the Slater condition is satisfied, Theorem 12.32 implies that a feasible (ζ,u) is optimal for (13.14a) if and only if there exists $y\in K^*\subseteq\mathbb{R}^m$ such that (noting that our primal problem (13.14a) is maximization corresponding to minimizing $-s^\mathsf{T}\zeta$)

$$\begin{bmatrix}-s\\ 0\end{bmatrix} = \begin{bmatrix}P^\mathsf{T}\\ Q^\mathsf{T}\end{bmatrix}y, \qquad y^\mathsf{T}\left(\begin{bmatrix}P & Q\end{bmatrix}\begin{bmatrix}\zeta\\ u\end{bmatrix} + d\right) = 0 \tag{13.15}$$

The first condition in (13.15) is stationarity $\nabla_{\zeta,u}L(\zeta,u,y) = 0$ and the second complementary slackness. Moreover such an y is optimal for (13.14b). It hence suffices to show that (13.15) is equivalent to (13.13b). The complementary slackness condition in (13.15) involves the primal variables (ζ,u), but we claim that it is implied by the stationarity condition in (13.15) and strong duality ($y^\mathsf{T}d = s^\mathsf{T}\zeta$) and therefore can be omitted:

$$y^\mathsf{T}\left(\begin{bmatrix}P & Q\end{bmatrix}\begin{bmatrix}\zeta\\ u\end{bmatrix} + d\right) = y^\mathsf{T}P\zeta + y^\mathsf{T}Qu + y^\mathsf{T}d = -s^\mathsf{T}\zeta + 0 + y^\mathsf{T}d = 0$$

Next recall that $s_l := s_l(x) := a_0^\mathsf{T}x - b_l$ and hence $s + P^\mathsf{T}y = 0$ is equivalent to $a_l^\mathsf{T}x + (P^\mathsf{T}y)_l = b_l$, $l = 1,\ldots,k$. We have thus shown that y is dual optimal if and only if (x,y) satisfies (13.13). This completes the proof. □

Remark 13.2 (Derivation strategy)**.** The proof of Theorem 13.1 illustrates the strategy outlined in Chapter 13.1.1. For parts 1 and 2, the subproblem (13.8) is solved explicitly. The equivalent feasibility condition $\bar{h}(x)\le 0$ takes the convex form given in the theorem. For part 3 the subproblem (13.8) is convex but cannot be solved explicitly. □

Example 13.2 (Conic uncertainty set)**.** The conic uncertainty set in part 3 of Theorem 13.1

$$Z := \{\zeta \in \mathbb{R}^k : \exists u \in \mathbb{R}^p \text{ s.t. } P\zeta + Qu + d \in K\}$$

is very general and includes the linear uncertainty in part 1 and conic uncertainty in part 2 as special cases. Specifically part 3 reduces to part 1 when $K := \mathbb{R}^m_+$ is the nonnegative quadrant, $Q = 0$, $d = \mathbf{1}$ of size $2k$ and P is $2k \times k$ with $P_{ll} = -1$, $P_{(k+l)l} = 1$ and $P_{il} = 0$ if $i \neq l, k+l$, such that $(P\zeta + d)_l = 1 - \zeta_l$ and $(P\zeta + d)_{k+l} = \zeta_l + 1$. The uncertainty set of part 2 can be expressed as the intersection of that of part 3 and an affine set (see Exercise 13.5).

A particularly simple case is $Z := \{\zeta \in \mathbb{R}^k : \zeta \in K\}$ in which case the robust linear program (13.11) is equivalent to the following conic program:

$$\min_{(x,y)\in\mathbb{R}^{n+m}} c^\mathsf{T} x \qquad \text{s.t.} \qquad a_0^\mathsf{T} x \le b_0,\ a_l^\mathsf{T} x + y_l = b_l,\ y \in K^*,\ l = 1, \ldots, k$$

where the first inequality corresponds to the nominal system and the other inequalities correspond to uncertain perturbations. □

13.1.3 Robust second-order cone program

We study the robust counterpart of an uncertain second-order cone program studied in Chapters 8.4.4 and 12.8.3. It takes the form

$$\min_{x\in\mathbb{R}^n} \ c^\mathsf{T} x \ \text{ s.t. } \ \|A(\zeta)x + b(\zeta)\|_2 \le \alpha^\mathsf{T}(\zeta)x + \beta(\zeta), \ \ \forall \zeta \in Z \subseteq \mathbb{R}^k \tag{13.16a}$$

where $(A(\zeta), b(\zeta))$ and $(\alpha(\zeta), \beta(\zeta))$ are affine functions of ζ:

$$A(\zeta) := A_0 + \sum_{l=1}^k \zeta_l A_l \in \mathbb{R}^{m\times n}, \qquad b(\zeta) := b_0 + \sum_{l=1}^k \zeta_l b_l \in \mathbb{R}^m \tag{13.16b}$$

$$\alpha(\zeta) = \alpha_0 + \sum_{l=1}^k \zeta_l \alpha_l \in \mathbb{R}^n, \qquad \beta(\zeta) := \beta_0 + \sum_{l=1}^k \zeta_l \beta_l \in \mathbb{R} \tag{13.16c}$$

Hence x is feasible if the affine transformation of x defined by $(A(\zeta), b(\zeta), \alpha(\zeta), \beta(\zeta))$ is in the second-order cone in $\mathbb{R}^{m+1}$ for all ζ in an uncertainty set Z. The form of uncertainty in (13.16b)(13.16c) does not lose generality because with $k = mn$ and appropriate choices of $(A_l, b_l, \alpha_l, \beta_l)$ we can perturb each entry of $(A_l(\zeta), b_l(\zeta), \alpha_l(\zeta), \beta_l(\zeta))$ independently around its nominal value.

If $Z = \text{conv}(\zeta^1, \ldots, \zeta^p) \subseteq \mathbb{R}^k$ then these constraints are equivalent to a set of p second-order cone constraints

$$\|A(\zeta^i)x + b(\zeta^i)\|_2 \le \alpha^\mathsf{T}(\zeta^i)x + \beta(\zeta^i), \qquad i = 1, \ldots, p$$

Otherwise, (13.16) is generally a semi-infinite set of constraints. Writing (13.16) as a

bi-level problem:

$$\min_{x\in\mathbb{R}^n} \; c^{\mathsf{T}}x \qquad \text{s.t.} \qquad \max_{\zeta\in Z} h(x,\zeta) \;\le\; 0$$

It can be easily shown that, for any fixed $x \in \mathbb{R}^n$, the constraint $h(x,\zeta) \le 0$ can be written as a SOC constraint, and hence convex, in ζ (Exercise 13.7):

$$\left\|\hat{A}(x)\zeta \;+\; \hat{b}(x)\right\|_2 \le \hat{\alpha}^{\mathsf{T}}(x)\,\zeta \;+\; \hat{\beta}(x), \qquad \forall \zeta \in Z \tag{13.17}$$

for some $\hat{A}(x) \in \mathbb{R}^{m\times k}$, $\hat{b}(x) \in \mathbb{R}^m$, $\hat{\alpha}(x) \in \mathbb{R}^k$, $\hat{\beta}(x) \in \mathbb{R}$. In particular $\hat{\beta}(x) := \alpha_0^{\mathsf{T}}x + \beta_0$ which will be used in Theorem 13.2. This means that the maximization of the convex $h(x,\cdot)$ over Z, and hence robust SOCP (13.16), is generally computationally hard except for special Z, e.g., $Z = \mathrm{conv}(\zeta^1,\ldots,\zeta^p)$. We now present two other classes of Z with decoupled uncertainties for which (13.16) is a tractable problem.

Suppose the dependence on the uncertain parameter $\zeta := (\zeta^{\mathrm{l}},\zeta^{\mathrm{r}}) \in Z^{\mathrm{l}}\times Z^{\mathrm{r}}$ in (13.16) is decoupled in that the left-hand side depends only on ζ^{l} and the right-hand side depends only on ζ^{r}. Specifically consider the robust SOCP:

$$\min_{x\in\mathbb{R}^n} \; c^{\mathsf{T}}x \quad \text{s.t.} \quad \left\|A(\zeta^{\mathrm{l}})x + b(\zeta^{\mathrm{l}})\right\|_2 \le \alpha^{\mathsf{T}}(\zeta^{\mathrm{r}})x + \beta(\zeta^{\mathrm{r}}), \;\; \forall \zeta^{\mathrm{l}} \in Z^{\mathrm{l}},\; \zeta^{\mathrm{r}} \in Z^{\mathrm{r}} \tag{13.18}$$

where $A(\zeta^{\mathrm{l}}) \in \mathbb{R}^{m\times n}$, $b(\zeta^{\mathrm{l}}) \in \mathbb{R}^m$, $\alpha(\zeta^{\mathrm{r}}) \in \mathbb{R}^n$ and $\beta(\zeta^{\mathrm{r}}) \in \mathbb{R}$. An $x \in \mathbb{R}^n$ is feasible for (13.18) if and only if there exists a variable τ such that

$$\max_{\zeta^{\mathrm{l}}\in Z^{\mathrm{l}}} \|A(\zeta^{\mathrm{l}})x + b(\zeta^{\mathrm{l}})\|_2 \;\le\; \tau \;\le\; \min_{\zeta^{\mathrm{r}}\in Z^{\mathrm{r}}} \alpha^{\mathsf{T}}(\zeta^{\mathrm{r}})x + \beta(\zeta^{\mathrm{r}}) \tag{13.19}$$

Fix any $x \in \mathbb{R}^n$. The semi-infinite constraint on x is tractable if both subproblems in (13.19):

$$\max_{\zeta^{\mathrm{l}}\in Z^{\mathrm{l}}} \|A(\zeta^{\mathrm{l}})x + b(\zeta^{\mathrm{l}})\|_2 \;\le\; \tau \tag{13.20a}$$

$$\min_{\zeta^{\mathrm{r}}\in Z^{\mathrm{r}}} \alpha^{\mathsf{T}}(\zeta^{\mathrm{r}})x + \beta(\zeta^{\mathrm{r}}) \ge \tau \tag{13.20b}$$

have finite convex representations. We discuss two classes of $(Z^{\mathrm{l}},Z^{\mathrm{r}})$ for which this is the case. In both cases, to maintain convexity of both subproblems, their objective functions are affine in ζ^{l} and ζ^{r} respectively. The feasible set Z^{l} for the maximization is affine in ζ^{l} and the feasible set Z^{r} for the minimization is defined by conic constraints. Even though the form (13.16b)(13.16c) of the uncertain parameters $(A(\zeta^{\mathrm{l}}),b(\zeta^{\mathrm{l}}),\alpha(\zeta^{\mathrm{r}}),\beta(\zeta^{\mathrm{r}}))$ is general, it is sometimes convenient to allow them to take other forms.

Interval + conic uncertainty.

Consider the robust SOCP (13.18). Suppose that:

1 *Left-side uncertainty*: $A(\zeta^{\mathrm{l}}) = A_0 + \Delta A \in \mathbb{R}^{m\times n}$ and $b(\zeta^{\mathrm{l}}) = b_0 + \Delta b \in \mathbb{R}^m$ with the uncertainty set

$$Z^{\mathrm{l}} \;:=\; \left\{\zeta^{\mathrm{l}} := [\Delta A \;\Delta b] : |\Delta A_{ij}| \le \delta_{ij},\; |\Delta b_i| \le \delta_i,\; \forall i,j\right\} \tag{13.21a}$$

i.e., each parameter $[A(\zeta^{\mathrm{l}})]_{ij}$, $[b(\zeta^{\mathrm{l}})]_i$ is perturbed independently of other parameters around its nominal value.[2] The first subproblem (13.20a) then becomes:

$$\max_{[\Delta A\ \Delta b]\in Z^{\mathrm{l}}} \|(A_0x+b_0)+(\Delta Ax+\Delta b)\|_2 \ \le\ \tau$$

2 *Right-side uncertainty*: $\alpha(\zeta^{\mathrm{r}}) := \alpha_0+\sum_{l=1}^{k_{\mathrm{r}}}\zeta_l\alpha_l \in \mathbb{R}^n$ and $\beta(\zeta^{\mathrm{r}}) := \beta_0+\sum_{l=1}^{k_{\mathrm{r}}}\zeta_l\beta_l \in \mathbb{R}$ with the uncertain parameter ζ^{r} in the conic uncertainty set in Theorem 13.1:

$$Z^{\mathrm{r}} \ :=\ \left\{\zeta^{\mathrm{r}}\in\mathbb{R}^{k_{\mathrm{r}}} : \exists u \text{ s.t. } P\zeta^{\mathrm{r}}+Qu+d\in K\right\} \tag{13.21b}$$

where $K\subseteq\mathbb{R}^p$ is a closed convex pointed cone for some P,Q,d,u of appropriate dimensions. Suppose Z^{r} satisfies the Slater condition, i.e., Z^{r} is nonempty and either K is polyhedral or there is $(\bar\zeta^{\mathrm{r}},\bar u)$ such that $P\bar\zeta^{\mathrm{r}}+Q\bar u+d\in\mathrm{ri}(K)$. The second subproblem (13.20b) then becomes:

$$\min_{\zeta^{\mathrm{r}}\in Z^{\mathrm{r}},\,t\in\mathbb{R}}\left\{t : (\alpha_0^{\mathsf{T}}x+\beta_0)+\sum_{l=1}^{k_{\mathrm{r}}}(\alpha_l^{\mathsf{T}}x+\beta_l)\zeta_l\le t\right\} \ \ge\ \tau$$

Fix an $x\in\mathbb{R}^n$. The first subproblem is of the form

$$\max_{\Delta w:|\Delta w_i|\le\epsilon_i}\|w+\Delta w\|_2^2 \ =\ \sum_i \max_{\Delta w_i:|\Delta w_i|\le\epsilon_i}(w_i+\Delta w_i)^2 \le \tau^2$$

which can be solved in closed form. Since Z^{l} is a simple box constraint, the maximum value of each term is $(|w_i|+\epsilon_i)^2$ and is attained at $\Delta w_i=\pm\epsilon_i$. Hence the first subproblem (13.20a) is equivalent to: $\exists z\in\mathbb{R}^m$ such that

$$z_i = \left|\sum_j [A_0]_{ij}x_j+[b_0]_i\right| + \sum_j|\delta_{ij}x_j|+\delta_i,\ \ i=1,\ldots,m,\quad \|z\|_2\ \le\ \tau$$

which is a linear constraint and a convex quadratic constraint in $z\in\mathbb{R}^m$. This leads to the constraint (13.23a) in Theorem 13.2 below. Rewrite the minimization in the second subproblem for the right-side uncertainty as:

$$\min_{\zeta^{\mathrm{r}},t,u}\ t\quad\text{s.t.}\quad \hat\alpha^{\mathsf{T}}(x)\zeta^{\mathrm{r}}+\hat\beta(x)-t\le 0,\ P\zeta^{\mathrm{r}}+Qu+d\in K\subseteq\mathbb{R}^p \tag{13.22}$$

where $\hat\alpha_l(x) := \alpha_l^{\mathsf{T}}x+\beta_l$ and $\hat\beta(x) := \alpha_0^{\mathsf{T}}x+\beta_0$. This is a convex problem similar to the problem (13.14a) in the proof of Theorem 13.1, with an additional affine constraint. The condition (13.20b) can therefore be characterized in the same way as in Theorem 13.1, leading to the constraint (13.23b) in the next theorem. The theorem shows that the robust SOCP (13.18) where $(Z^{\mathrm{l}},Z^{\mathrm{r}})$ are given by (13.21) is a conic program and hence tractable. It can be proved using Theorem 12.32, similarly for Theorem 13.1 (Exercise 13.8).

Theorem 13.2 (Tractable SOCP)**.** Consider the robust SOCP (13.18) where $(Z^{\mathrm{l}},Z^{\mathrm{r}})$ are given by (13.21) where Z^{r} satisfies the Slater condition. Suppose the minimum

[2] If uncertainty is expressed in the form of (13.16b)(13.16c), this corresponds to $\left|\sum_l\zeta_l[A_l]_{ij}\right|\le\delta_{ij}$, $|\sum_l\zeta_l[b_l]_i|\le\delta_i$.

value in (13.20b) is finite. Then $x \in \mathbb{R}^n$ is feasible for (13.18) if and only if there exist $(y,z) \in \mathbb{R}^{p+m}$ such that (x,y,z) satisfies

$$z_i = \left| \sum_j [A_0]_{ij} x_j + [b_0]_i \right| + \sum_j \delta_{ij} |x_j| + \delta_i, \quad i = 1,\ldots,m \tag{13.23a}$$

$$\|z\|_2 \le \hat{\beta}(x) - y^{\mathsf{T}} d, \quad y \in K^*, \quad P^{\mathsf{T}} y = \hat{\alpha}(x), \quad Q^{\mathsf{T}} y = 0 \tag{13.23b}$$

where $K^* \subseteq \mathbb{R}^p$ is the dual cone of K, $\hat{\alpha}_l(x) := \alpha_l^{\mathsf{T}} x + \beta_l$ and $\hat{\beta}(x) := \alpha_0^{\mathsf{T}} x + \beta_0$. Hence (13.18) is equivalent to the conic program:

$$\min_{(x,y,z) \in \mathbb{R}^{n+p+m}} c^{\mathsf{T}} x \quad \text{s.t.} \quad (13.23)$$

Bounded ℓ_2 norm + conic uncertainty.

Consider the robust SOCP (13.18). Suppose that:

1 *Left-side uncertainty*: $A(\zeta^{\mathrm{l}})x + b(\zeta^{\mathrm{l}})$ takes the form

$$A(\zeta^{\mathrm{l}})x + b(\zeta^{\mathrm{l}}) = (A_0 x + b_0) + L^{\mathsf{T}}(x)\zeta^{\mathrm{l}} r(x) \tag{13.24}$$

where $A(\zeta^{\mathrm{l}}) \in \mathbb{R}^{m \times n}$, $b(\zeta^{\mathrm{l}}) \in \mathbb{R}^m$, $L(x) \in \mathbb{R}^{k_1 \times m}$, $\zeta^{\mathrm{l}} \in \mathbb{R}^{k_1 \times k_2}$, $r(x) \in \mathbb{R}^{k_2}$. The first term $A_0 x + b_0$ is the nominal value and the second term $L^{\mathsf{T}}(x)\zeta^{\mathrm{l}} r(x)$ is the perturbation due to the uncertain matrix ζ^{l}. We impose the restriction that at most one of $L(x)$ and $r(x)$ depends on x and the other is a constant (see (13.27) below). Moreover the dependence of $L(x)$ or $r(x)$ is affine in x so that the constraints in (13.28b) and (13.28c) below are linear matrix inequalities in x. The uncertain parameter ζ^{l} is a matrix of bounded induced norm (maximum singular value) in the uncertainty set

$$Z^{\mathrm{l}} := \left\{ \zeta^{\mathrm{l}} \in \mathbb{R}^{k_1 \times k_2} : \left\|\zeta^{\mathrm{l}}\right\|_2 := \max_{u:\|u\|_2 \le 1} \left\|\zeta^{\mathrm{l}} u\right\|_2 \le 1 \right\} \tag{13.25}$$

The first subproblem (13.20a) then becomes:

$$\max_{\zeta^{\mathrm{l}} \in Z^{\mathrm{l}}} \|(A_0 x + b_0) + L^{\mathsf{T}}(x)\zeta^{\mathrm{l}} r(x)\|_2 \le \tau \tag{13.26}$$

2 *Right-side uncertainty*: Z^{r} is given by (13.21b) and satisfies the Slater condition.

Since Z^{r} is the same as that in Theorem 13.2, the second subproblem (13.20b) can be characterized in the same way, leading to the constraint (13.28a) in Theorem 13.3. We will show that the first subproblem (13.26) is equivalent to an explicit system of linear matrix inequalities (LMIs) (13.28b)(13.28c) in Theorem 13.3. They imply that robust SOCP (13.18) with bounded-norm and conic uncertainty is equivalen to a semidefinite program. We separate explicitly (13.24) into two cases:

$$A(\zeta^{\mathrm{l}})x + b(\zeta^{\mathrm{l}}) = (A_0 x + b_0) + L^{\mathsf{T}}(x)\zeta^{\mathrm{l}} r \tag{13.27a}$$

where $L(x)$ a matrix affine in x and $r \neq 0$ is a constant vector and

$$A(\zeta^{\mathrm{l}})x + b(\zeta^{\mathrm{l}}) = (A_0 x + b_0) + L^{\mathsf{T}}\zeta^{\mathrm{l}} r(x) \tag{13.27b}$$

where $L \neq 0$ is a constant matrix and $r(x)$ is a vector affine in x.

Theorem 13.3. Consider the robust SOCP (13.18) where Z^{l} is given by (13.25)(13.27) and Z^{r} is given by (13.21b). Suppose the minimum value in (13.26) is finite and Z^{r} satisfies the Slater condition. An $x \in \mathbb{R}^n$ is feasible for (13.18) if and only if there exist $y \in \mathbb{R}^p$ and $(\tau, \lambda) \in \mathbb{R}^2$ such that (x, y, τ, λ) satisfies

$$y \in K^*, \quad \tau \leq \hat{\beta}(x) - y^{\mathsf{T}} d, \quad P^{\mathsf{T}} y = \hat{\alpha}(x), \quad Q^{\mathsf{T}} y = 0 \tag{13.28a}$$

with $\hat{\alpha}_l(x) := \alpha_l^{\mathsf{T}} x + \beta_l$ and $\hat{\beta}(x) := \alpha_0^{\mathsf{T}} x + \beta_0$, and the following linear matrix inequalities:

- when $A(\zeta^{\mathrm{l}})x + b(\zeta^{\mathrm{l}})$ is given by (13.27a):

$$\lambda \geq 0, \qquad \begin{bmatrix} \tau - \lambda \|r\|_2^2 & (A_0 x + b_0)^{\mathsf{T}} & 0 \\ A_0 x + b_0 & \tau \mathbb{I}_m & L^{\mathsf{T}}(x) \\ 0 & L(x) & \lambda \mathbb{I}_{k_1} \end{bmatrix} \succeq 0 \tag{13.28b}$$

- when $A(\zeta^{\mathrm{l}})x + b(\zeta^{\mathrm{l}})$ is given by (13.27b):

$$\lambda \geq 0, \qquad \begin{bmatrix} \tau & (A_0 x + b_0)^{\mathsf{T}} & r^{\mathsf{T}}(x) \\ A_0 x + b_0 & \tau \mathbb{I}_m - \lambda L^{\mathsf{T}} L & 0 \\ r(x) & 0 & \lambda \mathbb{I}_{k_2} \end{bmatrix} \succeq 0 \tag{13.28c}$$

Hence (13.18) is equivalent to the semidefinite program:

$$\min_{(x, y, \tau, \lambda) \in \mathbb{R}^{n+p+2}} c^{\mathsf{T}} x \quad \text{s.t.} \quad (13.28)$$

□

The subproblem (13.26) is the constraint $(A(\zeta^{\mathrm{l}})x + b(\zeta^{\mathrm{l}}), \tau) \in K_{\mathrm{soc}}$ for all $\zeta \in Z^{\mathrm{l}}$. The proof that this is equivalent to (13.28b)(13.28c) then uses the following three ideas:

1 *Second-order cone in terms of K_{sdp}.* A vector $(y, t) \in K_{\mathrm{soc}} \subseteq \mathbb{R}^{l+1}$, i.e., $\|y\|_2 \leq t$, if and only if

$$\begin{bmatrix} t & y^{\mathsf{T}} \\ y & t\mathbb{I}_\ell \end{bmatrix} \succeq 0 \tag{13.29}$$

where $\mathbb{I}_\ell$ is the identity matrix of size ℓ. This follows from the following property of the Schur complement of the "arrow matrix" in (13.29): a matrix is (necessarily symmetric and) positive definite if both a principal submatrix and the Schur complement of the principal submatrix are positive definite (see Theorem A.4 in Chapter A.3.1).

2 *l_2-norm matrix minimization*. It can be proved using singular-value decomposition that (Exercise 13.9)

$$-\rho\|a_1\|_2\|a_2\|_2 \;=\; \min_{X:\|X\|_2\le\rho} a_1^{\mathsf{T}} X a_2 \tag{13.30}$$

3 *S-lemma*. Let A, B be symmetric matrices of the same size and $\bar{x}^{\mathsf{T}} A\bar{x} > 0$ for some $\bar{x}$. Then the implication $x^{\mathsf{T}} Ax \ge 0 \Rightarrow x^{\mathsf{T}} Bx \ge 0$ holds if and only if $\exists \lambda \ge 0$ such that $B \ge \lambda A$. Note that neither B nor A needs to be positive semidefinite, but $B - \lambda A$ is. This lemma is proved in Chapter 13.1.5. (The result originates from stability analysis of nonlinear systems and hence S in S-lemma.)

Proof of Theorem 13.3 Fix an $x \in \mathbb{R}^n$. It is feasible for (13.18) if and only if there exists a variable $\tau \in \mathbb{R}$ such that both subproblems in (13.20) have finite convex representations. Since Z^{r} is the same as that in Theorem 13.2 the second subproblem (13.20b) is equivalent to (13.28a). We now show that the first subproblem (13.20a), or (13.26), is equivalent to (13.28b)(13.28c), using the three ideas above.

Consider the case (13.27a) and let $g(x) := A_0 x + b_0 \in \mathbb{R}^m$. First, apply (13.29) to write (13.26) as

$$\begin{bmatrix} \tau & \left(g(x) + L^{\mathsf{T}}(x)\zeta^{\mathrm{l}} r\right)^{\mathsf{T}} \\ g(x) + L^{\mathsf{T}}(x)\zeta^{\mathrm{l}} r & \tau \mathbb{I}_m \end{bmatrix} \;\ge\; 0, \qquad \zeta^{\mathrm{l}} \in Z^{\mathrm{l}}$$

Therefore

$$(z_1)^2\tau + 2z_2^{\mathsf{T}}\left(g(x) + L^{\mathsf{T}}(x)\zeta^{\mathrm{l}} r\right) z_1 + (z_2^{\mathsf{T}} z_2)\tau \;\ge\; 0, \quad \forall\, z_1 \in \mathbb{R},\; z_2 \in \mathbb{R}^m,\; \zeta^{\mathrm{l}} \in Z^{\mathrm{l}}$$

Or, for all $z_1 \in \mathbb{R}$ and $z_2 \in \mathbb{R}^m$,

$$(z_1)^2\tau + 2z_2^{\mathsf{T}} g(x) z_1 + (z_2^{\mathsf{T}} z_2)\tau + \min_{\zeta^{\mathrm{l}}:\|\zeta^{\mathrm{l}}\|_2\le 1} (2L(x)z_2)^{\mathsf{T}} \zeta^{\mathrm{l}}(z_1 r) \;\ge\; 0 \tag{13.31}$$

Second, use (13.30) twice to eliminate ζ^{l}:

$$\min_{\zeta^{\mathrm{l}}:\|\zeta^{\mathrm{l}}\|_2\le 1} (2L(x)z_2)^{\mathsf{T}} \zeta^{\mathrm{l}}(z_1 r) \;=\; -2\|L(x)z_2\|_2\|z_1 r\|_2 \;=\; \min_{X:\|X\|_2\le\|z_1 r\|_2} (2L(x)z_2)^{\mathsf{T}} X \tag{1}$$

where the second equality uses (13.30) with $X \in \mathbb{R}^{k_1\times 1}$. Substituting into (13.31) we have, for all $z_1 \in \mathbb{R}$, $z_2 \in \mathbb{R}^m$ and $X \in \mathbb{R}^{k_1}$, if $z_1^2\|r\|_2^2 - X^{\mathsf{T}} X \ge 0$ then

$$(z_1)^2\tau \;+\; 2z_2^{\mathsf{T}} g(x) z_1 \;+\; (z_2^{\mathsf{T}} z_2)\tau \;+\; 2X^{\mathsf{T}} L(x) z_2 \;\ge\; 0$$

This is equivalent to: for $(z_1, z_2, X) \in \mathbb{R}^{1+m+k_1}$

$$\begin{bmatrix} z_1 \\ z_2 \\ X \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \|r\|_2^2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -\mathbb{I}_{k_1} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ X \end{bmatrix} \ge 0 \quad \Rightarrow \quad \begin{bmatrix} z_1 \\ z_2 \\ X \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \tau & g^{\mathsf{T}}(x) & 0 \\ g(x) & \tau\mathbb{I}_m & L^{\mathsf{T}}(x) \\ 0 & L(x) & 0 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ X \end{bmatrix} \ge 0$$

Third, there clearly exists $z_1 > 0$ such that $z_1^2\|r\|_2^2 > 0$ since $r \ne 0$. Hence we can apply the S-lemma to the two $(1+m+k_1)\times(1+m+k_1)$ matrices above to conclude that (13.20a) is equivalent to (13.28b).

The case of (13.27b) is similar. Applying (13.29) to write (13.26) as for all $\zeta^1 \in Z^1$,

$$\begin{bmatrix} \tau & \big(g(x) + L^\mathsf{T}\zeta^1 r(x)\big)^\mathsf{T} \\ g(x) + L^\mathsf{T}\zeta^1 r(x) & \tau \mathbb{I}_m \end{bmatrix} \succeq 0$$

Therefore, for all $z_1 \in \mathbb{R}$ and $z_2 \in \mathbb{R}^m$,

$$(z_1)^2\tau + 2z_2^\mathsf{T} g(x) z_1 + (z_2^\mathsf{T} z_2)\tau + \min_{\zeta^1 : \|\zeta^1\|_2 \le 1} (2Lz_2)^\mathsf{T}\zeta^1 (z_1 r(x)) \;\ge\; 0 \tag{13.32}$$

Use (13.30) twice to eliminate ζ^1 ((13.27a) and (13.27b) differ mainly in the second equality below):

$$\min_{\zeta^1 : \|\zeta^1\|_2 \le 1} (2Lz_2)^\mathsf{T}\zeta^1 (z_1 r(x)) \;=\; -2\|Lz_2\|_2 \|z_1 r(x)\|_2 \;=\; \min_{X : \|X\|_2 \le \|Lz_2\|_2} (2z_1 r(x))^\mathsf{T} X(1)$$

where $X \in \mathbb{R}^{k_2 \times 1}$. Substituting into (13.32) we have, for all $z_1 \in \mathbb{R}$, $z_2 \in \mathbb{R}^m$ and $X \in \mathbb{R}^k_2$, if $z_2^\mathsf{T}(L^\mathsf{T}L)z_2 - X^\mathsf{T}X \ge 0$ then

$$(z_1)^2\tau + 2z_2^\mathsf{T} g(x) z_1 + (z_2^\mathsf{T} z_2)\tau + 2X^\mathsf{T} r(x) z_1 \;\ge\; 0$$

This is equivalent to: for $(z_1, z_2, X) \in \mathbb{R}^{1+m+k_2}$

$$\begin{bmatrix} z_1 \\ z_2 \\ X \end{bmatrix}^\mathsf{T} \begin{bmatrix} 0 & 0 & 0 \\ 0 & L^\mathsf{T}L & 0 \\ 0 & 0 & -\mathbb{I}_{k_2} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ X \end{bmatrix} \ge 0 \quad \Rightarrow \quad \begin{bmatrix} z_1 \\ z_2 \\ X \end{bmatrix}^\mathsf{T} \begin{bmatrix} \tau & g^\mathsf{T}(x) & r^\mathsf{T}(x) \\ g(x) & \tau\mathbb{I}_m & 0 \\ r(x) & 0 & 0 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ X \end{bmatrix} \ge 0$$

Hence we can apply the S-lemma to the two $(1+m+k_2)\times(1+m+k_2)$ matrices above to conclude that (13.20a) is equivalent to (13.28c). □

13.1.4 Robust semidefinite program

We study the robust counterpart of an uncertain semidefinite program (SDP) studied in Chapter 8.4.5. Consider a standard SDP

$$\min_{x \in \mathbb{R}^n} \quad f(x) \qquad \text{s.t.} \qquad h_0(x) \in K_{\text{psd}} \tag{13.33a}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a real-valued convex function, $K_{\text{psd}} \subseteq \mathbb{S}^m$ is the closed convex pointed cone of positive semidefinite matrices in the vector space $\mathbb{S}^m \subset \mathbb{R}^{m\times m}$ of symmetric matrices, and the matrix-valued function $h_0 : \mathbb{R}^n \to \mathbb{S}^m$ is given by:

$$h_0(x) \;:=\; B_0 + \sum_{i=1}^n x_i A_0^i \;\in\; \mathbb{S}^m \tag{13.33b}$$

where $A_0^i, B_0 \in \mathbb{S}^m$ are given symmetric matrices for $i = 0, 1, \ldots, n$. This is the nominal problem where the parameters $(A_0^i, B_0, i \ge 0)$ that define h_0 are certain and given.

The robust counterpart of (13.33) is

$$\min_{x \in \mathbb{R}^n} \quad f(x) \qquad \text{s.t.} \qquad h_0(x) + h(x, \zeta) \in K_{\text{psd}}, \qquad \forall \zeta \in Z \tag{13.34a}$$

where $h_0(x)$ is given by (13.33b), $h(x,\zeta)$ is a symmetric matrix in $\mathbb{S}^m$ as a function of x indexed by ζ, and ζ is the uncertain parameter that takes value in an uncertainty set Z. We assume that the matrix-valued function $h(x,\zeta)$ is an affine function of x for each fixed $\zeta \in Z$ so that the constraints in (13.34a) are linear matrix inequalities (LMIs) in x. For example $h(x,\zeta)$ may take the form:

$$h(x,\zeta) := \sum_{l=1}^{k} \zeta_l \left(B_l + \sum_{i=1}^{n} x_i A_l^i \right) \in \mathbb{S}^m, \qquad \forall \zeta \in Z \subseteq \mathbb{R}^k$$

for a given set of symmetric matrices $(A_l^i, B_l,\ i = 1,\ldots,n,\ l = 1,\ldots,k)$ in $\mathbb{S}^m$. For a general uncertainty set Z, it is a semi-infinite set of LMIs and hence the robust SDP (13.34a) is generally computationally intractable. There are two exceptions. The first is when $Z := \mathrm{conv}(\zeta^1,\ldots,\zeta^p)$ is the convex hull of p given vectors $\zeta^1,\ldots,\zeta^p \in \mathbb{R}^k$. In this case the semi-infinite set of LMIs reduces to a set of p LMIs and the robust SDP (13.34a) reduces to the following convex problem

$$\min_{x\in\mathbb{R}^n} \quad f(x) \qquad \text{s.t.} \qquad h_0(x) + h(x,\zeta^i) \in K_{\mathrm{psd}}, \qquad i = 1,\ldots,p$$

for any affine functions $h_{\zeta^i}(x)$ of x, indexed by $\zeta^1,\ldots,\zeta^p \in Z$.

The second exception is when the affine function $h(x,\zeta)$ is given by

$$h(x,\zeta) := L^{\mathsf{T}}(x)\zeta R(x) + R^{\mathsf{T}}(x)\zeta^{\mathsf{T}} L(x) \in \mathbb{S}^m \tag{13.34b}$$

where ζ is a $k_1 \times k_2$ matrix with bounded spectral norm in the uncertainty set

$$Z := \left\{ \zeta \in \mathbb{R}^{k_1\times k_2} \ :\ \|\zeta\|_2 := \max_{u:\|u\|_2=1} \|\zeta u\|_2 \ \le\ \rho \right\} \tag{13.34c}$$

and both $L(x) \in \mathbb{R}^{k_1\times m}$ and $R(x) \in \mathbb{R}^{k_2\times m}$ are affine functions of x with at least one of them being independent of x so that (13.34b) is an LMI (cf. the left-side uncertainty set in (13.25)(13.27) for robust SOCP). The semi-infinite contraint in (13.34a) is then:

$$h_0(x) + L^{\mathsf{T}}(x)\zeta R + R^{\mathsf{T}}\zeta^{\mathsf{T}} L(x) \ \in\ K_{\mathrm{psd}}, \quad \forall \zeta \in Z$$

Example 13.3 (SDP relaxation of OPF)**.** For notational simplicity we will formulate our problem in the complex domain, i.e., $\mathbb{S}^m$ is the set of Hermitian matrices and K_{psd} is the closed convex pointed cone of semidefinite matrices in the vector space $\mathbb{S}^m$ over the field $\mathbb{R}$ (not $\mathbb{C}$). It can be converted to the real domain (see Remark 9.2).

The semidefinite relaxation (10.20a) in Chapter 10.1 of optimal power flow (OPF) (9.16) is given by (omitting line flow constraints for simplicity):

$$\min_{W\in K_{\mathrm{psd}}} \mathrm{tr}\,(C_0 W) \ \text{ s.t. } \ \mathrm{tr}\,(\Phi_j W) \le p_j^{\max}, \quad -\mathrm{tr}\,(\Phi_j W) \le -p_j^{\min}, \quad j \in \overline{N} \tag{13.35a}$$

$$\mathrm{tr}\,(\Psi_j W) \le q_j^{\max}, \quad -\mathrm{tr}\,(\Psi_j W) \le -q_j^{\min}, \quad j \in \overline{N} \tag{13.35b}$$

$$\mathrm{tr}\,(J_j W) \le v_j^{\max}, \quad -\mathrm{tr}\,(J_j W) \le -v_j^{\min}, \quad j \in \overline{N} \tag{13.35c}$$

where $K_{\mathrm{psd}} \subset \mathbb{S}^{N+1}$,

$$\Phi_j := \frac{1}{2}\left(Y_0^{\mathsf{H}} e_j e_j^{\mathsf{T}} + e_j e_j^{\mathsf{T}} Y_0\right), \ \ \Psi_j := \frac{1}{2\mathbf{i}}\left(Y_0^{\mathsf{H}} e_j e_j^{\mathsf{T}} - e_j e_j^{\mathsf{T}} Y_0\right), \ \ J_j := e_j e_j^{\mathsf{T}} \tag{13.35d}$$

and $e_j \in \{0,1\}^{N+1}$ is the unit vector with a single 1 in its jth entry. Here $Y_0 \in \mathbb{C}^{(N+1)\times(N+1)}$ is a given nominal admittance matrix. This problem is of the form (8.73), reproduced here:

$$\min_{Z \in K_{\text{psd}}} \operatorname{tr}\left(B_0^{\mathsf{H}} Z\right) \quad \text{s.t.} \quad \operatorname{tr}\left(A_0^{i\mathsf{H}} Z\right) \le c_i,\ i = 1,\ldots,n := 6(N+1)$$

for some $B_0, A_0^i \in \mathbb{S}^{N+1}$, $i \ge 1$. The dual problem of (13.35) is therefore of the form (from (8.74b)):

$$-\min_{x \in \mathbb{R}^n} c^{\mathsf{T}} x \quad \text{s.t.} \quad x \ge 0,\ h_0(x) \in K_{\text{psd}} \tag{13.36a}$$

where $h_0(x) \in \mathbb{C}^{(N+1)\times(N+1)}$ for each $x \in \mathbb{R}^{6(N+1)}$, defined as:

$$h_0(x) := C_0 + \sum_{i=1}^{N+1} \left((x_{2i-1} - x_{2i})\, \Phi_i + \left(x_{2(N+1)+2i-1} - x_{2(N+1)+2i}\right) \Psi_i \right) \tag{13.36b}$$

$$+ \sum_{i=1}^{N+1} \left(x_{4(N+1)+2i-1} - x_{4(N+1)+2i}\right) J_i \tag{13.36c}$$

which takes the form of the nominal SDP problem (13.33).

Suppose the admittance matrix Y in (13.35d) is uncertain with $Y = Y_0 + \Delta Y$ where ΔY is the uncertain parameter that takes value in an uncertainty set $Z \subseteq \mathbb{C}^{(N+1)\times(N+1)}$. Let $\Delta\Phi_i := \left(\Delta Y^{\mathsf{H}} e_i e_i^{\mathsf{T}} + e_i e_i^{\mathsf{T}} \Delta Y\right)/2$ and $\Delta\Psi_i := \left(\Delta Y^{\mathsf{H}} e_i e_i^{\mathsf{T}} - e_i e_i^{\mathsf{T}} \Delta Y\right)/2\mathbf{i}$. Then the robust counterpart of (13.36) is

$$-\min_{x \in \mathbb{R}^n} c^{\mathsf{T}} x \quad \text{s.t.} \quad x \ge 0,\ h_0(x) + h(x, \Delta Y) \in K_{\text{psd}} \tag{13.37a}$$

where $h(x, \Delta Y) := L^{\mathsf{H}}(x)\Delta Y + \Delta Y^{\mathsf{H}} L(x)$ is a linear function in x and

$$L(x) := \sum_{i=1}^{N+1} \left(\frac{1}{2}\left(x_{2i-1} - x_{2i}\right) + \frac{1}{2\mathbf{i}} \left(x_{2(N+1)+2i-1} - x_{2(N+1)+2i}\right) \right) e_i e_i^{\mathsf{T}} \tag{13.37b}$$

If the perturbation ΔY has bounded spectral norm then this is the uncertainty model in (13.34) with $R(x) := \mathbb{I}_{N+1}$. □

The next result says that the robust semidefinite program (13.34) whose uncertain parameter ζ has a bounded spectral norm is computationally tractable.

Theorem 13.4. Consider the robust SDP (13.34).

1 If $h(x,\zeta) := L^{\mathsf{T}}(x)\zeta R + R^{\mathsf{T}}\zeta^{\mathsf{T}} L(x)$ with $R \ne 0$, then x is feasible for (13.34) if and only if there exists λ such that $(x,\lambda) \in \mathbb{R}^{n+1}$ satisfies

$$\lambda \ge 0, \qquad \begin{bmatrix} h_0(x) - \lambda R^{\mathsf{T}} R & \rho L^{\mathsf{T}}(x) \\ \rho L(x) & \lambda \mathbb{I}_{k_1} \end{bmatrix} \succeq 0 \tag{13.38a}$$

2 If $h(x,\zeta) := L^\mathsf{T}\zeta R(x) + R^\mathsf{T}(x)\zeta^\mathsf{T} L$ with $L \neq 0$, then x is feasible for (13.34) if and only if there exists λ such that $(x,\lambda) \in \mathbb{R}^{n+1}$ satisfies

$$\lambda \;\geq\; 0, \qquad \begin{bmatrix} h_0(x) - \lambda L^\mathsf{T} L & \rho R^\mathsf{T}(x) \\ \rho R(x) & \lambda \mathbb{I}_{k_2} \end{bmatrix} \;\geq\; 0 \tag{13.38b}$$

Hence the robust SDP (13.34) is equivalent to the semidefinite program:

$$\min_{(x,\lambda)\in\mathbb{R}^{n+1}} \quad f(x) \qquad \text{s.t.} \qquad (13.38)$$

Proof Suppose $h(x,\zeta) := L^\mathsf{T}(x)\zeta R + R^\mathsf{T}\zeta^\mathsf{T} L(x)$ with nonzero R. Fix an $x \in \mathbb{R}^n$. It is feasible for (13.34) if and only

$$y^\mathsf{T}\Big(h_0(x) + L^\mathsf{T}(x)\zeta R + R^\mathsf{T}\zeta^\mathsf{T} L(x)\Big)\, y \geq 0, \;\; \forall y \in \mathbb{R}^m, \; \forall \Big(\zeta \in \mathbb{R}^{k_1\times k_2} : \|\zeta\|_2 \leq \rho\Big)$$

Hence

$$y^\mathsf{T} h_0(x) y + 2 \min_{\zeta:\|\zeta\|_2\leq\rho} (L(x)y)^\mathsf{T}\zeta(Ry) \;\geq\; 0, \qquad \forall y \in \mathbb{R}^m \tag{13.39}$$

As in the proof of Theorem 13.3, apply (13.30) twice to eliminate ζ from (13.39):

$$\min_{\zeta:\|\zeta\|_2\leq\rho} (L(x)y)^\mathsf{T}\zeta(Ry) \;=\; -\rho\|L(x)y\|_2\|Ry\|_2 \;=\; \min_{X\in\mathbb{R}^{k_1}:\|X\|_2\leq\|Ry\|_2} (\rho L(x)y)^\mathsf{T} X(1) \tag{13.40}$$

Substituting into (13.39) we have

$$y^\mathsf{T}(R^\mathsf{T}R)y - X^\mathsf{T}X \geq 0 \implies y^\mathsf{T}h_0(x)y + 2y^\mathsf{T}(\rho L(x))^\mathsf{T}X \geq 0, \quad \forall (y,X) \in \mathbb{R}^{m+k_1}$$

This is equivalent to

$$\begin{bmatrix} R^\mathsf{T}R & 0 \\ 0 & -\mathbb{I}_{k_1} \end{bmatrix} \geq 0 \quad \implies \quad \begin{bmatrix} h_0(x) & \rho L^\mathsf{T}(x) \\ \rho L(x) & 0 \end{bmatrix}$$

Clearly there exists y such that $y^\mathsf{T}R^\mathsf{T}Ry > 0$ since R is nonzero. Hence we can apply the S-lemma to conclude (13.38a).

The case of $h(x,\zeta) := L^\mathsf{T}\zeta R(x) + R^\mathsf{T}(x)\zeta^\mathsf{T} L$ with nonzero L is similar. The main difference is that (13.40) becomes

$$\min_{\zeta:\|\zeta\|_2\leq\rho} (Ly)^\mathsf{T}\zeta(R(x)y) \;=\; -\rho\|Ly\|_2\|R(x)y\|_2 \;=\; \min_{X\in\mathbb{R}^{k_2}:\|X\|_2\leq\|Ly\|_2} (1)X^\mathsf{T}(\rho R(x)y)$$

Hence

$$y^\mathsf{T}(L^\mathsf{T}L)y - X^\mathsf{T}X \;\geq\; 0 \quad \implies \quad y^\mathsf{T}h_0(x)y + 2X^\mathsf{T}(\rho R(x))y \;\geq\; 0, \qquad \forall (y,X) \in \mathbb{R}^{m+k_2}$$

This is equivalent to

$$\begin{bmatrix} L^\mathsf{T}L & 0 \\ 0 & -\mathbb{I}_{k_2} \end{bmatrix} \geq 0 \quad \implies \quad \begin{bmatrix} h_0(x) & \rho R^\mathsf{T}(x) \\ \rho R(x) & 0 \end{bmatrix}$$

Then S-lemma implies (13.38b). □

13.1.5 Appendix: proof of S-lemma

Lemma 13.5 (S-lemma)**.** Let A, B be $n \times n$ symmetric matrices and $\bar{x}^\mathsf{T} A \bar{x} > 0$ for some $\bar{x} \in \mathbb{R}^n$. Then the following are equivalent:

(i) $x^\mathsf{T} A x \geq 0 \Rightarrow x^\mathsf{T} B x \geq 0$.
(ii) $\exists \lambda \geq 0$ such that $B \geq \lambda A$.

Proof Suppose (ii) holds. Then $x^\mathsf{T} B x - x^\mathsf{T} \lambda A x = x^\mathsf{T}(B - \lambda A)x \geq 0$, implying (i).

To prove (i) $\Rightarrow$ (ii), consider the following subsets of $\mathbb{R}^2$:

$$S := \left\{ \begin{bmatrix} x^\mathsf{T} A x \\ x^\mathsf{T} B x \end{bmatrix} \in \mathbb{R}^2 : x \in \mathbb{R}^n \right\}, \qquad T := \left\{ \begin{bmatrix} u \\ v \end{bmatrix} \in \mathbb{R}^2 : u \geq 0,\ v < 0 \right\}$$

Suppose (i) holds. We will establish (ii) in 4 steps:

1 Show that $S \cap T = \emptyset$.
2 Show that S is a cone.
3 Show that S is convex.
4 Use the Separating Hyperplane Theorem 8.11 of Chapter 8.2.4 to prove (ii).

The Slater condition $\bar{x}^\mathsf{T} A \bar{x} > 0$ in the lemma serves the same purpose as in the Slater theorem of ensuring that the separating hyperplane is not vertical. The result is illustrated in Figure 13.1. Let $u(x) := x^\mathsf{T} A x$ and $v(x) := x^\mathsf{T} B x$ for $x \in \mathbb{R}^n$. Then $(u(x), v(x)) \in S$

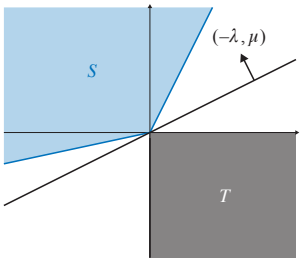


Figure 13.1 S-lemma: S, T and their separation.

by definition for all $x \in \mathbb{R}^n$.

1 $S \cap T = \emptyset$. Since (i) says $u(x) \geq 0 \Rightarrow v(x) \geq 0$, $(u(x), v(x)) \notin T$. On the other hand let $(a, b) \in T$, i.e., $a \geq 0$ and $b < 0$. If $(a, b) \in S$, then $a = u(x)$ and $b = v(x)$ for some x, and hence $a \geq 0$ and $b < 0$ contradicts (i). This shows that $S \cap T = \emptyset$.
2 *S is a cone*. Suppose $(u(x), v(x)) = (x^\mathsf{T} A x, x^\mathsf{T} B x) \in S$. For any $\lambda^2 > 0$ we have

$$\lambda^2 \begin{bmatrix} u(x) \\ v(x) \end{bmatrix} = \begin{bmatrix} (\lambda x)^\mathsf{T} A (\lambda x) \\ (\lambda x)^\mathsf{T} B (\lambda x) \end{bmatrix} = \begin{bmatrix} u(\lambda x) \\ v(\lambda x) \end{bmatrix}$$

i.e., $\lambda^2(u(x), v(x)) \in S$ and hence S is a cone.

3 *S is convex.* Let $y_1 := (u(x_1), v(x_1))$ and $y_2 := (u(x_2), v(x_2))$ be in S. Fix any $\alpha \in (0,1)$. We separate two cases.

- *Case 1: y_1, y_2 are linearly dependent.* Suppose $y_1 = cy_2$ for some $c \neq 0$. Then

$$\alpha y_1 + (1-\alpha)y_2 \;=\; (c\alpha + (1-\alpha))y_2 \;=\; \left(\frac{c\alpha + (1-\alpha)}{c}\right) y_1$$

  i.e., $\alpha y_1 + (1-\alpha)y_2$ is on the ray of y_1 and y_2 (which are on the same ray). It therefore must be in S, because if $c\alpha + (1-\alpha) > 0$ then $(c\alpha + (1-\alpha))y_2 \in S$ since S is a cone. If $c\alpha + (1-\alpha) < 0$ then both c and $c\alpha + (1-\alpha)$ must be negative and hence $((c\alpha + (1-\alpha))/c)y_1 \in S$ since S is a cone.

- *Case 2: y_1, y_2 are linearly independent.* We have to show that there exist $\bar{x} \in \mathbb{R}^n$ such that

$$\begin{bmatrix} u(\bar{x}) \\ v(\bar{x}) \end{bmatrix} \;=\; \alpha y_1 + (1-\alpha)y_2 \;=\; \alpha \begin{bmatrix} x_1^\mathsf{T} A x_1 \\ x_1^\mathsf{T} B x_1 \end{bmatrix} + (1-\alpha) \begin{bmatrix} x_2^\mathsf{T} A x_2 \\ x_2^\mathsf{T} B x_2 \end{bmatrix}$$

  which implies that $\alpha y_1 + (1-\alpha)y_2 \in S$. Since S is a cone it suffices to construct $\bar{x}$ such that, for some $\lambda > 0$,

$$\begin{bmatrix} u(\bar{x}) \\ v(\bar{x}) \end{bmatrix} \;=\; \lambda(\alpha y_1 + (1-\alpha)y_2) \tag{13.41}$$

  We will seek $\bar{x}$ of the form $\bar{x} = \alpha x_1 + \beta x_2$, i.e., we will derive $\beta \in \mathbb{R}$ such that (13.41) is satisfied for some $\lambda > 0$, given α, x_1, x_2. By definition

$$\begin{bmatrix} u(\bar{x}) \\ v(\bar{x}) \end{bmatrix} \;=\; \begin{bmatrix} (\alpha x_1 + \beta x_2)^\mathsf{T} A (\alpha x_1 + \beta x_2) \\ (\alpha x_1 + \beta x_2)^\mathsf{T} B (\alpha x_1 + \beta x_2) \end{bmatrix} \;=\; \begin{bmatrix} \alpha^2 u(x_1) + \beta^2 u(x_2) + 2\alpha\beta x_1^\mathsf{T} A x_2 \\ \alpha^2 v(x_1) + \beta^2 v(x_2) + 2\alpha\beta x_1^\mathsf{T} B x_2 \end{bmatrix}$$
$$\;=\; \alpha^2 y_1 + \beta^2 y_2 + 2\alpha\beta \begin{bmatrix} x_1^\mathsf{T} A x_2 \\ x_1^\mathsf{T} B x_2 \end{bmatrix}$$

  where the second equality uses the fact that $A^\mathsf{T} = A$ and $B^\mathsf{T} = B$. Since y_1, y_2 form a basis of $\mathbb{R}^2$ we can express

$$\begin{bmatrix} x_1^\mathsf{T} A x_2 \\ x_1^\mathsf{T} B x_2 \end{bmatrix} \;=:\; a y_1 + b y_2$$

  for some $a, b \in \mathbb{R}$. Therefore

$$\begin{bmatrix} u(\bar{x}) \\ v(\bar{x}) \end{bmatrix} \;=\; (\alpha^2 + 2\alpha\beta a) y_1 + (\beta^2 + 2\alpha\beta b) y_2 \;=\; (\alpha + 2a\beta) \left(\alpha y_1 + \frac{\beta^2 + 2\alpha b \beta}{\alpha + 2a\beta} y_2 \right)$$

  Substituting into (13.41) with $\lambda := \alpha + 2a\beta$, we therefore seek $\beta \in \mathbb{R}$ such that

$$\alpha + 2a\beta > 0, \quad \beta^2 + 2\alpha b \beta \;=\; (1-\alpha)(\alpha + 2a\beta) \tag{13.42}$$

  The quadratic equation in (13.42) is

$$\beta^2 + 2(\alpha b - (1-\alpha)a)\beta - \alpha(1-\alpha) \;=\; 0$$

with roots

$$\beta \;=\; -(\alpha b-(1-\alpha)a)\pm\sqrt{(\alpha b-(1-\alpha)a)^2+\alpha(1-\alpha)}$$

one of which is positive and the other negative since $\alpha\in(0,1)$. Choose the root β such that $a\beta\geq 0$. Then $\alpha+2a\beta>0$ and (13.42) is satisfied.

This completes the proof that S is convex.

4 Since S and T are convex and disjoint the Separating Hyperplane Theorem 8.11 of Chapter 8.2.4 implies there exists a nonzero $(-\lambda,\mu)\in\mathbb{R}^2$ such that

$$-\lambda u+\mu v \;\geq\; -\lambda a+\mu b, \qquad \forall (u,v)\in S,\ (a,b)\in T$$

Since $0\in S$ we have $-\lambda a+\mu b\leq 0$ for all $(a,b)\in T$. Taking $(a,b)\to 0$ we have $-\lambda u+\mu v\geq 0$ for all $(u,v)\in S$. Hence substituting $(u,v)=(x^{\mathsf{T}}Ax, x^{\mathsf{T}}Bx)$ we have

$$-\lambda x^{\mathsf{T}}Ax+\mu x^{\mathsf{T}}Bx \;\geq\; 0 \;\geq\; -\lambda a+\mu b, \qquad \forall x\in\mathbb{R}^n,\ (a,b)\in T \tag{13.43}$$

Taking $a=1$ and $b\to 0$ yields $\lambda\geq 0$. Taking $a=0$ yields $\mu\geq 0$ since $b<0$. If $\mu=0$ then $\lambda>0$ since $(-\lambda,\mu)\neq 0$. By assumption $\bar{x}^{\mathsf{T}}A\bar{x}>0$, implying that $-\lambda\bar{x}^{\mathsf{T}}A\bar{x}<0$, contradicting (13.43). Hence $\mu>0$ and we can normalize $(-\lambda,\mu)$ to become $(-\lambda,1)$ to obtain from (13.43)

$$x^{\mathsf{T}}(B-\lambda A)x \;\geq\; 0, \qquad \forall x\in\mathbb{R}^n$$

i.e., $B-\lambda A\geq 0$ for some $\lambda\geq 0$.

□

13.2 Chance constrained optimization

Consider the optimization problem:

$$\min_{x\in X\subseteq\mathbb{R}^n} \;\; c(x) \tag{13.44a}$$

$$\text{s.t.} \quad \mathbb{P}\left(h_i(x,\zeta)\leq 0,\ i=1,\ldots,m\right) \;\geq\; p \tag{13.44b}$$

where $c:\mathbb{R}^n\to\mathbb{R}$ is a cost function, $h_i:\mathbb{R}^n\times\mathbb{R}^k\to\mathbb{R}$, $i=1,\ldots,m$, are constraint functions, $\zeta\in\mathbb{R}^k$ is a random vector and $\mathbb{P}$ is a probability measure defined on some probability space[3], $p\in[0,1]$, and $X\subseteq\mathbb{R}^n$ is nonempty. The constraint (13.44b) is called a *chance constraint* or a *probabilistic constraint*. The problem (13.44) is

[3] Formally a *probability space* is a triple $(\Omega,\mathcal{F},\mathbb{P})$ where the *sample space* Ω is an arbitrary nonempty set. The *σ-algebra* $\mathcal{F}\subseteq 2^{\Omega}$ is a collection of subsets $A\subseteq\Omega$ called *events* that satisfies: (i) $\Omega\in\mathcal{F}$; (ii) if $A\in\mathcal{F}$ then $\Omega\setminus A\in\mathcal{F}$; and (iii) if $A_i\in\mathcal{F}$ for $i=1,2,\ldots,$ then $\cup_i A_i\in\mathcal{F}$. The *probability measure* $\mathbb{P}:\mathcal{F}\to[0,1]$ is a function such that (i) $\mathbb{P}(\Omega)=1$; and (ii) if $A_i\in\mathcal{F}$ for $i=1,2,\ldots$ are pairwise disjoint, then $\mathbb{P}(\cup_i A_i)=\sum_i\mathbb{P}(A_i)$. A *random variable* or *random vector* Z defined on the probability space $(\Omega,\mathcal{F},\mathbb{P})$ is a function $Z:\Omega\to\mathbb{R}^m$ such that $\mathbb{P}(\{\omega\in\Omega: Z(\omega)\leq z\})$ is called the *probability of the event* $\{Z\leq z\}$ and sometimes denoted by $\mathbb{P}(Z\leq z)$. The *probability distribution function* or *distribution function* $F_Z:\mathbb{R}^m\to[0,1]$ of the random variable Z is the function defined by the probability measure $\mathbb{P}$, $F_Z(z)=\mathbb{P}(\{\omega\in\Omega: Z(\omega)\leq z\})$.

a deterministic optimization problem called a *chance constrained program*. It is generally intractable because the chance constraint (13.44b) is often nonconvex.

Compared with the robust program (13.2), the chance constrained program (13.44) allows the dependence on the uncertain parameter ζ of different constraints $h_i(x,\zeta) \le 0$ to be coupled across i and is more general than $h_i(x,\zeta_i) \le 0$. More importantly (13.44) is less conservative in the sense that the constraints $h_i(x,\zeta) \le 0$ for all i need not hold for almost all uncertain parameter values ζ, but only with a probability greater than or equal to p.

In this section we introduce two techniques to deal with the chance constrained program (13.44). When the constraint functions h_i and the probability measure $\mathbb{P}$ have certain concavity properties then the chance constraint (13.44b) is convex and (13.44) is tractable. This is studied in Chapter 13.2.1. When these concavity conditions may not hold, we derive bounds on the tail probability of a random variable, called concentration inequalities, and show how these ideas can provide inner approximations of the feasible set defined by the chance constraint (13.44b). These inner approximations may be tractable or easier to solve. This is studied in Chapter 13.2.2. In Chapter 13.3 we approximate the chance constraint by a finite set of random constraints.

13.2.1 Tractable instances: convexity, strong duality and optimality

In this subsection we studied conditions under which the chance constrained program is tractable. Unless otherwise specified (see Remark 13.3), we assume that the chance constraint is separable in the decision variable x and the random vector ζ, i.e., we consider the following special case of (13.44) where the constraint function takes the form $\zeta \le h(x)$:

$$\min_{x\in X} \; c(x) \qquad \text{s.t.} \qquad \mathbb{P}(\zeta \le h(x)) \ge p$$

where $h : \mathbb{R}^n \to \mathbb{R}^m$, $\zeta \in \mathbb{R}^m$ and $X \subseteq \mathbb{R}^n$ is a nonempty convex set. In this case the chance constraint can be expressed in terms of the (probability) distribution function $F_\zeta : \mathbb{R}^m \to [0,1]$ of ζ and the chance constrained program becomes:

$$\min_{x\in X} \; c(x) \qquad \text{s.t.} \qquad F_\zeta(h(x)) \ge p \tag{13.45}$$

The function $F_\zeta(z)$ is also called a cumulative distribution function. A (probability) density function, if exists, is denoted by $f_\zeta(z)$. A distribution function F_ζ is nondecreasing, i.e., $F_\zeta(z_1) \le F_\zeta(z_2)$ if $z_1 \le z_2$, and *upper semicontinuous*, i.e., if $z_k \to z$ then

$$F_\zeta(z) \;\ge\; \limsup_k F_\zeta(z_k) \tag{13.46}$$

In this book we will ignore measurability issues, i.e., we will assume all random variables or processes encountered are well defined, they generate appropriate σ-algebra on which appropriate probability measures are defined, and all functions encountered are measurable. When we say two sets are the same, we mean they differ only by a measure-zero set.

We next study two equivalent formulations of (13.45) for convexity analysis that mainly differ in their specification of the feasible set. The first formulation hides both the constraint function h and the distribution function F_ζ in the feasible set X_p for x:

$$\min_{x\in X} \; c(x) \quad \text{s.t.} \quad x \in X_p := \left\{x \in \mathbb{R}^n : F_\zeta(h(x)) \geq p\right\} \tag{13.47a}$$

where $X \subseteq \mathbb{R}^n$ is a nonempty convex set. The second formulation allows the structure of h to play a more explicit role in the optimality condition and uses the p-level set Z_p of the distribution function $F_\zeta(z)$, defined by:

$$Z_p := \left\{z \in \mathbb{R}^m : F_\zeta(z) \geq p\right\} \tag{13.47b}$$

The chance constrained problem (13.45) is then a minimization over both x and z:

$$\min_{(x,z)\in X\times Z_p} \; c(x) \quad \text{s.t.} \quad h(x) \geq z \tag{13.47c}$$

with the explicit constraint $h(x) \geq z$ that can be used for optimality analysis. The main issue for the first formulation is the convexity of X_p and that for the second formulation is conditions for strong duality and saddle point optimality We study them in turn.

Convexity of X_p.

Suppose components h_i, $\forall i$, of $h : \mathbb{R}^n \to \mathbb{R}^m$ and the distribution function $F_\zeta : \mathbb{R}^m \to [0,1]$ are real-valued and concave functions. Then the feasible set X_p in (13.47a) is convex (Exercise 13.10). Important distribution functions however may not be concave, as the next example shows.

Example 13.4 (Gaussian distribution)**.** The multivariate Gaussian random vector $Z \in \mathbb{R}^m$ has a density function

$$f_\zeta(z) \;:=\; \frac{1}{\sqrt{(2\pi)^m \det(\Sigma)}} \exp\left(-\frac{1}{2}(z-\mu)^\mathsf{T}\Sigma^{-1}(z-\mu)\right)$$

with a mean $\mu \in \mathbb{R}^m$ and a positive definite covariance matrix $\Sigma \in \mathbb{R}^{m\times m}$. Then

$$\ln f_\zeta(z) \;=\; -\frac{1}{2}(z-\mu)^\mathsf{T}\Sigma^{-1}(z-\mu) - \frac{1}{2}\ln\left((2\pi)^m \det(\Sigma)\right)$$

and hence f_ζ is log-concave. It can be shown that its distribution function $F_\zeta(z)$ is also log-concave (see (13.48) below). □

Example 13.4 motivates a more general notion of concavity under which the feasible set X_p remains convex.

Definition 13.2 (α-concavity)**.** Let $\Omega \subseteq \mathbb{R}^m$ be a convex set. A nonnegative function $f : \Omega \to \mathbb{R}_+$ is called *α-concave* with $\alpha \in [-\infty, \infty]$ if for all $x, y \in \Omega$ such that $f(x) > 0$

and $f(y) > 0$ and all $\lambda \in [0,1]$, we have

$$f(\lambda x + (1-\lambda)y) \;\geq\; \begin{cases} (\lambda f^{\alpha}(x) + (1-\lambda) f^{\alpha}(y))^{1/\alpha} & \text{if} \quad \alpha \notin \{0, -\infty, \infty\} \\ f^{\lambda}(x)\, f^{1-\lambda}(y) & \text{if} \quad \alpha = 0 \\ \min\{f(x), f(y)\} & \text{if} \quad \alpha = -\infty \\ \max\{f(x), f(y)\} & \text{if} \quad \alpha = \infty \end{cases}$$

The class of α-concave functions includes several commonly used function classes as special cases. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called *concave* if for all $x, y \in \mathbb{R}^n$ and all $\lambda \in [0,1]$ we have $f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda) f(y)$; this corresponds to 1-concavity. More generally, for $\alpha \notin \{0, -\infty, \infty\}$, f is α-concave if and only if f^{α} is concave. The function f is called *log-concave* if $\log f$ is concave with respect to any base; this corresponds to 0-concavity. The function f is called *quasi-concave* if $f(\lambda x + (1-\lambda)y) \geq \min\{f(x), f(y)\}$; this corresponds to $-\infty$-concavity. The function f is ∞-concave if and only if it is a constant function.

One can also define α-concavity for a probability measure $\mathbb{P}$ which is a stronger property in the sense that an α-concave probability measure implies an α-concave distribution function F_{ζ}, but the converse may not hold. Unless otherwise specified we assume the chance constraint is separable in x and ζ in which case the α-concavity of F_{ζ} is sufficient for our purposes (cf. Remark 13.3). The α-concavity of a probability density function $f_{\zeta}(z)$ induces a probability measure, and hence its distribution function $F_{\zeta}(z)$, that is β-concave for some β. Specifically it can be shown (see [142, Corollary 4.16, p.106]) that if the probability density function $f_{\zeta}(z)$ defined on $\Omega \subseteq \mathbb{R}^m$, with $\int_{\Omega} f_{\zeta}(z)dz = 1$, is α-concave with $\alpha \in [-1/m, \infty]$ and if $f_{\zeta}(z) > 0$ in the interior of Ω, then the probability measured $\mathbb{P}$ defined by

$$\mathbb{P}(A) \;:=\; \int_A f_{\zeta}(z)dz, \qquad A \subseteq \Omega \tag{13.48a}$$

is β-concave with

$$\beta \;:=\; \begin{cases} \frac{\alpha}{1+m\alpha} & \text{if} \quad \alpha \in (-1/m, \infty) \\ -\infty & \text{if} \quad \alpha = -1/m \\ 1/m & \text{if} \quad \alpha = \infty \end{cases} \tag{13.48b}$$

This implies that, since the Gaussian density function f_{ζ} of Example 13.4 is log-concave ($\alpha = 0$), so is its distribution function F_{ζ}.

The following properties of α-concavity are important in determining the convexity of the feasible set X_p in (13.47a) (Exercise 13.11).

Lemma 13.6 (α-concavity)**.** Let $\Omega \subseteq \mathbb{R}^m$ be a convex set and consider nonnegative function $f : \Omega \to \mathbb{R}_+$.

1 For $\alpha \in [-\infty, \infty]$, $(a,b) \in \mathbb{R}^2_+$ with $a > 0$, $b > 0$, and $\lambda \in [0,1]$, define

$$m_\alpha(a,b,\lambda) := \begin{cases} (\lambda a^\alpha + (1-\lambda) b^\alpha)^{1/\alpha} & \text{if} \quad \alpha \notin \{0, -\infty, \infty\} \\ a^\lambda b^{1-\lambda} & \text{if} \quad \alpha = 0 \\ \min\{a,b\} & \text{if} \quad \alpha = -\infty \\ \max\{a,b\} & \text{if} \quad \alpha = \infty \end{cases}$$

Therefore f being α-concave is equivalent to $f(\lambda x + (1-\lambda)y) \geq m_\alpha(f(x), f(y), \lambda)$. Then for each (a,b,λ), the mapping $\alpha \to m_\alpha(a,b,\lambda)$ is nondecreasing and continuous.

2 If f is α-concave then it is β-concave for all $\beta \leq \alpha$; in particular concavity implies log-concavity which implies quasi-concavity.

3 If f is α-concave for some $\alpha > -\infty$ then f is continuous on $\mathrm{ri}(\Omega)$.

4 Let $h_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m$. If all h_i are concave and $f : \mathbb{R}^m \to \mathbb{R}$ is nonnegative, nondecreasing (i.e., $x \leq y \in \mathbb{R}^m \Rightarrow f(x) \leq f(y)$), and α-concave for some $\alpha \in [-\infty, \infty]$, then $f \circ h : \mathbb{R}^n \to \mathbb{R}_+$ is α-concave.

5 Let $f : \mathbb{R}^{n_1+n_2} \to \mathbb{R}_+$. Suppose there exists an $\alpha \in [-\infty, \infty]$ such that, for all $y \in Y \subseteq \mathbb{R}^{n_2}$, $f(x,y)$ is α-concave in x on a convex set $X \subseteq \mathbb{R}^{n_1}$. Then $g(x) := \inf_{y \in Y} f(x,y)$ is α-concave on X.

Consider a concave function $f : \mathbb{R}^m \to \mathbb{R}$ (and is therefore proper as an extended real-valued function). We say $y \in \mathbb{R}^m$ is a *subgradient of* f at $\bar{x} \in \mathbb{R}^m$ if $-y$ is a subgradient of the convex function $-f$, i.e., if

$$f(x) \;\leq\; f(\bar{x}) + y^\mathsf{T}(x - \bar{x}), \qquad x \in \mathbb{R}^m$$

The set of all subgradients of the concave function f at $\bar{x}$ is the *subdifferential* $\partial f(\bar{x})$ *of* f at $\bar{x}$. Then $x^* \in \mathbb{R}^m$ is an optimal solution of $\sup_{x \in \mathbb{R}^m} f(x)$ if and only if $0 \in \partial f(x^*)$. Moreover Lemma 12.15 applies directly to the real-valued concave function f. In particular $f(x)$ is continuous on $\mathrm{ri}(\mathrm{dom}(f)) = \mathbb{R}^m$. For each $x \in \mathbb{R}^m$, $\partial f(x) \subseteq \mathbb{R}^m$ is a nonempty convex compact set. If $X \subset \mathbb{R}^m$ is nonempty and compact, then $\partial_X f := \cup_{x \in X} \partial f(x)$ is nonempty and bounded; moreover f is Lipschitz continuous over X with Lipschitz constant $L := \sup_{\xi \in \partial_X f} \|\xi\|_2$, i.e., $\|f(x) - f(y)\|_2 \leq L\|x - y\|_2$ for all $x, y \in \mathbb{R}^m$. More generally if f is α-concave with $\alpha > -\infty$ then it is continuous on the relative interior of its domain according to Lemma 13.6. A quasi-concave function ($\alpha = -\infty$) need not be continuous.

In general the feasible set X_p in (13.47a) is not convex or even connected. The following result provides a sufficient condition for the feasible set to be convex and closed.

Theorem 13.7 (Convexity of X_p)**.** Suppose all components h_i of $h : \mathbb{R}^n \to \mathbb{R}^m$ are concave and the distribution function $F_\zeta : \mathbb{R}^m \to [0,1]$ is α-concave for some $\alpha \in [-\infty, \infty]$, then the feasible set $X_p := \{x \in \mathbb{R}^n : F_\zeta(h(x)) \geq p\}$ in (13.47a) is convex and closed.

Proof Lemma 13.6 implies that the function $H : \mathbb{R}^n \to [0,1]$ defined by $H(x) :=$

$F_\zeta(h(x))$ is a nonnegative α-concave function. To show that the set X_p for a fixed $p \in [0,1]$ is convex consider $x_1, x_2 \in X_p$ with $H(x_1) \geq p$ and $H(x_2) \geq p$ and $x := \lambda x_1 + (1-\lambda)x_2$ for any $\lambda \in [0,1]$. We have

$$H(x) \;\geq\; m_\alpha(H(x_1), H(x_2), \lambda)$$

If $\alpha = -\infty$, i.e., $H(x)$ is quasi-concave, then $H(x) \geq \min\{H(x_1), H(x_2)\} \geq p$, i.e., $x \in X_p$. Since the mapping $\alpha \to m_\alpha(a,b,\lambda)$ for each (a,b,λ) is nondeacreasing in α by Lemma 13.6, if $H(x)$ is α-concave for any $\alpha \in [-\infty, \infty]$, it is quasi-concave and hence $x \in X_p$. This proves that X_p is convex.

To show that X_p is closed, consider any sequence $x_k \in X_p$ with $x_k \to x$. We have

$$H(x) \;:=\; F_\zeta(h(x)) \;=\; F_\zeta\left(\lim_k h(x_k)\right) \;\geq\; \limsup_k F_\zeta(h(x_k)) \;\geq\; p$$

where the second equality follows from the continuity of h since h is concave on $\mathbb{R}^m$ (Lemma 13.6), the first inequality follows from the uppersemicontinuity of distribution functions from (13.46), and the last inequality follows from $x_k \in X_p$. □

Remark 13.3 (Inseparable chance constraint)**.** Theorem 13.7 generalizes to the case where the chance constraint in (13.45) is not separable in the decision variable x and the random vector ζ and takes the form $H(x) := \mathbb{P}(h_i(x,\zeta) \geq 0,\ i = 1,\ldots,m) \geq p$. It can be shown ([142, Theorem 4.39, p.115]) that if $h_i(x,\zeta)$, $i = 1,\ldots,m$, are jointly quasi-concave in $(x,\zeta) \in \mathbb{R}^{n+k}$ and if ζ has a probability measure that is α-concave, then $H(x)$ is α-concave on $\{x \in \mathbb{R}^n : \exists \zeta \in \mathbb{R}^k \text{ s.t. } h_i(x,\zeta) \geq 0, \forall i\}$. This implies that the feasible set $X := \{x \in \mathbb{R}^n : H(x) \geq p\}$ is convex and closed, because for all $x, y \in X$ and $\lambda \in [0,1]$,

$$H(\lambda x + (1-\lambda)y) \;\geq\; m_\alpha(H(x), H(y), \lambda) \;\geq\; \min\{H(x), H(y)\} \;\geq\; p$$

where the first inequality follows from the α-concavity of H, the second inequality follows from the monotonicity of the mapping $\alpha \to m_\alpha(a,b,\lambda)$, and the last inequality follows from $x, y \in X$. Compared with Theorem 13.7, the functions $h_i(x,\zeta)$ are required only to be quasi-concave ($\alpha = -\infty$) which is weaker than concavity, but the probability measure of ζ is required to be α-concave which is stronger than requiring only its distribution function F_ζ to be α-concave. □

Duality and optimality.

Fix $p \in (0,1)$. We now study the second formulation in (13.47b)(13.47c) where h plays a more explicit role in the optimality condition. Recall the p-level set Z_p of the distribution function $F_\zeta(z)$:

$$Z_p := \left\{z \in \mathbb{R}^m : F_\zeta(z) \geq p\right\}$$

and the chance constrained formulation:

$$c^* \;:=\; \min_{(x,z) \in X \times Z_p} c(x) \quad \text{s.t.} \quad h(x) \geq z \tag{13.49a}$$

where $c : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}^m$ are real-valued, and $X \subseteq \mathbb{R}^n$ is nonempty convex. The Lagrangian is

$$L(x, z, \mu) = c(x) \; + \; \mu^{\mathsf{T}} (z - h(x))$$

the dual function is

$$d(\mu) := \inf_{(x,z) \in X \times Z_p} L(x, z, \mu) \;\; = \;\; \underbrace{\inf_{x \in X} \Big(c(x) - \mu^{\mathsf{T}} h(x) \Big)}_{d_X(\mu)} \;\; + \;\; \underbrace{\inf_{z \in Z_p} \mu^{\mathsf{T}} z}_{d_Z(\mu)}, \qquad \mu \in \mathbb{R}^m$$

and the dual problem is

$$d^* \;\; := \;\; \sup_{\mu \geq 0} d(\mu) := \sup_{\mu \geq 0} \left(d_X(\mu) + d_Z(\mu) \right) \tag{13.49b}$$

where

$$d_X(\mu) \;\; := \;\; \inf_{x \in X} \Big(c(x) - \mu^{\mathsf{T}} h(x) \Big), \qquad d_Z(\mu) \;\; := \;\; \inf_{z \in Z_p} \mu^{\mathsf{T}} z \tag{13.49c}$$

Since we only partially dualize the primal problem (13.49a) we cannot characterize a primal-dual optimal point (x^*, z^*, μ^*) by the KKT condition, but we can characterize it as a saddle point. Recall that $(x^*, z^*, \mu^*) \in X \times Z_p \times \mathbb{R}^m_+$ is a saddle point if and only if

$$\sup_{\mu \geq 0} L(x^*, z^*, \mu) \;\; = \;\; L(x^*, z^*, \mu^*) \;\; = \;\; \inf_{(x,z) \in X \times Z_p} L(x, z, \mu^*) \tag{13.50}$$

By the definition of L, d_X, d_Z, the minimization in (13.50) is equivalent to

$$d_X(\mu^*) \;\; = \;\; c(x^*) - \mu^{*\mathsf{T}} h(x^*), \qquad\qquad d_Z(\mu^*) \;\; = \;\; \mu^{*\mathsf{T}} z^*$$

It is shown in Theorem 13.8 that the maximization in (13.50) is equivalent to complementary slackness, given $h(x^*) \geq z^*$ (or see Exercise 8.15).

Even though c and h are real-valued the dual function $d(\mu)$ can be extended real-valued. Moreover $d(\mu)$ may not be differentiable even if c and h are because the minimizer (x, z) of the Lagrangian function may not be unique. It is however always concave hence always subdifferentiable for any c and h. This is a nonsmooth convex optimization problem studied in Chapter 12. In particular the problem (13.49) takes the same form as the nonsmooth convex problem (12.42). We next use the Slater Theorem 12.28 to provide sufficient conditions for strong duality and dual optimality and the Saddle-point Theorem 12.20 to characterize a primal-dual optimal point. We make the following assumptions:

C13.1 *Convexity*:
- c is convex; h is concave (i.e., each component h_i is concave);
- X is nonempty convex;
- The distribution function $F_\zeta(z)$ is α-concave for an $\alpha \in [-\infty, \infty]$.

C13.2 *Slater condition*: one of the following holds:
- CQ1: There exists $(\bar{x}, \bar{z}) \in X \times Z_p$ such that $h(\bar{x}) > \bar{z}$; or
- CQ2: h is affine and there exists $(\bar{x}, \bar{z}) \in \mathrm{ri}(X \times Z_p)$ such that $h(\bar{x}) \geq \bar{z}$.

The α-concavity of F_ζ implies that F_ζ is quasi-concave (Lemma 13.6) and hence Z_p is a nonempty convex set (since $p \in (0,1)$).

Theorem 13.8 (Strong duality and optimality)**.** Suppose the chance constrained program and its dual (13.49) satisfy conditions C13.1 and C13.2. Then

1 *Strong duality and dual optimality*. If $c^* > -\infty$ then there exists a dual optimal solution $\mu^* \geq 0$ that closes the duality gap, i.e., $c^* = d^* = d(\mu^*)$. The set of dual optimal solutions μ^* is convex and closed; it is also compact under CQ1.
2 *Saddle point characterization*. A point $(x^*, z^*, \mu^*) \in X \times Z_p \times \mathbb{R}^m_+$ is primalp-dual optimal and closes the duality gap, i.e., $c(x^*) = c^* = d^* = d(\mu^*)$ if and only if

$$d_X(\mu^*) \;=\; c(x^*) - \mu^{*\mathsf{T}} h(x^*), \quad d_Z(\mu^*) \;=\; \mu^{*\mathsf{T}} z^*, \quad \mu^{*\mathsf{T}}(z^* - h(x^*)) \;=\; 0 \tag{13.51}$$

Such a point is a saddle point.

Proof Since c is real-valued, $\mathrm{dom}(c) = \mathbb{R}^n$. The Slater Theorem 12.28 in Chapter 12.7 then implies that strong duality holds and there is a dual optimal μ^* that attains dual optimality. Moreover the set of dual optimal solutions is convex and closed, and also bounded (and hence compact) under CQ1.

To characterize a primal-dual optimal we apply the Saddle-point Theorem 12.20 which states that $(x^*, z^*, \mu^*) \in X \times Z_p \times \mathbb{R}^m_+$ is primal-dual optimal and closes the duality gap if and only if it is a saddle point, i.e., if and only if it satisfies (13.50). As discussed above, the second equality in (13.50) is equivalent to the first two conditions in (13.51). We next show that the first equality in (13.50) is equivalent to the complementary slackness condition in (13.51).

First we claim that, if $(x^*, z^*, \mu^*) \in X \times Z_p \times \mathbb{R}^m_+$ is a primal-dual optimal or a saddle point, then $h(x^*) \geq z^*$. If (x^*, z^*, μ^*) is primal-dual optimal then (x^*, z^*) is primal feasible and hence $h(x^*) \geq z^*$. If (x^*, z^*, μ^*) is a saddle point then, if $h_i(x^*) < z_i^*$ for any i, then $\sup_{\mu \geq 0} L(x^*, z^*, \mu) = \infty$ contradicting that $\sup_{\mu \geq 0} L(x^*, z^*, \mu) = L(x^*, z^*, \mu^*)$. Then the first equality in (13.50) yields

$$L(x^*, z^* \mu^*) \;=\; \sup_{\mu \geq 0} L(x^*, z^*, \mu) \;=\; \sup_{\mu \geq 0} \Big(c(x^*) \;+\; \mu^{\mathsf{T}} (z^* - h(x^*)) \Big) \;\leq\; c(x^*)$$

with equality if and only if $\sup_{\mu \geq 0} \mu^{\mathsf{T}} (z^* - h(x^*)) = 0$. Since $\mu^* \geq 0$ attains the maximum of $L(x^*, z^*, \mu)$, the complementary slackness condition in (13.51) is established. □

Remark 13.4 (Primal optimality and dual differentiability)**.** 1 Denote the sets of minimizers in (13.49c) by

$$X(\mu) := \{x \in X : d_X(\mu) = c(x) - \mu^{\mathsf{T}} h(x)\}, \quad Z(\mu) := \{z \in Z_p : d_Z(\mu) = \mu^{\mathsf{T}} z\}$$

Theorem 13.8 holds even if $X(\mu)$ and $Z(\mu)$ are empty, i.e., primal optimality may not be attained. If X and Z_p are nonempty, convex and compact, then the sets

$X(\mu)$ and $Z(\mu)$ of primal optimal solutions are nonempty, convex and compact and hence the dual function $d(\mu)$ is a real-valued concave function. Moreover the subdifferentials of d_X, d_Z are

$$\partial d_X(\mu) \;=\; \mathrm{conv}(-h(x) : x \in X(\mu)), \qquad \partial d_Z(\mu) \;=\; Z(\mu)$$

and hence $\partial d(\mu) = \mathrm{conv}(-h(x) : x \in X(\mu)) + Z(\mu)$. These results are derived in Exercise 13.12 using Theorem 12.19 and Theorem 12.26.

2 See Exercise 13.13 for an alternative proof of the saddle-point characterization (13.51). It applies Theorem 12.21 to the dual (13.49b) and illustrates basic techniques in nonsmooth analysis that are used to reduce optimality conditions to a saddle-point characterization. □

13.2.2 Concentration inequalities and safe approximation

In Chapter 13.2.1 we study conditions, e.g., α-concavity of the distribution function F_Z, under which the chance constrained program (13.45) is convex. In this subsection we introduce the idea of solving a safe approximation of (13.45) that is more conservative but easier to solve. We illustrate this idea with the chance constrained linear program (cf. the robust linear program (13.11)):

$$\min_{x\in\mathbb{R}^n} \; c^\mathsf{T} x \quad \text{s.t.} \quad \mathbb{P}\left(\sum_{l=1}^{k} \left(a_l^\mathsf{T} x - b_l\right) \zeta_l \le -(a_0^\mathsf{T} x - b_0)\right) \;\ge\; 1-\epsilon \tag{13.52a}$$

where the uncertain parameter is the random vector $\zeta := (\zeta_l, l = 1,\ldots,k)$. We will show that, if the moment generating functions of ζ_l are upper bounded by those of Gaussian random variables, then the following second-order cone program is a safe approximation of (13.52a):

$$\min_{x\in\mathbb{R}^n} \; c^\mathsf{T} x \quad \text{s.t.} \quad r\|\hat{A}x - \hat{b}\|_2 \le -(\hat{a}_0^\mathsf{T} x - \hat{b}_0) \tag{13.52b}$$

where $\hat{A}, \hat{b}, \hat{a}_0, \hat{b}_0$ depend on $(a_l, b_l, l \ge 0)$ and r depends on ϵ. The second-order cone program (13.52b) is generally much simpler to solve than the chance constrained problem (13.52a). It is a safe approximation in the sense that an x that is feasible, or optimal, for (13.52b) will always satisfy the chance constraint in (13.52a).

The derivation of a safe approximation generally boils down to deriving an explicit convex feasible set of the approximation that is a subset (inner approximation) of the feasible set of the chance constrained problem. It relies on two techniques. First we upper bound the violation probability of the chance constraint in terms of distribution properties of the uncertain parameter ζ_l such as its variance or its moment generating function ψ_{ζ_l} (e.g. Chernoff bound). Then we upper bound these distribution properties by known properties (e.g., the moment generating function of the Gaussian distribution). In the rest of this subsection we derive some basic bounds on the tail probability of a random variable, study properties of sub-Gaussian random variables, and then use

these techniques to derive the safe approximation (13.52b) of the chance constrained linear program (13.52a). These bounds are the most basic inequalities in probability and widely applicable, e.g., used in Chapter 13.3.5 to derive sample complexity of scenario programs.

In this subsection we follow the usual notation in the probability literature where capital letters typically denote random variables, e.g., Y, and small letters their values, e.g., y.

Markov's inequality.

Let Y be any *nonnegative* random variable with finite mean $EY < \infty$. Let $\delta(x)$ denote the indicator function where $\delta(x) = 1$ if x is true and 0 otherwise (different definition from $\delta(x)$ in Chapter 12). Observe that, for all $t > 0$, $Y/t \geq \delta(Y \geq t)$. Taking expectation on both sides we obtain the Markov's inequality: for all $t > 0$,

$$\mathbb{P}(Y \geq t) \;\leq\; \frac{EY}{t} \tag{13.53a}$$

Let $R \subseteq \mathbb{R}$ be any interval and let $\phi : R \to \mathbb{R}_+$ be a nonnegative nondecreasing function on R. Since $\delta(Y \geq t) = \delta(\phi(Y) \geq \phi(t))$, (13.53a) implies, for any t with $\phi(t) > 0$,

$$\mathbb{P}(Y \geq t) \;=\; \mathbb{P}(\phi(Y) \geq \phi(t)) \;\leq\; \frac{E(\phi(Y))}{\phi(t)} \tag{13.53b}$$

Chebyshev's inequality.

Let $Y := |X - EX|$ be nonnegative where X is an arbitrary random variable with a finite variance $\mathrm{var}(X) < \infty$. Let $R := (0, \infty)$ and $\phi(t) = t^2$. Then the Markov's inequality (13.53b) implies the Chebyshev's inequality: for any $t > 0$,

$$\mathbb{P}(|X - EX| \geq t) \;\leq\; \frac{\mathrm{var}(X)}{t^2} \tag{13.54a}$$

For the sample mean $n^{-1}\sum_i X_i$ of a sequence of independent random variables $X_1, \ldots, X_n$, since $\mathrm{var}(\sum_i X_i) = \sum_i \mathrm{var}(X_i)$, (13.54a) implies

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_i (X_i - EX_i)\right| \geq t\right) \;\leq\; \frac{\sum_i \mathrm{var}(X_i)}{n^2 t^2} \;=\; \frac{v_n}{nt^2} \tag{13.54b}$$

where $v_n := n^{-1}\sum_i \mathrm{var}(X_i)$ is the average variance. In particular if X_i are iid (independent and identically distributed) then $\sigma_n^2 = \mathrm{var}(X_1)$ and the tail probability decreases in n at the rate of n^{-1}.

Chernoff bound.

For a random variable Y with a finite expectation $EY < \infty$, $E(e^{\lambda Y})$ is called a *moment-generating function* of Y, as a function of $\lambda \in \mathbb{R}$. Let

$$\psi_Y(\lambda) := \ln E(e^{\lambda Y}), \qquad \lambda \in \mathbb{R} \tag{13.55a}$$

be the log moment-generating function of Y. Here $\ln := \log_e$ denotes the natural log and we sometimes use log if the base is clear from the context. The function $\psi_Y(\lambda)$ is convex in λ (Exercise 13.14). Recall the conjugate function defined (in Chapter 12.3.2) as

$$\psi_Y^*(t) := \sup_{\lambda \in \mathbb{R}} (t\lambda - \psi_Y(\lambda)), \qquad t \in \mathbb{R} \tag{13.55b}$$

Jensen's inequality says that, if f is a convex function, then $E(f(x)) \geq f(EX)$ (see Exercise 12.11). Hence the log moment-generating function $\psi_Y(\lambda)$ satisfies

$$\psi_Y(0) = 0, \qquad \psi_Y(\lambda) \geq \lambda EY \tag{13.56}$$

We now bound the tail probability $\mathbb{P}(Y \geq t)$, in two equivalent forms ($t \geq EY$ and $t \in \mathbb{R}$). For $\lambda \geq 0$, the function $\phi(t) := e^{\lambda t}$ is a nonnegative nondecreasing function of t over $\mathbb{R}$ and hence the Markov's inequality (13.53b) implies $\mathbb{P}(Y \geq t) \leq E(e^{\lambda Y})/e^{\lambda t}$ for all $\lambda \geq 0$. Therefore, for $t \geq EY$,

$$\ln \mathbb{P}(Y \geq t) \leq -\sup_{\lambda \geq 0} (t\lambda - \psi_Y(\lambda)) = -\sup_{\lambda \in \mathbb{R}} (t\lambda - \psi_Y(\lambda)) = -\psi_Y^*(t) \tag{13.57a}$$

where the first equality follows because, for $\lambda \leq 0$ and $t \geq EY$, $t\lambda - \psi_Y(\lambda) \leq \lambda(t - EY) \leq 0 = -\psi_Y(0)$ by (13.56). Hence the Chernoff bound on the tail probability is:

$$\mathbb{P}(Y \geq t) \leq e^{-\psi_Y^*(t)}, \qquad t \geq EY \tag{13.57b}$$

where the conjugate function $\psi_Y^*(t)$ is defined in (13.55). Note that (13.57) holds for $t \geq EY$. For $t \leq EY$, (13.56) implies that $\psi_Y(\lambda) - t\lambda \geq \lambda(EY - t) \geq 0$ if (and only if) $\lambda \geq 0$ and hence $-\sup_{\lambda \geq 0}(t\lambda - \psi_Y(\lambda)) \geq 0$ in (13.57a) is a trivial upper bound. Therefore the Chernoff bound that holds for all $t \in \mathbb{R}$ takes the following forms:

$$\ln \mathbb{P}(Y \geq t) \leq \inf_{\lambda \geq 0} \ln\left(e^{-\lambda t} E e^{\lambda Y}\right) \tag{13.58a}$$

$$\mathbb{P}(Y \geq t) \leq \exp\left(-\sup_{\lambda \geq 0} (t\lambda - \psi_Y(\lambda))\right), \qquad t \in \mathbb{R} \tag{13.58b}$$

where the infimum and supremum are taken over $\lambda \geq 0$ as opposed to $\lambda \in \mathbb{R}$ in (13.55b).

If $Y := n^{-1} \sum_i X_i$ is the sample mean of n independent random variables X_i with $EX_i < \infty$, $i = 1, \ldots, n$, then

$$\psi_Y(\lambda) = \sum_i \psi_{X_i}(\lambda/n) \tag{13.59a}$$

$$\psi_Y^*(t) = \sup_{\lambda \in \mathbb{R}} \sum_i \left(t\lambda - \psi_{X_i}(\lambda)\right) \leq \sum_i \psi_{X_i}^*(t) \tag{13.59b}$$

with equality if X_i are iid. The sample mean of n independent random variables X_i satisfies the Chernoof bound:

$$\mathbb{P}\left(\frac{1}{n}\sum_i X_i \geq t\right) \leq e^{-\psi_Y^*(t)} = e^{-nI_n(t)}, \qquad t \geq \frac{1}{n}\sum_i EX_i \tag{13.60a}$$

where $I_n(t)$ is called a rate function defined as:

$$I_n(t) := \sup_{\lambda\in\mathbb{R}}\left(t\lambda - \frac{1}{n}\sum_i \psi_{X_i}(\lambda)\right), \qquad t \geq \frac{1}{n}\sum_i EX_i \tag{13.60b}$$

The rate function $I_n(t) \leq (1/n)\sum_i \psi_{X_i}^*(t)$ with equality if X_i are iid. For arbitrary $t \in \mathbb{R}$, the rate function is (from (13.58)):

$$I_n(t) := \sup_{\lambda\geq 0}\left(t\lambda - \frac{1}{n}\sum_i \psi_{X_i}(\lambda)\right), \qquad t \in \mathbb{R} \tag{13.60c}$$

Therefore the tail probability decays exponentially in n when X_i are independent and $I_n(t)$ is independent of n. Indeed if X_i are iid then $I_n(t) = I(t) = \sup_{\lambda\in\mathbb{R}}\left(t\lambda - \psi_{X_1}(\lambda)\right)$ and

$$\mathbb{P}\left(\frac{1}{n}\sum_i X_i \geq t\right) \leq e^{-n\psi_{X_1}^*(t)}, \qquad t \geq EX_1$$

The Chernoff bound is extremely useful. We will use it to derive a safe approximation of chance constrained linear program below and sample complexity results in Chapter 13.3.5.

Sub-Gaussian random variable.

Gaussian random variable is useful for bounding other random variables because its log moment-generating function $\phi_Y(\lambda)$ is particularly simple (quadratic). Therefore the supremum in the Chernoff bounds (13.57) (13.58) (13.60) can be computed in closed form.

Example 13.5 (Gaussian random variable)**.** Consider the Gaussian random variable Y with mean $\mu := EY$ and standard deviation $\sigma := \sqrt{\text{var}(Y)}$. Its log moment-generating function is

$$\psi_G(\lambda) := \ln E(e^{\lambda Y}) = \ln\left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{\lambda y} e^{-(y-\mu)^2/2\sigma^2}\, dy\right) = \mu\lambda + \frac{1}{2}\sigma^2\lambda^2 \tag{13.61a}$$

Its conjugate function is

$$\psi_G^*(t) := \sup_{\lambda\in\mathbb{R}}\left(t\lambda - \mu\lambda - \frac{1}{2}\sigma^2\lambda^2\right) = \frac{(t-\mu)^2}{2\sigma^2} \tag{13.61b}$$

where the maximizer $\lambda^* = (t-\mu)/\sigma^2$. For $t := \mu + r\sigma$ with $r \geq 0$, the Chernoff bound is (from (13.57b))

$$\mathbb{P}(Y > \mu + r\sigma) \;\leq\; e^{-r^2/2}, \qquad r \geq 0 \tag{13.61c}$$

i.e., the tail probability that the Gaussian random variable Y is r standard deviations above its mean decays exponentially in r^2.

Consider a weighted sum $Y := \sum_i a_i X_i$ of independent Gaussian random variables $X_1, \ldots, X_n$ with parameter (μ_i, σ_i^2). Then Y is Gaussian with parameter $\left(\sum_i a_i \mu_i, \sum_i a_i^2 \sigma_i^2\right)$. Hence (13.61) implies

$$\psi_Y(\lambda) \;=\; \ln E e^{\lambda Y} \;=\; \lambda \sum_i a_i \mu_i + \frac{\lambda^2}{2} \sum_i a_i^2 \sigma_i^2, \qquad \lambda \in \mathbb{R}$$

$$\psi_Y^*(t) \;=\; \sup_{\lambda \in \mathbb{R}} (t\lambda - \phi_Y(\lambda)) \;=\; \frac{(t - \sum_i a_i \mu_i)^2}{2 \sum_i a_i^2 \sigma_i^2}, \qquad t \in \mathbb{R}$$

and the Chernoff bound

$$\mathbb{P}\left(\sum_i a_i (X_i - \mu_i) > r \sqrt{\sum_i a_i^2 \sigma_i^2} \right) \;\leq\; e^{-r^2/2}, \qquad r \geq 0 \tag{13.62}$$

A special case is the sample mean $Y := n^{-1} \sum_i X_i$ of n independent Gaussian random variables X_i, $i = 1, \ldots, n$, with finite parameters (μ_i, σ_i^2). The tail probability satisfies the Chernoff bound (from (13.62)):

$$\mathbb{P}\left(\frac{1}{n} \sum_i (X_i - \mu_i) > t \right) \;\leq\; e^{-nt^2/2v_n}, \qquad t \geq 0$$

where $v_n := (1/n) \sum_i \sigma_i^2$ is the average variance. Compared with (13.54), the Chernoff bound $(2e^{-nt^2/2v_n})$ generally decays more rapidly than Chevyshev's bound (v_n/nt^2). If X_i are iid with parameter (μ, σ^2), this reduces to

$$\mathbb{P}\left(\frac{1}{n} \sum_i X_i - \mu > t \right) \;\leq\; e^{-nt^2/2\sigma^2}, \qquad t \geq 0$$

□

A random variable Y is called *sub-Gaussian* with parameter (μ, σ^2) if

$$\psi_Y(\lambda) \;:=\; \ln E(e^{\lambda Y}) \;\leq\; \mu\lambda + \frac{\sigma^2}{2}\lambda^2 \;=:\; \psi_{\mathrm{G}}(\lambda), \qquad \lambda \in \mathbb{R} \tag{13.63a}$$

i.e., if the log moment-generating function is upper bounded by that of a Gaussian random variable with mean μ and variance σ^2. This is equivalent to $E(e^{\lambda Y}) \leq \exp\left(\mu\lambda + \frac{\sigma^2}{2}\lambda^2\right)$ for all $\lambda \in \mathbb{R}$. If Y has zero mean $EY = 0$ then Y is called *sub-Gaussian* with variance factor σ^2 if

$$\psi_Y(\lambda) \;:=\; \ln E(e^{\lambda Y}) \;\leq\; \frac{\sigma^2}{2}\lambda^2 \;=:\; \psi_{\mathrm{G}}(\lambda), \qquad \lambda \in \mathbb{R} \tag{13.63b}$$

where $\psi_{\mathrm{G}}(\lambda)$ denotes the log moment-generating function of a zero-mean Gaussian random variable. Since $\psi_Y^*(t) \geq \psi_{\mathrm{G}}^*(t)$ for $t \in \mathbb{R}$ where $\psi_{\mathrm{G}}^*(t)$ is defined in (13.61b), (13.57) implies

$$\mathbb{P}(Y \geq t) \;\leq\; e^{-\psi_Y^*(t)} \;\leq\; e^{-(t-\mu)^2/2\sigma^2}, \qquad t \geq EY \tag{13.64}$$

Hence the tail probability $\mathbb{P}(Y \geq t)$ for $t \geq EY$ of a sub-Gaussian random variable Y decays more rapidly than that of the bounding Gaussian random variable.

Given a sequence $X_1, \ldots, X_n$ of sub-Gaussian random variables, we can bound the tail probability of its weighted sum $\sum_i a_i X_i$ and its maximum $\max_i X_i$.

1 Let $Y := \sum_i a_i X_i$, $a_i \in \mathbb{R}$. Suppose $X_1, \ldots, X_n$ are independent sub-Gaussian random variables with parameter (μ_i, σ_i^2), i.e.,

$$\phi_{X_i}(\lambda) \;\leq\; \phi_{\mathrm{G}}(\lambda) \;=\; \mu_i\lambda + \frac{\sigma_i^2}{2}\lambda^2, \qquad \lambda \in \mathbb{R}$$

Then its weighted sum Y is sub-Gaussian whose parameter $(\mu, \sigma^2) := \left(\sum_i a_i\mu_i, \sum_i a_i^2\sigma_i^2\right)$ is the weighted sum of individual parameters:

$$\psi_Y(\lambda) \;=\; \ln E\left(\prod_i e^{a_i\lambda X_i}\right) \;=\; \sum_i \ln E e^{a_i\lambda X_i} \;=\; \sum_i \psi_{X_i}(a_i\lambda) \;\leq\; \mu\lambda + \frac{\sigma^2}{2}\lambda^2$$

where the second equality follows since X_i are independent. Hence (13.64) implies that Y satisfies the Chernoff bound:

$$\mathbb{P}\left(\sum_i a_i X_i \geq t\right) \;\leq\; \exp\left(-\frac{(t - \sum_i a_i\mu_i)^2}{2\sum_i a_i^2\sigma_i^2}\right), \qquad t \geq EY \tag{13.65}$$

Comparing with (13.62) we see that the tail probability of a sub-Gaussian weighted sum is bounded by the Chernoff bound for the bounding Gaussian weighted sum. The corresponding bound for $t \in \mathbb{R}$ (as opposed to $t \geq EY$) will be established in the derivation of a safe approximation of the chance constrained linear program (Theorem 13.9). Therefore as far as Chernoof bound is concern, a sub-Gaussian random variable behaves like its bounding Gaussian random variable.

2 Let $Y := \max_i X_i$. Suppose $Y \geq 0$ and $X_1, \ldots, X_n$ are sub-Gaussian random variables with a common variance factor σ^2, i.e., for all i, $\psi_{X_i}(\lambda) \leq \sigma^2\lambda^2/2$, $\lambda \in \mathbb{R}$. Note that X_i are not necessarily independent. It can be shown that (Exercise 13.16)

$$E\left(\max_{i=1,\ldots,n} X_i\right) \;\leq\; \sigma\sqrt{2\ln n} \tag{13.66}$$

The Markov's inequality (13.53a) then implies a concentration inequality for the maximum of finitely many sub-Gaussian random variables:

$$\mathbb{P}\left(\max_{i=1,\ldots,n} X_i \geq t\right) \;\leq\; \frac{\sigma\sqrt{2\ln n}}{t}, \qquad t > 0$$

provided $Y := \max_i X_i \geq 0$.

Safe approximation.

A chance constrained problem is often intractable. We now use the Chernoff bound to derive a tractable safe approximation of a chance constrained linear program when the uncertain parameters are independent sub-Gaussian random variables.

Consider the chance constrained linear program (cf. the robust linear program (13.11)):

$$\min_{x\in\mathbb{R}^n} \; c^\mathsf{T}x \quad \text{s.t.} \quad \mathbb{P}\left(\sum_{l=1}^{k}\left(a_l^\mathsf{T}x - b_l\right)\zeta_l \le -(a_0^\mathsf{T}x - b_0)\right) \;\ge\; 1-\epsilon \tag{13.67a}$$

where $\epsilon \in (0,1)$, the probability measure $\mathbb{P}$ is on the random vector $\zeta := (\zeta_l, l = 1,\ldots,k)$ and c, $(a_l, b_l) \in \mathbb{R}^n \times \mathbb{R}$, $l = 0,\ldots,k$, are given. We say that an optimization problem is a *safe approximation* of the chance constrained program (13.67b) if the feasible set of the optimization problem is contained in the feasible set of (13.67b). This implies that any optimal solution of the safe approximation will satisfy the chance constraint in (13.67b).

Let $A := [a_1 \;\cdots\; a_k]^\mathsf{T} \in \mathbb{R}^{k\times n}$ and $b := (b_1,\ldots,b_k) \in \mathbb{R}^k$. Then (13.67a) becomes:

$$\min_{x\in\mathbb{R}^n} \; c^\mathsf{T}x \quad \text{s.t.} \quad \mathbb{P}\left(\zeta^\mathsf{T}(Ax-b) \le -(a_0^\mathsf{T}x - b_0)\right) \;\ge\; 1-\epsilon \tag{13.67b}$$

Theorem 13.9 (Safe approximation: LP)**.** Suppose the random variables $\zeta_l, l = 1,\ldots,k$, in the chance constrained program (13.67b) are independent and sub-Gaussian with parameters (μ_l, σ_l^2), $\sigma_l > 0$, i.e.,

$$\psi_{\zeta_l}(\lambda) \;:=\; \ln E_{\zeta_l}\left(e^{\lambda\zeta_l}\right) \;\le\; \mu_l\lambda + \frac{\sigma_l^2}{2}\lambda^2, \qquad \lambda\in\mathbb{R} \tag{13.68}$$

Then the following second-order cone program is a safe approximation of (13.67b):

$$\min_{x\in\mathbb{R}^n} \; c^\mathsf{T}x \quad \text{s.t.} \quad r\|\sqrt{\Sigma}(Ax-b)\|_2 \le -(\hat{a}_0^\mathsf{T}x - \hat{b}_0) \tag{13.69}$$

where $r := \sqrt{2\ln(1/\epsilon)}$, $\hat{a}_0 := a_0 + A^\mathsf{T}\mu \in \mathbb{R}^n$, $\hat{b}_0 := b_0 + b^\mathsf{T}\mu \in \mathbb{R}$, $\mu := (\mu_1,\ldots,\mu_k)$ and $\Sigma := \operatorname{diag}\left(\sigma_1^2,\ldots,\sigma_k^2\right)$.

Proof Fix an $x \in \mathbb{R}^n$. Let $c_l(x) := a_l^\mathsf{T}x - b_l$, $l = 0,\ldots,k$, and let $Y(x) := \sum_{l=1}^k c_l(x)\zeta_l$ be the weighted sum of the independent sub-Gaussian random variables ζ_l. The violation probability is then $\mathbb{P}(Y(x) > -c_0(x))$. The derivation of the Chernoff bound (13.65) shows that $Y(x)$ is sub-Gaussian with parameter

$$\left(\mu(x), \sigma^2(x)\right) \;:=\; \left(\sum_l c_l(x)\mu_l, \sum_l c_l^2(x)\sigma_l^2\right)$$

that is the weighted sum of the individual parameters, i.e.,

$$\psi_{Y(x)}(\lambda) \;\le\; \mu(x)\lambda + \frac{\sigma^2(x)}{2}\lambda^2 \tag{13.70}$$

Even though we do not know whether $-c_0(x) \geq EY(x)$, we will show directly that the Chernoff bound (13.65) still bounds the violation probability. Substituting (13.70) into (13.58a) we have

$$\ln \mathbb{P}(Y(x) > -c_0(x)) \;\leq\; \inf_{\lambda \geq 0} \psi_{Y(x)}(\lambda) + c_0(x)\lambda \;\leq\; \inf_{\lambda \geq 0} (c_0(x) + \mu(x))\lambda + \frac{\sigma^2(x)}{2}\lambda^2$$

If $c_0(x) + \mu(x) \geq 0$ then the minimum on the right-hand side is 0 (a trivial bound on the tail probability), attained at the minimizer $\lambda(x) := 0$. If $c_0(x) + \mu(x) < 0$ and $\sigma^2(x) > 0$, then the minimum is $-(c_0(x) + \mu(x))^2/(2\sigma^2(x))$, attained at the minimizer $\lambda(x) := -(c_0(x) + \mu(x))/\sigma^2(x)$. Finally if $c_0(x) + \mu(x) < 0$ but $\sigma^2(x) = 0$, then $c_l(x) = 0$ for all l (since $\sigma_l > 0$). Hence $Y(x) = 0$ and $c_0(x) + \mu(x) = c_0(x) < 0$, and therefore the violation probability $\mathbb{P}(Y(x) > -c_0(x)) = \mathbb{P}(c_0(x) > 0) = 0$. This means that if $\eta(x) < 0$ and $\gamma(x) = 0$, then x is feasible for (13.67b). In all cases we therefore have

$$\ln \mathbb{P}(Y(x) > -c_0(x)) \;\leq\; -\frac{(c_0(x) + \mu(x))^2}{2\sigma^2(x)} \tag{13.71}$$

but the bound is useful only when $c_0(x) + \mu(x) < 0$.

Since $\epsilon \in (0, 1)$, $\ln \epsilon < 0$. A sufficient condition for the chance constraint in (13.67b) to hold is therefore $c_0(x) + \mu(x) < 0$ and (13.71) holds, i.e.,

$$-\frac{(c_0(x) + \mu(x))^2}{2\sigma^2(x)} \;\leq\; \ln \epsilon$$

(If $\sigma^2(x) = 0$, then x is feasible as discussed above and this inequality holds trivially.) Hence x is feasible for (13.67b) if $\sqrt{2\ln(1/\epsilon)}\sigma(x) \leq -(c_0(x) + \mu(x))$, or

$$\sqrt{2\ln(1/\epsilon)}\sqrt{\sum_l \sigma_l^2 c_l^2(x)} \;\leq\; -\left(c_0(x) + \sum_l \mu_l c_l(x)\right)$$

Substituting $c_l(x) := a_l^\mathsf{T} x - b_l$, $l = 0, \ldots, k$, yields the constraint in the second-order cone program (13.69). □

We compare three formulations of an uncertain linear program in the next example.

Example 13.6 (LPs with bounded uncertainty)**.** Consider the uncertain linear program

$$\min_{x \in \mathbb{R}^n} \; c^\mathsf{T} x \quad \text{s.t.} \quad (a_0 + a_1\zeta_1 + a_2\zeta_2)^\mathsf{T} x \;\leq\; 0 \tag{13.72}$$

where $c, a_l \in \mathbb{R}^n$ and $\zeta := (\zeta_1, \zeta_2)$ is an uncertain parameter taking value in $Z_\infty := \{\zeta \in \mathbb{R}^2 : |\zeta_l| \leq 1, l = 1, 2\}$. We consider three formulations of the uncertain linear program.

1 The robust counterpart of (13.72) is:

$$\min_{x \in \mathbb{R}^n} \; c^\mathsf{T} x \quad \text{s.t.} \quad a_0^\mathsf{T} x + \max_{\zeta \in Z_\infty} (a_1\zeta_1 + a_2\zeta_2)^\mathsf{T} x \;\leq\; 0 \tag{13.73}$$

Theorem 13.1 says that the robust counterpart is equivalent to the linear program: $\min_{x\in\mathbb{R}^n} c^\mathsf{T}x$ s.t. $x \in X_1$ where

$$X_1 := \left\{x \in \mathbb{R}^n : a_0^\mathsf{T}x + \hat{A}x \le 0\right\} \quad \text{with} \quad \hat{A} := \begin{bmatrix} (+a_1+a_2)^\mathsf{T} \\ (+a_1-a_2)^\mathsf{T} \\ (-a_1+a_2)^\mathsf{T} \\ (-a_1-a_2)^\mathsf{T} \end{bmatrix}$$

2 The chance constrained formulation of (13.72) is:

$$\min_{x\in\mathbb{R}^n} \; c^\mathsf{T}x \quad \text{s.t.} \quad \mathbb{P}\left((a_0 + a_1\zeta_1 + a_2\zeta_2)^\mathsf{T}x \le 0\right) \;\ge\; 1-\epsilon \tag{13.74}$$

where $\epsilon \in (0,1)$ and $\mathbb{P}$ defines a probability distribution on Z_∞. Denote the chance constrained feasible set by X_2.

3 Suppose ζ_l are independent zero-mean random variables. Since each ζ_l takes value in a bounded interval $[-1,1]$, Hoeffding's Lemma 13.10 below implies that ζ_l are (independent) sub-Gaussian with variance factor $(b-a)^2/4 := 1$, i.e., they satisfy (13.68) with $\mu_l := 0$ and $\sigma_l^2 := 1$, so that $\hat{a}_0 = a_0$ and Σ is the identity matrix in (13.69). Theorem 13.9 then implies that a safe approximation of (13.74) is the following second-order cone program:

$$\min_{x\in\mathbb{R}^n} \; c^\mathsf{T}x \quad \text{s.t.} \quad a_0^\mathsf{T}x + r\|Ax\|_2 \le 0 \tag{13.75}$$

where $r := \sqrt{2\ln(1/\epsilon)}$ and $A := [a_1 \; a_2]^\mathsf{T}$. The feasible set X_3 is the pre-image of the standard second order cone K_{soc} under an affine transformation:

$$X_3 \;:=\; \left\{x \in \mathbb{R}^n : \begin{bmatrix} A \\ -(1/r)a_0^\mathsf{T} \end{bmatrix} x \in K_{\text{soc}}\right\}$$

and is itself a convex cone.

Both X_1 and X_3 are convex and contained in the feasible set X_2 of (13.74) which may be nonconvex. It does not however necessarily hold that $X_1 \subseteq X_3$, i.e., the robust formulation may not be more conservative than the safe approximation. To see this, Theorem 13.1 says that the second-order cone program (13.75) is equivalent to a robust linear program with the SOC uncertainty set $Z_2 := \{\zeta \in \mathbb{R}^2 : \|\zeta\|_2 \le \sqrt{2\ln(1/\epsilon)}\}$:

$$\min_{x\in\mathbb{R}^n} \; c^\mathsf{T}x \quad \text{s.t.} \quad a_0^\mathsf{T}x + \max_{\zeta\in Z_2} (a_1\zeta_1 + a_2\zeta_2)^\mathsf{T}x \;\le\; 0$$

Compared with (13.73), neither Z_∞ nor Z_2 may contain the other, depending on the value of ϵ, and hence neither X_1 nor X_3 may contain the other. This is illustrated in Figure 13.2 for $n = 2$ and $e^{-1} < \epsilon < e^{-1/2}$. □

Hoeffding's lemma for bounded Y.

We have seen above sub-Gaussian random variables have convenient Chernoff bounds. Hoeffding's lemma shows that a *zero-mean* random variable with bounded support $[a,b]$ is always sub-Gaussian with variance factor $(b-a)^2/4$. It is used in Example

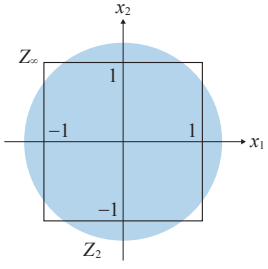


Figure 13.2 Example 13.6: neither Z_∞ nor Z_2 may contain the other, depending on ϵ.

13.6 and will be used to prove Theorem 13.12 that bounds the tail probability of a martingale with bounded increments. The proof of the Hoeffding's lemma relies on a useful technique called change of measure, which we now explain.

Given a probability measure represented by the distribution function F_Z, let a function $L(x)$ and another probability measure on the same probability space, represented by the distribution function F_Z, satisfy

$$dF_Z(x) \;=\; L(x)dF_Y(x) \tag{13.76a}$$

which means that $\int_A dF_Z(x) = \int_A L(x)dF_Z(x)$ for any (measurable) set A. If they have probability density functions f_Z and f_Y respectively then (13.76) means

$$f_Z(x) \;=\; L(x)f_Y(x)$$

The function $L(x)$ is called the *likelihood ratio* of the distribution functions F_Z and F_Y. A consequence of (13.76a) is that for any (measurable) function g, the expectation $E_Z(g(Z))$ under distribution F_Z can be computed under F_Y instead according to

$$E_Z(g(Z)) \;:=\; \int g(z)dF_Z(z) \;=\; \int g(y)L(y)dF_Y(y) \;=:\; E_Y(g(Y)L(Y)) \tag{13.76b}$$

This is used e.g. in importance sampling to speed up simulations where a rare event under distribution F_Z can be much more efficiently sampled under a modified distribution F_Y, i.e., instead of generating N samples $\{z_i\}$ under F_Z (a rare event) to estimate $E_Z(g(Z))$ by $(1/N)\sum_i g(z_i)$ we generate n samples $\{y_i\}$ under F_Y (not a rare event) to estimate $E_Y(g(Y)L(Y))$ by $(1/n)\sum_i g(y_i)L(y_i)$ (Exercise 13.17). The required number n of samples can be much smaller than N for the same variance. Due to (13.76b) we refer to (13.76) as a *change of measure* from F_Z to F_Y through the likelihood ratio $L(x)$. For the change of measure to be well defined, the probability measures and the likelihood ratio must satisfy two conditions:

- It is necessary that any event that is impossible under the probability measure (represented by) F_Y is also impossible under F_Z, i.e., for any A,

$$\int_A dF_Y(y) = 0 \quad\Rightarrow\quad \int_A dF_Z(y) = 0 \tag{13.77a}$$

In this case the probability measure F_Z is said to be *absolutely continuous* with respect to F_Y. The *Radon Nikodym theorem* says that absolutely continuity is also sufficient, i.e., if F_Z is absolutely continuous with respective to F_Y then there exists a likelihood ratio $L(x)$ such that they satisfy (13.76). The likelihood ratio is also called the *Radon Nikodym* derivative of probability measure F_Z with respect to F_Y and denoted by $\frac{dF_Z(x)}{dF_Y(x)} = L(x)$. This condition implies that, e.g., we can change a Gaussian distribution $F_Z := N(\mu, \sigma^2)$ to a standard Gaussian $F_Y := N(0,1)$, but not to an exponential distribution $F_Y(y) = 1 - e^{-\lambda y}$ which is nonzero only for $y > 0$. (An exponential distribution is absolutely continuous with respect to an Gaussian distribution, but not vice versa.)

- The likelihood ratio $L(x)$ must satisfy $L(x) \geq 0$ (almost surely with respect to F_Y) and be normalized:

$$\int L(x) dF_Y(x) \;=\; E_Y(L(Y)) \;=\; 1 \tag{13.77b}$$

Lemma 13.10 (Hoeffding's lemma)**.** Let Y be a zero-mean random variable taking values in a bounded interval $[a,b]$. Then

$$\psi_Y(\lambda) \;:=\; \ln E(e^{\lambda Y}) \;\leq\; \frac{(b-a)^2}{8}\lambda^2, \qquad \lambda \in \mathbb{R}$$

i.e., Y is sub-Gaussian with variance factor $(b-a)^2/4$.

Proof First observe that any random variable Z with bounded support on $[a,b]$, whether or not $EZ = 0$, satisfies $\mathrm{var}(Z) \leq (b-a)^2/4$ because

$$\left| Z - \frac{a+b}{2} \right| \;\leq\; \frac{b-a}{2}$$

and hence $\mathrm{var}(Z) = \mathrm{var}(Z - (a+b)/2) \leq (b-a)^2/4$ because for any random variable X, $|X| \leq c$ implies that $E(X - EX)^2 \leq c^2$.

Second, since Y takes value in a bounded set, the bounded convergence theorem implies that $\frac{d}{d\lambda} E(g(Y)) = E\left(\frac{d}{d\lambda} g(Y)\right)$ for any (measurable) function g on $\mathbb{R}$. Hence

$$\psi_Y''(\lambda) \;=\; E_Y\left(Y^2 \cdot \frac{e^{\lambda Y}}{E_Y e^{\lambda Y}}\right) - \left(E_Y\left(Y \cdot \frac{e^{\lambda Y}}{E_Y e^{\lambda Y}}\right)\right)^2, \qquad \lambda \in \mathbb{R} \tag{13.78}$$

where we have written E_Y to emphasize that the expectation is taken with respect to the probability distribution F_Y of the random variable Y. Consider a random variable Z that takes value in the same bounded interval $[a,b]$ whose distribution function F_Z is obtained from F_Y according to the following change of measure:

$$dF_Z(x) \;=\; \frac{e^{\lambda x}}{E_Y(e^{\lambda Y})} dF_Y(x) \;=:\; L(x) dF_Y(x)$$

In particular F_Z is absolutely continuous with respect to F_Y. The likelihood ratio

$L(x) := e^{\lambda x}/E_Y(e^{\lambda Y}) \geq 0$ for all x and satisfies $E_Y(L(Y)) = 1$. Hence (13.77) is satisfied. Therefore (13.76b) implies

$$E_Y\left(g(Y)\cdot\frac{e^{\lambda Y}}{E_Y(e^{\lambda Y})}\right) \;=\; E_Y(g(Y)L(Y)) \;=\; E_Z(g(Z))$$

for any function g. Substituting into (13.78) we have

$$\psi_Y''(\lambda) \;=\; E_Z\left(Z^2\right) - (E_Z Z)^2 \;=\; \mathrm{var}(Z) \;\leq\; \frac{(b-a)^2}{4}, \qquad \lambda\in\mathbb{R} \tag{13.79}$$

where the inequality follows since Z takes value in the bounded interval $[a,b]$.

Finally notice that $EY = 0$ implies that $\psi_Y(0) = 0$ and $\psi_Y'(0) = 0$. Hence Taylor expansion implies that, for some $\mu\in[0,\lambda]$,

$$\psi_Y(\lambda) \;=\; \psi_Y(0)+\psi_Y'(0)\lambda+\frac{1}{2}\psi_Y''(\mu)\lambda^2 \;\leq\; \frac{(b-a)^2}{8}\lambda^2, \qquad \lambda\in\mathbb{R}$$

where the inequality follows from (13.79). □

Azuma-Hoeffding inequality.

The Azuma-Hoeffding inequality is useful in bounding the sum of bounded random variables $\sum_i X_i$. We will first derive a bound for when X_i are independent zero-mean random variables and then extend it to the case where X_i need not be independent but forms a martingale with bounded increment $|X_i - X_{i-1}|$.

Let $Y_n := (1/n)\sum_i (X_i - EX_i)$ be the sample mean of independent and centered random variables $X_i - EX_i$ with $EX_i < \infty$. The conjugate of its log-moment generating function is, from (13.59b),

$$\psi_Y^*(t) \;=\; \sup_{\lambda\in\mathbb{R}} \sum_i (t\lambda - \psi_i(\lambda)) \tag{13.80}$$

where ψ_i are the log moment-generating functions of the centered random variables $X_i - EX_i$. The application of Hoeffding's Lemma 13.10 leads to a concentration inequality for the sample mean Y_n.

Theorem 13.11 (Azuma-Hoeffding inequality)**.** Let $X_1,\ldots,X_n$, be independent with $X_i\in[a_i,b_i]$, then

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (X_i - EX_i) \geq t\right) \;\leq\; \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i-a_i)^2}\right), \qquad t\geq 0$$

Proof Let $Y_n := (1/n)\sum_{i=1}^n (X_i - EX_i)$ be the sample mean of the independent and

centered random variables $X_i - EX_i$. Chernoff bound gives, for $t \geq 0$,

$$\begin{aligned}\mathbb{P}(Y_n \geq t) &\leq e^{-\psi^*_{Y_n}(t)} = \exp\left(\inf_{\lambda\in\mathbb{R}} \sum_i (\psi_i(\lambda) - t\lambda)\right) \\ &\leq \exp\left(\inf_{\lambda\in\mathbb{R}}\left(\lambda^2 \sum_i \frac{(b_i-a_i)^2}{8} - nt\lambda\right)\right) = \exp\left(-\frac{2n^2t^2}{\sum_i (b_i-a_i)^2}\right)\end{aligned}$$

where the first equality follows from (13.80) and the second inequality follows from Hoeffding's Lemma 13.10 since $X_i - EX_i \in [a_i - EX_i, b_i - EX_i]$. □

The bound in Theorem 13.11 can be generalized to the case where X_i are not necessarily independent, but form a martingale. A discrete-time stochastic process $X_0, X_1, \ldots,$ is a *martingale* if

- $E|X_n| < \infty$.
- $E(X_n|X_0, \ldots, X_{n-1}) = X_{n-1}$.

This implies that the total change $X_n - X_0$ by any time n has zero mean:

$$E(X_n - X_0) = E(E(X_n - X_0)|X_0, \ldots, X_{n-1}) = E(X_{n-1} - X_0) = \cdots = 0$$

The application of Hoeffding's Lemma 13.10 leads to a concentration inequality for a martingale with bounded increments.

Theorem 13.12 (Azuma-Hoeffding inequality)**.** Let $X_0, X_1, \ldots,$ be a martingale with bounded increments $|X_n - X_{n-1}| \leq \sigma_n$. Then for any $n \geq 1$,

$$\mathbb{P}(X_n - X_0 \geq t) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right), \qquad t \geq 0$$

Proof Without loss of generality we can assume $X_0 = 0$; otherwise we can consider the martingale $Y_0, Y_1, \ldots,$ with $Y_0 := 0$ and $Y_n := X_n - X_0$. Chernoff bound gives

$$\begin{aligned}\mathbb{P}(X_n \geq t) &\leq \min_{\lambda\in\mathbb{R}} E\frac{\exp(\lambda X_n)}{\exp(\lambda t)} = \min_{\lambda\in\mathbb{R}} e^{-\lambda t} E\exp\left(\lambda \sum_{i=1}^n (X_i - X_{i-1})\right) \\ &= \min_{\lambda\in\mathbb{R}} e^{-\lambda t} E\left(\exp\left(\lambda \sum_{i=1}^{n-1}(X_i - X_{i-1})\right) E\left(\exp(\lambda(X_n - X_{n-1}))|X_0, \ldots, X_{n-1}\right)\right)\end{aligned} \tag{13.81}$$

where the second equality uses $E(g(X)Y) = E(g(X)E(Y|X))$. Since $X_0, X_1, \ldots,$ is a martingale with bounded increment, $E(X_n|X_0, \ldots, X_{n-1}) = X_{n-1}$. Hence, given $X_0, \ldots, X_{n-1}$, $X_n - X_{n-1}$ is a zero-mean random variable that takes value in $[-\sigma_n, \sigma_n]$. Hoeffindg's Lemma 13.10 implies that

$$E\left(\exp(\lambda(X_n - X_{n-1}))|X_0, \ldots, X_{n-1}\right) \leq \exp\left(\frac{\sigma_n^2}{2}\lambda^2\right)$$

Substitute into (13.81) to get

$$\mathbb{P}(X_n \geq t) \;\leq\; \min_{\lambda\in\mathbb{R}} e^{-\lambda t} E\left(\exp\left(\lambda \sum_{i=1}^{n-1}(X_i - X_{i-1})\right)\right)\exp\left(\frac{\sigma_n^2}{2}\lambda^2\right)$$

Repeating this calculation for $X_{n-1} - X_{n-2}, \ldots, X_1 - X_0$, we arrive at

$$\mathbb{P}(X_n \geq t) \;\leq\; \min_{\lambda\in\mathbb{R}} e^{-\lambda t}\exp\left(\frac{\lambda^2}{2}\sum_{i=1}^{n}\sigma_i^2\right) \;=:\; \min_{\lambda\in\mathbb{R}}\exp\left(\frac{s_n^2}{2}\lambda^2 - t\lambda\right)$$

where $s_n^2 := \sum_{i=1}^n \sigma_i^2$. The minimizer is $\lambda_n := t/s_n^2$ and $\mathbb{P}(X_n \geq t) \leq \exp\left(-\frac{t^2}{2s_n^2}\right)$. □

The two-sided tail probabilities in Theorems 13.11 and 13.12 are bounded by twice the bounds in these theorems:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}|X_i - EX_i| \geq t\right) \;\leq\; 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

$$\mathbb{P}(|X_n - X_0| \geq t) \;\leq\; 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right)$$

Summary.

The inequalities introduced in this subsection are some of the most basic inequalities in probability and are summarized in Table 13.1.

13.3 Convex scenario optimization

Consider the robust program (13.5) studied in Chapter 13.1 with a linear cost: [4]

$$\text{RCP}: \qquad c^*_{\text{RCP}} \;:=\; \min_{x\in X\subseteq\mathbb{R}^n} c^{\mathsf{T}}x \quad \text{s.t.} \quad h(x,\zeta) \leq 0,\ \zeta \in Z \subseteq \mathbb{R}^k \tag{13.82}$$

where $c \in \mathbb{R}^n$, $\zeta \in \mathbb{R}^k$ is an uncertain parameter taking value in the uncertainty set Z, $h : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^m$ is a convex (and hence continuous) function in x for every $\zeta \in Z$, and X is a nonempty closed convex set.[5] Even though (13.82) is convex, it is semi-infinite and hence generally intractable. Moreover requiring constraint satisfaction for all possible uncertain parameters in Z can be too conservative. The chance constrained formulation studied in Chapter 13.2 is less conservative as it requires constraint satisfaction only

[4] The linear cost function does not lose generality; see Remark 13.1.

[5] We can also assume without loss of generality that $h : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}$ is a scalar-valued function because otherwise, $h(x,\zeta) \leq 0$ can be replaced by the single constraint $\max_i h_i(x,\zeta) \leq 0$. Note however that if h is scalar-valued then x is infeasible if $h(x,\zeta) > 0$, but if h is vector-valued then x is infeasible if $h_i(x,\zeta) > 0$ for at least one i, not $h(x,\zeta) > 0$.

	Inequality	Assumptions
Markov's	$\mathbb{P}(Y \geq t) \leq \frac{E(\phi(Y))}{\phi(t)}$	$\phi(Y) \geq 0$, $\phi(t) > 0$, $EY < \infty$
Chebyshev's	$\mathbb{P}(\lvert X - EX\rvert \geq t) \leq \mathrm{var}(X)/t^2$	$\mathrm{var}(X) < \infty$, $t > 0$
	$\mathbb{P}\left(\left\lvert \frac{1}{n}\sum_i (X_i - EX_i)\right\rvert \geq t\right) \leq \frac{(1/n)\sum_i \mathrm{var}(X_i)}{nt^2}$	$\mathrm{var}(X_i) < \infty$, independent X_i, $t > 0$
Chernoff	$\mathbb{P}(Y \geq t) \leq e^{-\psi_Y^*(t)}$	$EY < \infty$, $t \geq EY$
	$\mathbb{P}(Y \geq t) \leq \exp\left(-\sup_{\lambda \geq 0}(t\lambda - \psi_Y(\lambda))\right)$	$EY < \infty$, $t \in \mathbb{R}$
	$\mathbb{P}\left(\frac{1}{n}\sum_i X_i \geq t\right) \leq e^{-n\psi_{X_1}^*(t)}$	iid X_i, $EX_i < \infty$, $t \geq E(X_1)$
sub-Gaussian	$\mathbb{P}(Y \geq t) \leq e^{-(t-\mu)^2/2\sigma^2}$	sub-Gaussian Y, $EY < \infty$, $t \geq EY$
	$\mathbb{P}(\sum_i a_i X_i \geq t) \leq \exp\left(-\frac{(t-\sum_i a_i\mu_i)^2}{2\sum_i a_i^2\sigma_i^2}\right)$	indep. sub-Gaussian X_i, $EX_i < \infty$, $t \geq EY$
	$\mathbb{P}\left(\max_{i=1}^n X_i \geq t\right) \leq \sigma\sqrt{2\ln n}/t$	sub-Gaussian X_i, $t > 0$
Hoeffding's lemma	$\psi_Y(\lambda) \leq (1/8)(b-a)^2\lambda^2$	$EY = 0$, $Y \in [a,b]$ a.s.
Azuma-Hoeffding	$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \geq t\right) \leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i-a_i)^2}\right)$	independent zero-mean $X_i \in [a_i, b_i]$, $t \geq 0$
	$\mathbb{P}(X_n - X_0 \geq t) \leq \exp\left(-t^2/2\sum_{i=1}^n \sigma_i^2\right)$	martingale X_i, $\lvert X_i - X_{i-1}\rvert \leq \sigma_i$, $t \geq 0$

Table 13.1 Summary of concentration inequalities. $\psi_Y(\lambda) := \ln Ee^{\lambda Y}$ and $\psi_Y^*(t) := \sup_{\lambda \in \mathbb{R}}(t\lambda - \phi_Y(\lambda))$. Y is sub-Gaussian if $\psi_Y(\lambda) \leq \mu\lambda + (\sigma^2/2)\lambda^2$.

with high probability rather than with probability 1. Consider the chance constrained program with a linear cost:

$$\mathrm{CCP}(\epsilon): \quad c_{\mathrm{CCP}}^*(\epsilon) := \min_{x \in X \subseteq \mathbb{R}^n} c^\mathsf{T} x \quad \text{s.t.} \quad \mathbb{P}(h(x,\zeta) \leq 0) \geq 1 - \epsilon \tag{13.83}$$

where X, c, h are the same as those in (13.82), $\zeta \in Z \subseteq \mathbb{R}^k$ is a random vector and $\mathbb{P}$ is a probability measure defined on some probability space, and $\epsilon \in (0,1)$. Solving problem (13.83) however can be challenging as it requires the knowledge of the probability measure $\mathbb{P}$ which may not be available. Moreover it requires an efficient method to evaluate the probability in order to assess the feasibility of x.

This motivates the scenario approach to uncertain optimization where N independent samples $\zeta^1, \ldots, \zeta^N$ of the uncertain parameter ζ are drawn according to the probability measure $\mathbb{P}$, leading to the following problem, called a convex *scenario program*:

$$\mathrm{CSP}(N): \quad c_{\mathrm{CSP}}^*(N) := \min_{x \in X \subseteq \mathbb{R}^n} c^\mathsf{T} x \quad \text{s.t.} \quad h(x,\zeta^i) \leq 0,\ i = 1, \ldots, N \tag{13.84}$$

Since ζ^i are random samples, the scenario program (13.84) is a randomized problem in the sense that its solution is a random variable whose value depends on the values of ζ^i. It does not require the knowledge of $\mathbb{P}$, but only a way to obtain independent samples according to $\mathbb{P}$. For instance, the uncertain parameter ζ may represent power

demand and its realizations ζ^i may be measured from a real power system without knowing the underlying distribution.

Unlike RCP (13.82) and CCP(ϵ) (13.83) which are often intractable, the scenario program (13.84) is a finite convex program for each realization of the random samples $(\zeta^1, \ldots, \zeta^N)$ and therefore can be efficiently solved if N is not too large. There is therefore a tradeoff between small computational burden (when N is small) and high likelihood of constraint satisfaction (when N is large). In this section we will study three issues:

1 *Violation probability* (Chapter 13.3.1). Given a fixed vector $x \in X \subseteq \mathbb{R}^n$ the violation probability $V(x)$ is the probability of $h_i(x,\zeta) > 0$ for at least one i, a deterministic value. A solution x_N^* of the convex scenario program CSP(N) is random, depending on the random samples $(\zeta^1, \ldots, \zeta^N)$. The violation probability $V(x_N^*)$ of the random solution x_N^* is therefore not a deterministic value, but a random variable itself. We will bound the expected value and the tail probability of $V(x_N^*)$.
2 *Sample complexity* (Chapter 13.3.5). The more sampled constraints are included in CSP(N), the more likely its optimal solution x_N^* will satisfy the chance constraint of CCP(ϵ). We will use the bounds on the expected value and the probability of $V(x_N^*)$ to derive a threshold $N(\epsilon,\beta)$ to guarantee that the (random) solution x_N^* will be feasible for CCP(ϵ) with probability at least $1-\beta$.
3 *Optimality guarantee* (Chapter 13.3.6). We will show that the same threshold $N(\epsilon,\beta)$ that guarantees, with probability at least $1-\beta$, the feasibility of x_N^* for CCP(ϵ) also guarantees that the optimal value $c^\mathsf{T} x_N^*$ is close to the optimal values of RCP and CCP(ϵ).

13.3.1 Violation probability $V(x_N^*)$

Let $X_\zeta := \{x \in X \subseteq \mathbb{R}^n : h(x,\zeta) \leq 0\}$. We will refer to a constraint by $h(x,\zeta) \leq 0$ or X_ζ or ζ interchangeably. The assumption that X is a closed convex set and each component h_j of h is convex (and hence continuous) in x for any $\zeta \in Z$ implies that X_ζ is a closed convex set for every $\zeta \in Z$. We may interpret X_ζ either as a deterministic set determined by a realization of ζ in Z, or a random set whose value depends on the random variable ζ; the meaning should be clear from the context. For each $x \in X$, define the *violation probability* of x as

$$V(x) \;:=\; \mathbb{P}\left(\left\{\zeta \in Z : x \notin X_\zeta\right\}\right) \tag{13.85a}$$

For a fixed $x \in X$, $V(x)$ is a deterministic value in $[0,1]$. As we will see the feasibility and sample complexity results are independent of the fine structure of the constraint function h or the probability measure $\mathbb{P}$, except through the random constraint set X_ζ. The CCP(ϵ) (13.83) with the deterministic constraint $\mathbb{P}\left(x \in X_\zeta\right) \geq 1-\epsilon$ can be

equivalently stated as:

$$\text{CCP}(\epsilon): \qquad c^*_{\text{ccp}}(\epsilon) \;:=\; \min_{x\in X\subseteq\mathbb{R}^n} c^{\mathsf{T}}x \quad \text{s.t.} \quad V(x)\le\epsilon$$

For each integer $N \ge n$, we interpret $(\zeta^1,\ldots,\zeta^N) \in \mathbb{Z}^N$ either as deterministic vectors realized by independent samples of $\zeta \in Z \subseteq \mathbb{R}^k$ under the probability measure $\mathbb{P}$, or as iid random vectors with the product measure $\mathbb{P}^N$, depending on the context. The randomized problem CSP(N) (13.84) can be equivalently stated as:

$$\text{CSP}(N): \qquad c^*_{\text{CSP}}(N) \;:=\; \min_{x\in X\subseteq\mathbb{R}^n} c^{\mathsf{T}}x \quad \text{s.t.} \quad x \in X_{\zeta^1}\cap\cdots\cap X_{\zeta^N}$$

An optimal solution x^*_N of CSP(N), if exists, is feasible for the chance constrained program (13.83) when $V(x^*_N) \le \epsilon$. Note however that x^*_N is a random variable under probability measure $\mathbb{P}^N$, depending on $(\zeta^1,\ldots,\zeta^N)$, i.e., $V\left(x^*_N\right)$ is the *conditional* violation probability:

$$V\left(x^*_N\right) \;:=\; \mathbb{P}\left(\left\{\zeta\in Z : x^*_N\notin X_\zeta\right\}\middle|\left(\zeta^1,\ldots,\zeta^N\right)\right) \tag{13.85b}$$

Hence the violation probability $V(x^*_N)$ itself is a random variable under $\mathbb{P}^N$. It may be greater or smaller than ϵ, i.e., x^*_N may or may not be feasible for CCP(ϵ) (13.83). We emphasize that $V(x^*_N)$ is not the unconditional probability $\mathbb{P}^{N+1}\left(x^*_N \notin X_{\zeta^{N+1}}\right)$. While the former is a random variable with probability measure $\mathbb{P}^N$, the latter is a deterministic value. Their relation is

$$\mathbb{P}^{N+1}\left(x^*_N\notin X_{\zeta^{N+1}}\right) \;=\; \int_{Z^N} V\left(x^*_N\right)\mathbb{P}^N\left(d\zeta^1,\ldots,d\zeta^N\right) \;=\; E^N\left(V\left(x^*_N\right)\right) \tag{13.86}$$

i.e., the expected value of the violation probability $V(x^*_N)$ turns out to be the unconditional probability $\mathbb{P}^{N+1}\left(x^*_N \notin X_{\zeta^{N+1}}\right)$.

Main result.

Intuitively a larger N will produce an optimal solution x^*_N that is more likely to satisfy the chance constraint $V(x^*_N) \le \epsilon$. A reasonable approach is then to choose N large enough to ensure that the expected value $E^N\left(V(x^*_N)\right) \le \beta$ under $\mathbb{P}^N$ for a sufficiently small β. Another approach is to ensure that the probability $\mathbb{P}^N\left(V(x^*_N) > \epsilon\right) \le \beta$. In this subsection we show in Theorems 13.14 and 13.15 that

$$E^N\left(V(x^*_N)\right) \;\le\; \frac{n}{N+1}, \qquad\qquad \mathbb{P}^N\left(V(x^*_N)>\epsilon\right) \;\le\; \sum_{i=0}^{n-1}\binom{N}{i}\epsilon^i(1-\epsilon)^{N-i}$$

and that both bounds are tight for a class of problems called fully supported problems defined in Definition 13.3. The bound on $E^N\left(V(x^*_N)\right)$ decreases at a rate $\sim 1/N$. The bound on $\mathbb{P}^N\left(V(x^*_N)>\epsilon\right)$ is a Binomial tail. Hence it is in $(0,1)$ as long as $N \ge n$ (equal to 1 if $N = n-1$) and decreases more rapidly as N increases. These bounds mean that if we solve CSP(N) (13.84) with a sufficiently large N, then we will obtain a random optimal solution x^*_N whose conditional violation probability $V(x^*_N)$ is small

either in expectation or probability. They translate into sample complexity studied in Chapter 13.3.5.

We make the following assumption

C13.3: Consider CSP(N) (13.84).

- X is a closed convex set and, for each $\zeta \in Z$, the components h_j of the constraint function $h(x,\zeta)$ are continuous and convex in x, so that X_ζ is a closed convex set.
- For each integer $N \geq n$ and each realization of $(\zeta^1,\ldots,\zeta^N)$, the feasible set of CSP(N) (13.84) (is nonempty and) has a nonempty interior. Moreover CSP(N) has a unique optimal solution denoted by x_N^*.

See Remark 13.6 when x_N^* may be non-unique.

Definition 13.3 (Uniformly supported problem)**.** Fix any $N \geq n$ and consider CSP(N) (13.84).

1 Consider a realization of $(\zeta^1,\ldots,\zeta^N) \in Z^N$. A constraint ζ^i is called a *support constraint for CSP*(N) (with respect to the realization) if its removal changes the optimal solution, i.e., $x_{N\setminus i}^* \neq x_N^*$ where $x_{N\setminus i}^*$ is the optimal solution of the scenario program CSP($N-1$) with the constraint X_{ζ^i} removed. A constraint that is not a support constraint is called a *non-support constraint* for CSP(N).
2 CSP(N) is called *uniformly supported* with s support constraints if every realization of $(\zeta^1,\ldots,\zeta^N) \in Z^N$ contains exactly $s \geq 0$ support constraints for CSP(N) with probability 1. It is called *fully supported* if it is uniformly supported with $s = n$ support constraints. It is said to have *no support constraint* if it is uniformly supported with $s = 0$ support constraint.

A support constraint must be an active constraint at the optimal point x_N^* but the converse may not hold, e.g., if $\zeta^i = \zeta^j$ (redundant constraints) then neither can be a support constraint. For a uniformly supported problem with $s \geq 1$ support constraints, the probability of $\zeta^i = \zeta^j$ must be zero. Since optimal solutions are unique (assumption C13.3), $x_{N\setminus i}^* \neq x_N^*$ is equivalent to $c^\mathsf{T} x_{N\setminus i}^* < c^\mathsf{T} x_N^*$ because otherwise, if $c^\mathsf{T} x_{N\setminus i}^* = c^\mathsf{T} x_N^*$ then both $x_{N\setminus i}^*$ and x_N^* are optimal solutions of CSP(N), contradicting the uniqueness of optimal solutions. If CSP(N) is uniformly supported with $s = 0$ support constraint, it means that, with probability 1, no realization of $(\zeta^1,\ldots,\zeta^N)$ has a single constraint that is a support constraint (e.g. all constraints are inactive at x_N^* or all active constraints are redundant). For a general problem that is not uniformly supported, different realizations of $(\zeta^1,\ldots,\zeta^N)$ may have different number of support constraints. Given a realization $(\zeta^1,\ldots,\zeta^N)$, by "the set of support constraints for CSP(N)" we mean the *unique* set of *all* support constraints for CSP(N).

An important observation is the following result of [144]. Its proof makes use of the linearity of the cost function $c^\mathsf{T} x$ and convexity of X_ζ.

Lemma 13.13. [144] For each $N \geq n$, consider CSP(N) (13.84) with a linear cost function and closed convex sets X_ζ for all $\zeta \in Z$. Then the number of support constraints is at most n for any realization of $(\zeta^1, \ldots, \zeta^N) \in Z^N$ as long as (13.84) is feasible.

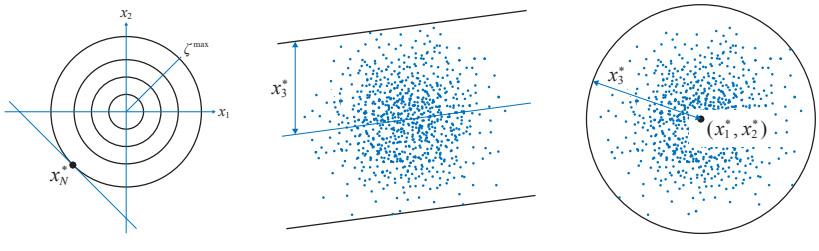
If (13.84) is infeasible the number of support constraints is upper bounded by $n+1$ [145].

Example 13.7 (Uniformly supported problems [146])**.** We consider three problems, a uniformly supported problem, a fully supported problem and a general problem; see Figure 13.3. We will derive their support constraints in Example 13.8.

1 *Uniformly supported problem*. Given N iid random squared radius ζ^i each taking value in $Z := \mathbb{R}_+$ according to an exponential distribution, we solve the scenario program CSP(N):

$$\min_{x \in \mathbb{R}^n} \sum_i x_i \quad \text{s.t.} \quad \|x\|_2^2 \leq \zeta^i, \quad i = 1, \ldots, N$$

For almost all $(\zeta^i, i = 1, \ldots N) \in \mathbb{R}_+^N$, there is exactly $s = 1$ support constraint and a unique optimal solution $x_N^* := \left(-\sqrt{\zeta^{\max}/n}, \ldots, -\sqrt{\zeta^{\max}/n}\right)$ where $\zeta^{\max} := \max_i \zeta^i$.



(a) Uniformly supported (b) Fully supported (c) General problem

Figure 13.3 Example 13.7.

2 *Fully supported problem*. We are given $N \geq 3$ points in $\mathbb{R}^2$ specified by their coordinates $\zeta^i := (a^i, b^i) \in Z := \mathbb{R}^2$, $i = 1, \ldots, N$, where (a^i, b^i) are iid samples under the Gaussian distribution over $\mathbb{R}^2$. To construct a strip of smallest vertical width that contains all the N points, we solve the CSP(N):

$$\min_{(x_1, x_2, x_3) \in \mathbb{R}^3} x_3 \quad \text{s.t.} \quad \left|b^i - (a^i x_1 + x_2)\right| \leq x_3, \quad i = 1, \ldots, N$$

See Figure 13.3. This problem is fully supported as CSP(N) has exactly $n = 3$ support constraints for *almost every* realization of $(a^i, b^i, i = 1, \ldots, N) \in \mathbb{R}^{2\times N}$.

3 *General problem*. Instead of the strip of smallest vertical width, suppose we wish to construct a circle of smallest radius that contains all the N points. Then we solve

the CSP(N):

$$\min_{(x_1,x_2,x_3)\in\mathbb{R}^3} x_3 \quad \text{s.t.} \quad \sqrt{(a^i-x_1)^2+(b^i-x_2)^2} \le x_3, \quad i=1,\ldots,N$$

with SOC constraints. This problem has 3 support constraints if the optimal circle is defined by three points on the circle or 2 support constraints if it is defined by two points on a diameter. □

The main characterization of the conditional violation probability $V(x_N^*)$ is given in the next two theorems.

Theorem 13.14 (Expectation of $V(x_N^*)$ [144, 147])**.** Fix any $N \ge n$ and suppose assumption C13.3 holds. Then

$$E^N\left(V\left(x_N^*\right)\right) = \mathbb{P}^{N+1}\left(x_N^* \notin X_{\zeta^{N+1}}\right) \le \frac{n}{N+1} \tag{13.87}$$

If CSP($N+1$) is uniformly supported with $0 \le s \le n$ support constraints then

$$E^N\left(V\left(x_N^*\right)\right) = \mathbb{P}^{N+1}\left(x_N^* \notin X_{\zeta^{N+1}}\right) = \frac{s}{N+1}$$

In particular if CSP($N+1$) has no support constraint then $E^N\left(V\left(x_N^*\right)\right)=0$.

Theorem 13.15 (Tail probability of $V(x_N^*)$ [146])**.** Fix any $N \ge n$ and suppose assumption C13.3 holds. Then

$$\mathbb{P}^N\left(V\left(x_N^*\right) > \epsilon\right) \le \sum_{i=0}^{n-1}\binom{N}{i}\epsilon^i(1-\epsilon)^{N-i} \tag{13.88}$$

If CSP(N) is uniformly supported with $1 \le s \le n$ support constraints then

$$\mathbb{P}^N\left(V\left(x_N^*\right) > \epsilon\right) = \sum_{i=0}^{s-1}\binom{N}{i}\epsilon^i(1-\epsilon)^{N-i}$$

In particular if CSP(N) has no support constraint then $\mathbb{P}^N\left(V\left(x_N^*\right)=0\right)=1$.

Remark 13.5 (Improved bounds)**.** 1 If an a.s. upper bound $s^{\max} \le n$ on the number of support constraints for CSP($N+1$) is known, then the bound in (13.87) can be improved to (see (13.94)):

$$E^N\left(V\left(x_N^*\right)\right) \le \frac{s^{\max}}{N+1}$$

2 If an a.s. upper bound $t^{\max} \le n$ on the number of "generalized support constraints" (see Definition 13.5) for CSP(N) is known, then the bound in (13.88) can be improved to (see (13.109)):

$$\mathbb{P}^N\left(V\left(x_N^*\right) > \epsilon\right) \le \sum_{i=0}^{t^{\max}-1}\binom{N}{i}\epsilon^i(1-\epsilon)^{N-i}$$

These improved bounds can be useful for power system applications because large

OPF problems often have a large n but very small $s^{\max}$ and $t^{\max}$, e.g., only a few lines are congested at only a few times a year in a multi-stage OPF problem. □

Remark 13.6 (Generality of bounds)**.** 1 It is remarkable that the bounds in Theorems 13.14 and 13.15 depend only on (n,N) and ϵ, and not on the probability measure $\mathbb{P}$, the cost function, or the structure of the constraint sets X_ζ. The cost function and the structure of the constraints only affect whether the problem is uniformly supported and hence the tightness of the bound. (The linearity of the cost function $c^\mathsf{T}x$ and convexity of X_ζ are used in the proof of Lemma 13.13 [144]. The linear cost does not lose generality as we can always replace a nonlinear cost $\min_x f(x)$ by the linear cost $\min_{x,t} t$ with the additional constraint $f(x) \le t$.)

2 The assumption in C13.3 on the existence and uniqueness of the optimal solution x_N^* is not important. It is shown in [144, 146] that if optimal solutions are nonunique, a tie-breaking rule can be used to produce a unique solution, e.g., choose the optimal solution with minimum Euclidean norm, and Theorems 13.14 and 13.15 hold unchanged. If optimal solutions may not exist then the expectation in Theorem 13.14 should be replaced by conditional expectation, conditioned on the subset of Z^N on which an optimal solution x_N^* exists, and the probability $\mathbb{P}^N(V(x_N^*) > \epsilon)$ in Theorem 13.15 should be replaced by $\mathbb{P}^N\left(x_N^* \text{ exists and } V(x_N^*) > \epsilon\right)$. (See also [145] for discussions on infeasible problems.) □

We prove Theorems 13.14 and 13.15 in the next two subsection.

13.3.2 Proof: bound on $E^N(V(x_N^*))$

Partitioning of Z^N.

The violation probability $V(x_N^*)$ is related to support constraints through the following useful characterization.

Lemma 13.16 ($V(x_N^*)$)**.** Consider CSP(N) and CSP($N+1$).

1 $x_N^* \notin X_{\zeta^{N+1}} \Leftrightarrow X_{\zeta^{N+1}}$ is support constraint for CSP($N+1$).

2 $V\left(x_N^*\right) := \mathbb{P}\left(x_N^* \notin X_{\zeta^{N+1}} \middle| \left(\zeta^1,\ldots,\zeta^N\right)\right)$ satisfies:

$$V\left(x_N^*\right) \;=\; \mathbb{P}\left(X_{\zeta^{N+1}} \text{ is support constraint for CSP}(N+1) \middle| \left(\zeta^1,\ldots,\zeta^N\right)\right)$$

Proof Suppose $x_N^* \notin X_{\zeta^{N+1}}$. Then $X_{\zeta^{N+1}}$ must be a support constraint of CSP($N+1$) with $N+1$ constraints because otherwise, $x_N^* = x_{N+1}^*$ where x_N^* is the optimal solution of CSP(N) after the constraint $X_{\zeta^{N+1}}$ is removed. This contradicts $x_N^* \notin X_{\zeta^{N+1}}$. Conversely, suppose $X_{\zeta^{N+1}}$ is a support constraint for CSP($N+1$). If $x_N^* \in X_{\zeta^{N+1}}$ then x_N^* is feasible, and hence optimal, for CSP($N+1$). Hence $x_N^* = x_{N+1}^*$ since optimal

solutions are unique (assumption C13.3). This contradicts $X_{\zeta^{N+1}}$ being a support constraint for CSP$(N+1)$, and hence $x_N^* \notin X_{\zeta^{N+1}}$.

Part 2 then follows from (13.85b). □

A key to the proof of both Theorems 13.14 and 13.15 is the partitioning of Z^N according to support constraints. Fix any $N \geq n$ and consider CSP(N). The independent samples $(\zeta^1, \ldots, \zeta^N)$ take values in Z^N. To simplify notation we will use $\tilde{\zeta} \in Z$ to denote a single vector and $\zeta := (\zeta^1, \ldots, \zeta^N) \in Z^N$ to denote a collection of vectors.

For $s = 1, \ldots, n$, let $I^s \subseteq \{1, \ldots, N\}$ be an index set with $|I^s| = s$ indices and let

$$Z^N(I^s) := \left\{\zeta \in Z^N : \left(X_{\zeta^i}, i \in I^s\right) \text{are all the support constraints in } \zeta\right\} \tag{13.89a}$$

$$Z^N(s) := \bigcup_{I^s} Z^N(I^s) \tag{13.89b}$$

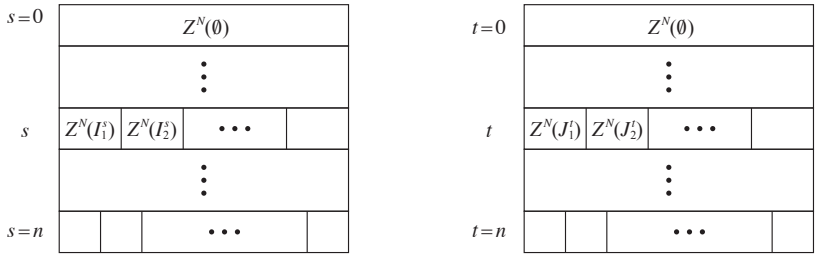
i.e., $Z^N(s)$ is the set of vectors $\zeta \in Z^N$ that contain exactly $1 \leq s \leq n$ support constraints (Lemma 13.13 implies $s \leq n$), and $Z^N(I^s)$ is the subset of $Z^N(s)$ whose support constraints are indexed by I^s. For $s = 0$, we define $I^0 := \emptyset$ and

$$Z^N(0) := Z^N(I^0) := \left\{\zeta \in Z^N : \text{CSP}(N) \text{ has no supp. const.}\right\} \tag{13.89c}$$

Clearly $Z^N(I^s)$ and $Z^N(J^s)$ are disjoint if I^s and J^s are distinct index sets each with s indices and there are $\binom{N}{s}$ distinct index sets. Moreover $Z^N(I^s)$ partition Z^N, first according to $Z^N(s)$ with $s = 0, \ldots, n$ support constraints and then according to different index sets I^s in $Z^N(s)$ (see Figure 13.4(a)):

$$Z^N = \bigcup_{s=0}^{n} Z^N(s) = \bigcup_{s=0}^{n} \bigcup_{I^s} Z^N(I^s) \tag{13.90}$$

This partitioning is useful in proving the bound on $E^N(V(x_N^*))$ in Theorem 13.14. The problem CSP(N) is uniformly supported with $0 \leq s \leq n$ support constraints if and



(a) Theorem 13.14 (b) Theorem 13.15

Figure 13.4 Partitioning of Z^N according to (a) support constraints in I^s and (b) generalized support constraints in J^t for CSP(N) that is not uniformly supported (see Chapter 13.3.4).

only if $Z^N = Z^N(s)$ and $Z^N(s') = \emptyset$ for all $s' \neq s$.[6] These concepts are illustrated in the next example.

Example 13.8 (Uniformly supported problems [146])**.** We partition the uncertainty sets Z^N for the three problems in Example 13.7.

1 The support constraint is (defined by) $\zeta^{\max} := \max_i \zeta^i$. The index sets I^1 for support constraints take the form $I^1 := \{i\}$ if $\zeta^i = \zeta^{\max}$, $i = 1, \ldots, N$, and

$$Z^N(\{i\}) \;=\; \left\{\zeta := (\zeta^1, \ldots, \zeta^N) \in \mathbb{R}^N_+ : \zeta^i \text{ is the support constraint}\right\}$$

Recall that the unique solution x^*_N depends on $\zeta^{\max}$. For the same x^*_N, any one of $\zeta^1, \ldots, \zeta^N$ can be the support constraint and therefore $\mathbb{P}(Z^N(\{i\})) = 1/N$ since ζ^i are iid. If more than one ζ^i attains the maximum in $\zeta^{\max}$, then none of $\zeta^1, \ldots, \zeta^N$ is a support constraint, but this is a zero-probability event under the exponential distribution (otherwise both $Z^N(1)$ and $Z^N(0)$ have nonzero probabilities and the problem is not uniformly supported). Therefore, for $s \neq 1$, $I^s = \emptyset$ and $Z^N(s) = \emptyset$ with probability 1, and

$$Z^N \;=\; Z^N(1) \;=\; \bigcup_{i=1}^{N} Z^N(\{i\}) \qquad \text{(with probability 1)}$$

2 The second problem in Example 13.7 is fully supported, i.e., it has $s = 3$ support constraints for almost every $\zeta := (\zeta^1, \ldots, \zeta^N)$ and hence $Z^N = Z^N(3)$. Suppose $I^3 := \{1,2,3\}$ are 3 support constraints, i.e., the three points $(a_1, b_1), (a_2, b_2), (a_3, b_3)$ define the optimal strip $S \subseteq \mathbb{R}^2$ with minimum vertical width containing all the N points $\zeta := (\zeta^1, \ldots, \zeta^N)$. Then

$$Z^N(\{1,2,3\}) \;=\; \left\{\left(\zeta^1, \zeta^2, \zeta^3, \zeta^4, \ldots, \zeta^N\right) \in Z^N : \zeta^i \in S, i = 4, \ldots, N\right\}$$

and $\mathbb{P}(Z^N(\{1,2,3\}))$ is the probability that $\zeta^i \in S$, $i \geq 4$. Even though some of ζ^i, $i \geq 4$, may lie on the boundary of S in which case $\zeta^1, \zeta^2, \zeta^3$ may not be support constraints, these are zero-probability events under the Gaussian distribution, conditioned on I^3. Therefore for $s \neq 3$, $Z^N(s) = \emptyset$ for almost every $\zeta \in Z^N$. For the same x^*_N, the three support constraints (points on the boundary) that define the (same) optimal strip S can be any three of $\zeta^1, \ldots, \zeta^N$. Hence $\mathbb{P}(Z^N(\{1,2,3\})) = \binom{N}{3}^{-1}$.

3 For the third problem, the optimal circle $C \subset \mathbb{R}^2$ with minimum radius that contains all N points is defined either by three points on the circle or two points on a diameter. If the distribution is not Gaussian, but nonzero only at grid points $(i, j) \in \mathbb{R}^2$ for a finite number of integers i, j, then CSP(N) can have 0, 1, 2, or 3 support constraints and $Z^N = \sum_{s=0}^{3} Z^N(s)$. For the same x^*_N that defines a C, any one of $\zeta^1, \ldots, \zeta^N$ can be the support constraint conditioned on $Z^N(1)$, any two of them can be the

[6] This should be interpreted as $\mathbb{P}(Z^N(s')) = 0$ even when $Z^N(s') \neq \emptyset$ for $s' = s$. We often simplify exposition by omitting the qualification of "almost surely (a.s.)."

support constraints conditioned on $Z^N(2)$, and any three of them can be the support constraints conditioned on $Z^N(3)$. Hence $\mathbb{P}(Z^N(I^s)|Z^N(s))$ is $\binom{N}{s}^{-1}$. □

The next result formalizes the intuition in Example 13.8 that the conditional probability $\mathbb{P}^N\left(Z^N(I^s)\middle|Z^N(s)\right)$ is the same for all index sets I^s, provided $Z^N(s) \neq \emptyset$. This reflects the fact that the order of the constraints in CSP(N) defined by $\zeta := (\zeta^1, \ldots, \zeta^N) \in Z^N(s)$ does not matter. Furthermore the probability does not depend on the details of the distribution function or the constraint functions, but only N and s because there are $\binom{N}{s}$ index sets on $Z^N(s)$.s

Lemma 13.17 (Partitions $Z^N(I^s)$ of $Z^N(s)$)**.** Fix any $N \geq n$ and suppose assumption C13.3 holds. For any $0 \leq s \leq n$, if $Z^N(s) \neq \emptyset$ then

$$\mathbb{P}^N\left(Z^N(I^s)\middle|Z^N(s)\right) = \binom{N}{s}^{-1}, \qquad \forall I^s \text{ with } |I^s| = s \tag{13.91}$$

where $Z^N(I^s)$ and $Z^N(s)$ are defined in (13.89).

Proof The problem may not be uniformly supported, but we will condition on $Z^N(s)$, i.e., consider only $\zeta := (\zeta^1, \ldots, \zeta^N) \in Z^N$ that contains s support constraints. The lemma holds for $s = 0$ by definition in (13.89c). Fix an arbitrary $1 \leq s \leq n$ with $Z^N(s) \neq \emptyset$. To avoid triviality we assume $N > s$.

Consider the set $[s] := \{1, 2, \ldots, s\}$ and

$$Z^N([s]) := \left\{\zeta \in Z^N(s) : \left(X_{\zeta^i}, i \in [s]\right) \text{ are the } s \text{ support constraints in } \zeta\right\}$$

For any $I^s \subseteq \{1, \ldots, N\}$ with s indices and

$$Z^N(I^s) := \left\{\zeta \in Z^N(s) : \left(X_{\zeta^i}, i \in I^s\right) \text{ are the } s \text{ support constraints in } \zeta\right\}$$

we will establish a one-one correspondence between $Z^N(I^s)$ and $Z^N([s])$. Since ζ^i are iid this implies that

$$\mathbb{P}^N\left(Z^N(I^s)\middle|Z^N(s)\right) = \mathbb{P}^N\left(Z^N([s])\middle|Z^N(s)\right), \qquad \forall I^s \text{ with } |I^s| = s$$

The lemma then follows since there are $\binom{N}{s}$ index sets I^s with s support constraints. Order the indices in I^s as $i_1 < i_2 < \cdots < i_s$. Let $\alpha := (1, 2, \ldots, N)$ and let $P \in \{0,1\}^{N \times N}$ be any permutation matrix such that $[P\alpha]_{i_k} = k$, i.e., P maps $1, \ldots, s$ to $i_1, \ldots, i_s$ respectively and the complement of $[s]$ to the complement of I^s. We also write this mapping defined by P as $\pi(1) = i_1, \ldots, \pi(s) = i_s, \ldots, \pi(N)$ and the inverse mapping defined by P^{-1} as $\pi^{-1}(1), \ldots, \pi^{-1}(N)$. Then given any $\zeta := (\zeta^1, \ldots, \zeta^N) \in Z^N([s])$, $\left(\zeta^{\pi^{-1}(1)}, \ldots, \zeta^{\pi^{-1}(N)}\right) \in Z^N(I^s)$; given any $\zeta := (\zeta^1, \ldots, \zeta^N) \in Z^N(I^s)$, $\left(\zeta^{\pi(1)}, \ldots, \zeta^{\pi(N)}\right) \in Z^N([s])$. Therefore the permutation matrix P defines a bijection between $Z^N(I^s)$ and $Z^N([s])$ and completes the proof of the lemma. □

We start by proving Theorems 13.14 and 13.15 for the simple case where CSP(N) is uniformly supported with $s = 0$ support constraints, i.e., it has no support constraint for all $\zeta \in Z^N$. In this case the violation probability is 0 with probability 1.

Lemma 13.18 (No support constraint)**.** Suppose CSP(N) has no support constraint for any realization of $\zeta \in Z^N$.

1 CSP(k) has no support constraint for $k \geq N \geq n$ (with probability 1).
2 $V\left(x_N^*\right) = 0$ with probability 1. Hence $E^N\left(V\left(x_N^*\right)\right) = 0$ and $\mathbb{P}^N\left(V\left(x_N^*\right) > \epsilon\right) = 0$ for any $\epsilon > 0$.

Proof Consider CSP($N+1$) and suppose for the sake of contradiction that there are $\left(\zeta^1, \ldots, \zeta^{N+1}\right)$ with nonzero probability that have s support constraints, i.e., $Z^{N+1}(s) \neq \emptyset$ for some $1 \leq s \leq n < N+1$ (this is weaker than CSP($N+1$) being uniformly supported with s support constraints). Then every realization $(\zeta^1, \ldots, \zeta^{N+1}) \in Z^{N+1}(s) \subseteq Z^{N+1}$ has exactly s support constraints and $N+1-s$ non-support constraints. Hence CSP(N) with one of the non-support constraints removed will still have the same s constraints as support constraints. Since the samples ζ^i are iid, this contradicts that CSP(N) has no support constraint. Hence CSP($N+1$) has no support constraint and part 1 is proved by induction.

Part 2 then follows from Lemma 13.16. □

Proof of Theorem 13.14.

We next bound the expectation $E^N\left(V(x_N^*)\right)$ of the violation probability when CSP(N) may not be uniformly supported or is uniformly supported with $s \geq 1$ support constraints.

Proof of Theorem 13.14 We have from (13.90)

$$Z^{N+1} = \bigcup_{s'=0}^{n} Z^{N+1}(s') = \bigcup_{s'=0}^{n} Z^{N+1}(s') \bigcup_{I^{s'}} Z^{N+1}(I^{s'})$$

where $I^{s'} \subseteq \{1, \ldots, N+1\}$ specifies $|I^{s'}| = s'$ support constraints for CSP($N+1$). Hence, conditioning on $\zeta \in Z^{N+1}$ having s' support constraints, we have

$$\begin{aligned} E^N\left(V(x_N^*)\right) &= \mathbb{P}^{N+1}\left(x_N^* \notin X_{\zeta^{N+1}}\right) \\ &= \mathbb{P}^{N+1}\left(X_{\zeta^{N+1}} \text{ is support constraint for CSP}(N+1)\right) \\ &= \sum_{s'=0}^{n} \mathbb{P}^{N+1}\left(Z^{N+1}(s')\right) \sum_{I^{s'}: N+1 \in I^{s'}} \mathbb{P}^{N+1}\left(Z^{N+1}(I^{s'}) \middle| Z^{N+1}(s')\right) \end{aligned} \quad (13.92)$$

where the first equality follows from (13.86), the second equality follows from Lemma 13.16, and the last equality follows because $Z^{N+1}(I^{s'})$ are disjoint across $I^{s'}$.

Suppose CSP$(N+1)$ is uniformly bounded with s support constraints. The case of $s=0$ (i.e., CSP$(N+1)$ has no support constraint) is proved in Lemma 13.18. Hence fix any $1 \le s \le n$. Then $Z^{N+1} = Z^{N+1}(s) = \bigcup_{I^s} Z^{N+1}(I^s)$ and $Z^{N+1}(s') = \emptyset$ (with probability 1) for $s' \neq s$. Applying Lemma 13.17 to (13.92) we have

$$\begin{aligned} E^N\left(V(x_N^*)\right) &= \sum_{I^s : N+1 \in I^s} \mathbb{P}^{N+1}\left(Z^{N+1}(I^s) \,\middle|\, Z^{N+1}(s)\right) \\ &= \binom{N}{s-1} \cdot \binom{N+1}{s}^{-1} = \frac{s}{N+1} \end{aligned} \tag{13.93}$$

where the second equality follows because, of all the $\binom{N+1}{s}$ index sets I^s, $\binom{N}{s-1}$ of them contain $N+1$.

For the general case where CSP$(N+1)$ may not be uniformly supported for any integer s, application of Lemma 13.17 and (13.93) to (13.92) gives

$$\begin{aligned} E^N\left(V(x_N^*)\right) &= \sum_{s=1}^{n} \mathbb{P}^{N+1}\left(Z^{N+1}(s)\right) \sum_{I^s : N+1 \in I^s} \mathbb{P}^{N+1}\left(Z^{N+1}(I^s) \,\middle|\, Z^{N+1}(s)\right) \\ &= \sum_{s=1}^{n} \frac{s}{N+1}\, \mathbb{P}^{N+1}\left(Z^{N+1}(s)\right) \\ &= \frac{1}{N+1}\, E^{N+1}\left(\text{number of support constraints for CSP}(N+1)\right) \\ &\le \frac{s^{\max}}{N+1} \end{aligned} \tag{13.94}$$

where $s^{\max}$ is an upper bound on the number of support constraints for CSP$(N+1)$. Theorem 13.14 follows since $s^{\max} \le n$ by Lemma 13.13. □

13.3.3 Proof: bound on $\mathbb{P}^N(V(x_N^*) > \epsilon)$ for uniformly supported problem

We first prove Theorem 13.15 when CSP(N) is uniformly supported and then extends the argument to the general case.

Uniformly supported case.

Suppose CSP(N) is uniformly supported with s support constraints. Theorem 13.15 follows from Lemma 13.18 if $s=0$. Hence suppose $1 \le s \le n$ and assume $N > s$ to avoid triviality. From (13.90) we have

$$Z^N = Z^N(s) = \bigcup_{I^s} Z^N(I^s)$$

where $Z^N(I^s)$ contains vectors $\zeta \in Z^N$ such that $(\zeta^i, i \in I^s)$ are s support constraints for CSP(N). Since the sets I^s partition $Z^N(s)$ we can intersect the event $\left(V(x_N^*) > \epsilon\right)$

with the events $Z^N(I^s)$ to get:

$$\mathbb{P}^N\left(V(x_N^*) > \epsilon\right) \;=\; \sum_{I^s} \mathbb{P}^N\left(\zeta : V(x_N^*) > \epsilon,\; \zeta \in Z^N(I^s)\right) \tag{13.95}$$

We will derive each summand $\mathbb{P}^N\left(\zeta : V(x_N^*) > \epsilon,\; \zeta \in Z^N(I^s)\right)$.

Fix any $I^s \subseteq \{1,\ldots,N\}$ with s indices. Each (realization of) $\zeta := (\zeta^1,\ldots,\zeta^N)$ defines a CSP(N) that has exactly s support constraints (they may not be in I^s unless $\zeta \in Z^N(I^s)$). Let $\zeta(I^s) := (\zeta^i, i \in I^s)$ denote the subset of constraints in ζ indexed by I^s. We will use the s constraints in $\zeta(I^s)$ to also define a scenario program CSP(s) and denote its (unique) optimal solution by x_s^*.

Even though CSP(N) is uniformly supported with s support constraints, a generic CSP(s) defined by arbitrary s iid samples may not be uniformly supported. Let $Z^s(s)$ denote the set of $\tilde{\zeta} := (\tilde{\zeta}^1,\ldots,\tilde{\zeta}^s) \in Z^s$ that are support constraints for CSP(s), equipped with the conditional distribution $\mathbb{P}(\cdot|Z^s(s))$. To emphasize, we will write the CSP(s) defined by a $\tilde{\zeta} \in Z^s(s)$ as CSP($\tilde{s}$). Denote by $\tilde{x}_s^*$ the unique optimal solution of CSP($\tilde{s}$). The violation probability of $\tilde{x}_s^*$ is the conditional probability (from (13.85b))

$$V\left(\tilde{x}_s^*\right) \;:=\; \mathbb{P}\left(\tilde{\zeta}^{s+1} \in Z : \tilde{x}_s^* \notin X_{\tilde{\zeta}^{s+1}} \middle| \tilde{\zeta} := (\tilde{\zeta}^1,\ldots,\tilde{\zeta}^s) \in Z^s(s)\right) \tag{13.96a}$$

conditioned on a $\tilde{\zeta} \in Z^s(s)$. This is a random variable with the product measure $\mathbb{P}^s(\cdot|Z^s(s))$. Let

$$F^s(v|Z^s(s)) \;:=\; \mathbb{P}^s\left(V\left(\tilde{x}_s^*\right) \le v|Z^s(s)\right), \qquad v \in [0,1] \tag{13.96b}$$

denote the distribution function of $V(\tilde{x}_s^*)$ condition on $Z^s(s)$ (not a $\tilde{\zeta} \in Z^s(s)$).[7]

Remark 13.7. We emphasize the difference between the two independent random solutions $\tilde{x}_s^*$ and x_s^*. CSP($\tilde{s}$) is defined by *s support constraints* $\tilde{\zeta} \in Z^s(s)$ for CSP($\tilde{s}$) and $\tilde{x}_s^*$ is its unique optimal solution. Given any $\zeta \in Z^N$ and an index set I with s indices, let $\zeta(I) := (\zeta^i, i \in I)$ denote the subset of constraints indexed by I. As we will see in Lemma 13.19, we are not interested in generic CSP(s) defined by arbitrary s iid samples, but the CSP(s) defined by $\zeta(I)$ obtained from N samples ζ and a given index set I. To emphasize this dependence, we will write CSP($\zeta(I)$) and x_I^* instead of CSP(s) and x_s^*. Lemma 13.19 shows that $\zeta \in Z^N(I^s)$, i.e., the support constraints in ζ are indexed by I^s, if and only if the optimal solution $x_{I^s}^*$ of CSP($\zeta(I^s)$) also satisfies constraints $(\zeta^i, i \notin I^s)$. These subtle details become important when we generalize the proof for the uniformly supported case to the general case in Chapter 13.3.4. □

We will prove Theorem 13.15 in three steps:

1 Show that $x_N^* = x_{I^s}^*$ where I^s is the set of support constraints in $\zeta \in Z$. Relate the violation probability of $x_{I^s}^*$ to that of $\tilde{x}_s^*$ and hence to F^s (Lemma 13.19).
2 Derive the distribution function $F^s(v|Z^s(s)) = v^s$ for $v \in [0,1]$ (Lemma 13.20).

[7] The distribution F^s is only for $\tilde{x}_s^*$ where s is the number of support constraints for CSP(N). In particular, violation probability of x_N^* is not given by F^N.

3 Apply Lemmas 13.19 and 13.20 to (13.95) to derive $\mathbb{P}^N\left(V(x_N^*) > \epsilon\right)$.

Step 1 makes crucial use of the fact that CSP(N) is uniformly supported and needs modification in the general case. Steps 2 and 3 extend to the general case directly.

Let

$$Y^N(I^s) := \left\{\zeta \in Z^N : x_{I^s}^* \text{ of } \mathrm{CSP}(\zeta(I^s)) \in X_{\zeta^i}, i \notin I^s\right\} \tag{13.97}$$

Informally, $\mathbb{P}\left(Y^N(I^s)\right)$ is the violation probability of $x_{I^s}^*$ where I^s is any index set with s indices, not necessarily a set of support constraints for CSP(N). We will relate it to the violation probability of $\tilde{x}_s^*$ and hence to F^s. Recall that $Z^N(I^s) := \left\{\zeta \in Z^N : (\zeta^i, i \in I^s) \text{ is the set of support constraints for } \mathrm{CSP}(N)\right\}$.

Lemma 13.19. Fix any $N \geq n$ and suppose assumption C13.3 holds. If CSP(N) is uniformly supported with s support constraints then for any $I^s \subseteq \{1, \ldots, N\}$

1 $x_s^* = x_{s+1}^* = \cdots = x_N^*$ for all $\zeta \in Z^N$ where x_k^* is the optimal solution of the resulting CSP(k) after $N-k$ *non-support* constraints are removed.
2 $Z^N(I^s) = Y^N(I^s)$ with probability 1 under $\mathbb{P}^N$.
3 We have

$$\mathbb{P}^N\left(Z^N(I^s)\right) = \mathbb{P}^N\left(Y^N(I^s)\right) = \int_0^1 (1-v)^{N-s} dF^s(v|Z^s(s)) \tag{13.98}$$

Proof Suppose $\zeta \in Z^N(I^s)$. Then I^s are support constraints and its complement $I^{sc} := \{i \notin I^s\}$ are not support constraints for CSP(N). If we remove a constraint from I^{sc}, since it is not a support constraint, x_N^* remains the optimal solution for CSP($N-1$) with the remaining $N-1$ constraints. If $N-1 = s$ then $x_s^* = x_N^*$ since optimal solutions are unique (assumption C13.3), and hence $x_s^* \in X_{\zeta^i}, i \notin I^s$. If $N-1 > s$ then the s constraints in I^s remain support constraints for CSP($N-1$). Moreover the $N-s-1$ constraints in its complement I^{sc} are not support constraints for CSP($N-1$), i.e., no $\zeta^i, i \in I^{sc}$ can become a support constraint for CSP($N-1$) when CSP(N) is uniformly supported (Exercise 13.19). [8] Repeating this process and we conclude that x_N^* remains the optimal solution for each CSP(k) after $N-k$ non-support constraints are removed from I^{sc}. Since the optimal solutions are unique for each $\zeta \in Z^N$ by assumption C13.3, $x_s^* = x_{s+1}^* = \cdots = x_N^*$. In particular $x_s^* \in X_{\zeta^i}$, $i \notin I^s$. Hence $\zeta \in Y^N(I^s)$.

Conversely suppose $\zeta \in Y^N(I^s)$. Since I^{sc} specifies the s constraints for CSP(s), $x_s^* \in X_{\zeta^i}, i \in I^s$. Moreover $x_s^* \in X_{\zeta^i}, i \notin I^s$, since $\zeta \in Y^N(I^s)$. Therefore x_s^* is feasible, and hence optimal, for CSP(k) after $N-k$ constraints are removed from I^{sc}, $k = s, \ldots, N$. By uniqueness of optimal solutions, we have $x_s^* = x_{s+1}^* = \cdots = x_N^*$. If any constraint in I^{sc} is a support constraint for CSP(N), then removing it will change the optimal solution, i.e., $x_{N-1}^* \neq x_N^*$, a contradiction. Hence none of the constraints in I^{sc} can be support constraints for CSP(N). Therefore all constraints in I^{sc} must

[8] If CSP(N) is not uniformly supported then this is not necessarily the case because of latent support constraints; see Definition 13.4.

be support constraints for CSP(N) since $\zeta \in Z^N(s)$. This proves $\zeta \in Z^N(I^s)$, and $Z^N(I^s) = Y^N(I^s)$ a.s.

The argument above shows that $\zeta \in Z^N(I^s)$ implies that $x^*_{I^s} = x^*_s = x^*_{s+1} = \cdots = x^*_N$ where x^*_k is the optimal solution of the resulting CSP(k) after $N-k$ *non-support* constraints are removed. (Recall that $x^*_{I^s}$ denotes the optimal solution of CSP($\zeta(I^s)$) defined by constraints in I^s and $x^*_{I^s} = x^*_s$ because $\zeta \in Z^N(I^s)$). Since this holds for all I^s and $Z^N = \cup_{I^s} Z^N(I^s)$ for uniformly supported CSP(N), $x^*_{I^s} = x^*_s = x^*_{s+1} = \cdots = x^*_N$ holds for all $\zeta \in Z^N$.

Finally, $Z^N(I^s) = Y^N(s)$ implies that any $\zeta \in Z^N(I^s) = Y^N(s)$ has its support constraints indexed by I^s and $x^*_{I^s} = x^*_s$ satisfies constraints $(\zeta^i, i \notin I^s)$. Moreover the event $(\zeta^i, i \in I^s)$ that defines $x^*_{I^s}$ and the event $(X_{\zeta^i}, i \notin I^s)$ in the definition of $Y^N(I^s)$ in (13.97) are independent because $\mathbb{P}\left(Y^N(I^s) | Z^N(s)\right) = \mathbb{P}\left(Y^N(I^s)\right)$ when $Z = Z^N(s)$, implying that

$$\mathbb{P}^N\left(Y^N(I^s)\right) \;=\; \mathbb{P}^N\left(\tilde{x}^*_s \in X_{\tilde{\zeta}^i}, i = s+1, \ldots, N \middle| Z^s(s)\right) \tag{13.99}$$

Here $x^*_{I^s} = \tilde{x}^*_s$ in probability since both are the unique optimal solutions of CSP(s) defined by two independent sets of s support constraints for CSP(s). [9]

Conditioned on a $\tilde{\zeta} \in Z^s(s)$, the probability that $\tilde{x}^*_s$ does not violate the constraints $(\tilde{\zeta}^i, i = s+1, \ldots, N)$ is $\left(1 - V\left(\tilde{x}^*_s\right)\right)^{N-s}$ since $\tilde{\zeta}^i$ are iid, i.e.,

$$\mathbb{P}^{N-s}\left(\tilde{x}^*_s \in X_{\tilde{\zeta}^i}, i = s+1, \ldots, N \middle| \tilde{\zeta} \in Z^s(s)\right) \;=\; \left(1 - V\left(\tilde{x}^*_s\right)\right)^{N-s}$$

where $V\left(\tilde{x}^*_s\right)$ is defined in (13.96). This is itself a random variable with probability measure $\mathbb{P}^s$ since $\tilde{x}^*_s$ depends on $\tilde{\zeta} \in Z^s(s)$. The probability that $\tilde{x}^*_s$ does not violate these constraints, conditioned on $Z^s(s)$ as opposed to a $\tilde{\zeta} \in Z^s(s)$, is

$$\begin{aligned}\mathbb{P}^N\left(\tilde{x}^*_s \in X_{\tilde{\zeta}^i}, i = s+1, \ldots, N \middle| Z^s(s)\right) \;&=\; \int_{Z^s} \left(1 - V\left(\tilde{x}^*_s\right)\right)^{N-s} \mathbb{P}^s\left(d\tilde{\zeta}^1, \ldots, d\tilde{\zeta}^s \middle| Z^s(s)\right) \\ &= \int_0^1 (1-v)^{N-s} dF^s(v | Z^s(s)) \end{aligned} \tag{13.100}$$

where the second equality follows from (13.96). Combining this with (13.99) and $Z^N(I^s) = Y^N(s)$ proves part 3 of the lemma. □

Consider the scenario program CSP($\tilde{s}$) defined by s iid samples $\tilde{\zeta} := (\tilde{\zeta}^1, \ldots, \tilde{\zeta}^s)$ that are support constraints of CSP($\tilde{s}$). The distribution function F^s of the violation probability $V(\tilde{x}^*_s)$ is defined in (13.96).

Lemma 13.20. $F^s(v | Z^s(s)) = v^s$ over $v \in [0,1]$.

[9] If CSP(N) is not uniformly supported then $(\zeta^i, i \in I^s)$ and $(X_{\zeta^i}, i \notin I^s)$ are dependent and (13.99) becomes an inequality in (13.107).

Proof We have from Lemma 13.19

$$\mathbb{P}^N\left(Z^N(I^s)\right) = \int_0^1 (1-v)^{N-s} dF^s(v|Z^s(s))$$

Substituting $\mathbb{P}^N\left(Z^N(I^s)|Z^N(s)\right) = \binom{N}{s}^{-1}$ from Lemma 13.17 we have

$$\binom{N}{s}\int_0^1 (1-v)^{N-s} dF^s(v|Z^s(s)) = 1$$

This is an integral equation in F^s. We show that $F^s(v|Z^s(s)) = v^s$ is the unique solution by substituting it into the left-hand side and integrating by part:

$$\begin{aligned}\binom{N}{s}\int_0^1 (1-v)^{N-s} d(v^s) &= \binom{N}{s}\left((1-v)^{N-s}v^s\Big|_0^1 + (N-s)\int_0^1 (1-v)^{N-s-1}v^s dv\right)\\ &= \binom{N}{s}\frac{N-s}{s+1}\int_0^1 (1-v)^{N-s-1} d(v^{s+1})\\ &= \binom{N}{s}\frac{(N-s)\cdots 1}{(s+1)\cdots N}\int_0^1 d(v^N) = 1\end{aligned}$$

which is equal to the right-hand side. □

We now use Lemmas 13.19 and 13.20 to bound the tail probability of $V(x_N^*)$ when CSP(N) is uniformly supported with s support constraints.

Proof of Theorem 13.15: uniformly supported case Suppose CSP(N) is uniformly supported with $1 \le s \le n$ support constraints. (The case of $s = 0$ follows from Lemma 13.18). Assume $N > s$ to avoid triviality.

The summands on the right-hand side of (13.95) are:

$$\begin{aligned}\mathbb{P}^N\left(V(x_N^*) > \epsilon, Z^N(I^s)\right) &= \mathbb{P}^N\left(V(x_{I^s}^*) > \epsilon, Y^N(I^s)\right)\\ &= \mathbb{P}^N\left(V(\tilde{x}_s^*) > \epsilon, \tilde{x}_s^* \in X_{\tilde{\zeta}^i}, i = s+1,\ldots,N \middle| Z^s(s)\right)\\ &= \int_{\{V(\tilde{x}_s^*)>\epsilon\}} \left(1 - V\left(\tilde{x}_s^*\right)\right)^{N-s} \mathbb{P}^s\left(d\tilde{\zeta}^1,\ldots,d\tilde{\zeta}^s \middle| Z^s(s)\right)\\ &= \int_\epsilon^1 (1-v)^{N-s} dF^s(v)\end{aligned}$$

where the first equality follows because $x_N^* = x_s^* = x_{I^s}^*$ and $Z^N(I^s) = Y^N(I^s)$ from Lemma 13.19, the second equality follows from (13.99), and the last equality follows (13.96). Substituting this and Lemma 13.20 into (13.95) and integrating by part, we

have (since there are $\binom{N}{s}$ index sets I^s):

$$\begin{aligned}
\mathbb{P}^N\left(V(x_N^*) > \epsilon\right) &= \binom{N}{s}\int_\epsilon^1 (1-v)^{N-s} d(v^s) \\
&= -\binom{N}{s}(1-\epsilon)^{N-s}\epsilon^s + \binom{N}{s+1}\int_\epsilon^1 (1-v)^{N-s-1} d(v^{s+1}) \\
&\vdots \\
&= -\sum_{i=s}^{N-1}\binom{N}{i}(1-\epsilon)^{N-i}\epsilon^i + \binom{N}{N}\int_\epsilon^1 d(v^N) \\
&= -\sum_{i=s}^{N-1}\binom{N}{i}(1-\epsilon)^{N-i}\epsilon^i + \left(1-\epsilon^N\right) \\
&= \sum_{i=0}^{s-1}\binom{N}{i}\epsilon^i(1-\epsilon)^{N-i} \qquad (13.101)
\end{aligned}$$

This completes the proof of Theorem 13.15 when CSP(N) is uniformly supported with $1 \le s \le n$ support constraints. □

13.3.4 Proof: bound on $\mathbb{P}^N(V(x_N^*) > \epsilon)$ for general problem

To prove the general case where the problem is not uniformly supported we need to study the partition Z^N more carefully. As pointed out in the proof of Lemma 13.19 (footnotes [8] and [9]) there are two difficulties. The first is that, given a $\zeta \in Z^N(I^s)$ with s support constraints, when we remove a non-support constraint from the complement I^{sc} of I^s, a remaining constraint $\zeta^i \in I^{sc}$ may become a support constraint for CSP($N-1$) if CSP(N) is not uniformly supported, i.e., CSP($N-1$) may have more than s support constraints. Then x_s^* violates the constraint X_{ζ^i} and therefore the given $\zeta \notin Y^N(I^s)$, violating $Z^N(I^s) = Y^N(I^s)$ in Lemma 13.19. This difficulty is overcome by considering generalized support constraints defined in Definition 13.5. The second difficulty is that, if CSP(N) is not uniformly supported, then the events $(\zeta^i, i \in I^s)$ and $(X_{\zeta^i}, i \notin I^s)$ in the definition of $Y^N(I^s)$ in (13.97) may no longer be independent. This means that (13.99), which expresses the violation probability of $x_{I^s}^*$ in terms of that of $\tilde{x}_s^*$, may no longer hold. Instead the conditional version of (13.99) becomes an inequality in (13.107), leading to an upper bound in Theorem 13.15.

Generalized support constraint.

A constraint that is not a support constraint for CSP(N) but becomes a support constraint for some CSP($N-k$) when some of the k constraints are removed is called a latent support constraint for CSP(N). Recall that, given any $\zeta \in Z^N$ and any index

set $J \subseteq \{1,\ldots,N\}$, $\zeta(J) := (\zeta^i, i \in J)$ denotes the subset of constraints indexed by J, CSP($\zeta(J)$) the scenario program defined by these constraints, and $x^*_{J^t}$ its unique optimal solution.

Definition 13.4 (Latent support constraint)**.** Fix a $\zeta \in Z^N$ and let $I^s = I^s(\zeta)$ denote its (unique) set of s support constraints. A set $L^\ell \subseteq \{i, i \notin I^s\}$ with ℓ indices is called a *set of latent support constraints* with respect to ζ if $\zeta(I^s \cup L^\ell)$ is the set of support constraints for CSP($I^s \cup L^\ell$). Each ζ^i (or X_{ζ^i}), $i \in L^\ell$, is called a *latent support constraint* for CSP(N).

A set L^ℓ with ℓ indices is a *maximal set of latent support constraints* with respect to ζ if L^ℓ is a set of latent support constraints with the largest number of indices. Instead of partitioning Z^N according to I^s of support constraints in the uniformly supported case, we will partition Z^N according to $I^s \cup L^\ell$ where, for each $\zeta \in Z^N$, I^s is the (unique) set of support constraints and L^ℓ is a maximal set of latent support constraints. A ζ however can have multiple maximal sets L^ℓ of latent support constraints. Even though they all have the same number ℓ of indices, the sets $Z^N(I^s \cup L^\ell)$, which is the set of all ζ whose support constraints are in I^s and latent support constraints in L^ℓ, do not form a partition of Z^N because maximal sets L^ℓ are non-unique. For instance, consider $\zeta \in Z^5$ defined by $\zeta^1 = a$, $\zeta^2 = \zeta^3 = b$, and $\zeta^4 = \zeta^5 = c$, where $I^1 := \{1\}$ is (the index of) the single support constraint. Suppose $L_{ij} := \{i, j\}$ for $i = 2,3$ and $j = 4,5$ are maximal sets of latent support constraint with 2 indices. Then ζ is in all four sets $Z^5(I^1 \cup L_{ij})$ for $i = 2,3$ and $j = 4,5$ (see Exercise 13.20 for more details). This can be resolved by choosing a unique representative among all maximal sets of latent support constraints for each ζ, e.g., according to the lexicographical order of these maximal sets. For the example above this representative is L_{24}.

Definition 13.5 (Generalized support constraint)**.** Fix a $\zeta \in Z^N$ and let $I^s = I^s(\zeta)$ denote its (unique) set of s support constraints.

1 A set $L^\ell \subseteq \{i : i \notin I^s\}$ with ℓ indices is called the (unique) *maximum set of latent support constraints* with respect to ζ if L^ℓ is a set of latent support constraints with the largest number of indices and it is the smallest of such sets in the lexicographical order. In this case ℓ is called the *maximum number* of latent support constraints with respect to ζ.
2 Let $J^t \subseteq \{1,\ldots,N\}$ be an index set with $s \le t \le n$ indices. We call $\zeta(J^t)$ or simply J^t *the set of generalized support constraints* for CSP(N) with respect to ζ if
   (a) $J^t = I^s \cup L^{t-s}$ where $L^{t-s} := L^{t-s}(\zeta)$ is the unique maximum set of $t-s$ latent support constraints with respect to ζ.
   (b) $\zeta(J^t)$ is the set of support constraints for CSP($\zeta(J^t)$).
3 We say $\zeta \in Z^N$ *has t generalized support constraints* if there exists an index set J^t with t indices such that $\zeta(J^t)$ is the set of generalized support constraints for CSP(N).

If CSP(N) is uniformly supported with $1 \le t \le n$ support constraints, then there

is no latent support constraints, i.e., $L^0 = \emptyset$ for all $\zeta \in Z^N = Z^N(t)$ (Exercise 13.19). For a general problem CSP(N) that may not be uniformly supported, any $\zeta \in Z^N(n)$ has no latent support constraint by Lemma 13.13. Exercise 13.19 proves some other properties of generalized support constraints.

Every $\zeta \in Z^N$ has a unique set I^s of support constraints and a unique maximum set L^ℓ of latent support constraints for CSP(N), and hence belongs to exactly one $Z^N(J^t)$ with $J^t := I^s \cup L^\ell$. This means that the sets $Z^N(J^t)$ form a partition of Z^N, the same way the set $Z^N(I^s)$ partition Z^N for the proof of Theorem 13.14 (see Figure 13.4). For $t = 0, 1, \ldots, n$, let $J^t \subseteq \{1, \ldots, N\}$ be an index set with t indices. We partition Z^N according to the number and identity of generalized support constraints:

$$Z^N(J^t) := \left\{\zeta \in Z^N : \zeta(J^t) \text{ are all the gen. supp. const. for CSP}(N)\right\} \tag{13.102a}$$

$$Z^N(t) := \bigcup_{J^t} Z^N(J^t) \tag{13.102b}$$

where the union ranges over all index sets J^t of t generalized support constraints, with $J^0 := \emptyset$ by definition. Then

$$Z^N = \bigcup_{t=0}^{n} Z^N(t) = \bigcup_{t=0}^{n} \bigcup_{J^t} Z^N(J^t) \tag{13.102c}$$

See Figure 13.4(b).

Therefore, as in (13.95) for the uniformly supported case, we can intersect the event $\left(V(x_N^*) > \epsilon\right)$ with the disjoint sets $Z^N(J^t)$:

$$\mathbb{P}^N\left(V(x_N^*) > \epsilon\right) = \sum_{t=0}^{n} \mathbb{P}^N(Z^N(t)) \sum_{J^t} \mathbb{P}^N\left(V(x_N^*) > \epsilon, Z^N(J^t) \middle| Z^N(t)\right) \tag{13.103}$$

These concepts are illustrated in Example 13.9 and Exercise 13.20.

Example 13.9 (Generalized support constraints and $\mathbb{P}(V(x_N^*) > \epsilon)$)**.** We are given two points $a, b \in \mathbb{R}^2$ on a plane. The random variable ζ is equal to a or b with nonzero probabilities p_a or $p_b := 1 - p_a$ respectively. Given N iid samples $(\zeta^1, \ldots, \zeta^N)$, $N \geq 4$, CSP(N) determines the smallest circle, specified by $x := (x_1, x_2, x_3) \in \mathbb{R}^3$, going through all N points $(\zeta^1, \ldots, \zeta^N)$.

1 Partition Z^N according to (13.102).
2 Derive $\mathbb{P}(V(x_N^*) > \epsilon)$ assuming $0 < \epsilon \leq \min\{p_a, p_b\}$. What if $\epsilon > \max\{p_a, p_b\}$?

Solution. The optimal circle is either C_a centered at a with zero radius (when $\zeta^i = a$ for all i), or C_b centered at b with zero radius (when $\zeta^j = b$ for all j), or C_{ab} with a and b on its diameter (when ζ^i takes both values a and b). Since $N \geq 4$, $Z^N(s) \neq \emptyset$ for only $s = 0$ or 1. In particular the maximum number $s^{\max} = 1$ of support constraints is less than $n = 3$. The partitioning of Z^N in (13.102) is summarized in Table 13.2. The

s	event	J^t	I^s	L^{t-s}	$\mathbb{P}^N$ (event)	x_N^*
0	$\zeta^i = a$ for all i	$\{i_{\min}\}$	$\emptyset$	$\{i\}$	p_a^N	C_a
	$\zeta^j = b$ for all j	$\{j_{\min}\}$	$\emptyset$	$\{j\}$	p_b^N	C_b
	$\zeta^i = a$ for $k \geq 2$ is, $\zeta^j = b$ for $N-k$ js	$\{i_{\min}, j_{\min}\}$	$\emptyset$	$\{i,j\}$	$p_a^k p_b^{N-k}$	C_{ab}
1	$\zeta^i = a$, $\zeta^j = b$ for all $j \neq i$	$\{i, j_{\min}\}$	$\{i\}$	$\{j\}$	$p_a p_b^{N-1}$	C_{ab}
	$\zeta^j = b$, $\zeta^i = a$ for all $i \neq j$	$\{j, i_{\min}\}$	$\{j\}$	$\{i\}$	$p_a^{N-1} p_b$	C_{ab}

Table 13.2 Example 13.9. The L^{t-s} column includes all maximal sets of latent support constraints.

sets $Z^N(1)$ and $Z^N(2)$ consist of all $\zeta \in Z^N$ with 1 and 2 respectively generalized support constraints. We have

$$\begin{aligned}\mathbb{P}^N\left(Z^N(1)\right) &= p_a^N + p_b^N \\ \mathbb{P}^N\left(Z^N(2)\right) &= \sum_{k=2}^{N-2}\binom{N}{k} p_a^k p_b^{N-k} + N p_a p_b^{N-1} + N p_a^{N-1} p_b\end{aligned}$$

and hence

$$\mathbb{P}\left(Z^N\right) = \mathbb{P}^N\left(Z^N(1)\right) + \mathbb{P}^N\left(Z^N(2)\right) = (p_a + p_b)^N = 1$$

as expected. Clearly $\mathbb{P}(J^1 \,|Z^N(1)) = 1/N$ and $\mathbb{P}(J^2 \,|Z^N(2)) = 2/(N(N-1))$.

For part 2, the violation probability is (from (13.103))

$$\mathbb{P}(V(x_N^*) > \epsilon) = \sum_{t=1}^{2}\sum_{J^t} \mathbb{P}^N\left(\zeta : V(x_N^*) > \epsilon, \zeta \in Z^N(J^t)\right) \tag{13.104}$$

Recall $V(x_N^*) := \mathbb{P}\left(x_N^* \notin X_{\zeta^{N+1}} \,|\zeta \in Z^N\right)$. Given a $\zeta \in Z^N(1)$ with one generalized support constraint, either $(\zeta^i = a\,\forall i)$ or $(\zeta^j = b\,\forall j)$. The former event happens with probability p_a^N, has the optimal solution $x_N^* = C_a$, and x_N^* satisfies constraint $X_{\zeta^{N+1}}$ if $\zeta^{N+1} = a$ and violates it if $\zeta^{N+1} = b$; similarly for the latter event. Therefore, since $\epsilon \leq \min\{p_a, p_b\}$,

$$\begin{aligned}P^N\left(\zeta : V(x_N^*) > \epsilon, \zeta^i = a\,\forall i\right) &= \mathbb{P}^N\left(\zeta^i = a\,\forall i\right)\mathbb{P}\left(\zeta^{N+1} = b\right) = p_a^N p_b \\ P^N\left(\zeta : V(x_N^*) > \epsilon, \zeta^j = b\,\forall j\right) &= \mathbb{P}^N\left(\zeta^j = b\,\forall j\right)\mathbb{P}\left(\zeta^{N+1} = b\right) = p_b^N p_a\end{aligned}$$

and $P^N\left(\zeta : V(x_N^*) > \epsilon, \zeta \in Z^N(J^2)\right) = 0$ since $x_N^* = C_{ab}$. Substituting into (13.104) we have $\mathbb{P}(V(x_N^*) > \epsilon) = p_a^N p_b + p_b^N p_a$. The upper bound on the number of generalized support constraints for this example is $t^{\max} = 2 < n$. Hence, in view of Remark 13.5, the bound in Theorem 13.15 is

$$\sum_{i=0}^{t^{\max}-1}\binom{N}{i}\epsilon^i(1-\epsilon)^{N-i} = (1-\epsilon)^N + N\epsilon(1-\epsilon)^{N-1} \overset{?}{\geq} p_a^N p_b + p_b^N p_a$$

since $1-\epsilon \geq \{p_a, p_b\}$ and $N\epsilon \geq p_a p_b$ (under appropriate conditions?! Something wrong?) If $\epsilon > \max\{p_a, p_b\}$ then $\mathbb{P}(V(x_N^*) > \epsilon) = 0$. □

The proof for the general case parallels that for the uniformly supported case, with $Z^N(J^t)$ and $Z^N(t)$ here playing the roles of $Z^N(I^s)$ and $Z^N(s)$ there. The main difference is Lemma 13.21 that extends Lemma 13.19 to the general case. Fix any $J^t \subseteq \{1,\ldots,N\}$ with t elements. Each (realization of) $\zeta := (\zeta^1,\ldots,\zeta^N) \in Z^N(t)$ defines a CSP(N) and has exactly t generalized support constraints (they may not be in J^t unless $\zeta \in Z^N(J^t)$). Moreover we will use the t constraints $\zeta(J^t) := (\zeta^i, i \in J^t)$ to also define CSP($\zeta(J^t)$), and denote its (random) optimal solution by $x^*_{J^t}$.

Consider again the scenario program CSP($\tilde{t}$) defined by separate t iid samples $\tilde{\zeta} := (\tilde{\zeta}^1,\ldots,\tilde{\zeta}^t) \in Z^t(t)$ that are support constraints for CSP($\tilde{t}$). The violation probability $V(\tilde{x}_t^*)$ of its unique optimal solution $\tilde{x}_t^*$ is defined in (13.96), reproduced here:

$$V(\tilde{x}_t^*) := \mathbb{P}\left(\tilde{\zeta}^{t+1} \in Z : \tilde{x}_t^* \notin X_{\tilde{\zeta}^{t+1}} \middle| \tilde{\zeta} := (\tilde{\zeta}^1,\ldots,\tilde{\zeta}^t) \in Z^t(t)\right) \tag{13.105a}$$

conditioned on a $\tilde{\zeta} \in Z^t(t)$ with distribution function

$$F^t(v|Z^t(t)) := \mathbb{P}^t\left(V(\tilde{x}_t^*) \leq v | Z^t(t)\right), \qquad v \in [0,1] \tag{13.105b}$$

condition on $Z^t(t)$ (as opposed to a $\tilde{\zeta} \in Z^t(t)$). The proof of Lemma 13.20 in terms of a certain scenario program CSP(N) with t support constraints applies here and shows that $F^t(v) = v^t$ over $[0,1]$.

We next extend Lemma 13.19 to the general case where CSP(N) may not be uniformly supported. Let $Y^N(J^t)$ and $Z^N(J^t)$ be the conditional version of these sets in the uniformly supported case (cf. (13.97)):

$$Y^N(J^t) := \{\zeta \in Z^N(t) : x^*_{J^t} \text{ of CSP}(\zeta(J^t)) \in X_{\zeta^i}, i \notin J^t\}$$
$$Z^N(J^t) := \{\zeta \in Z^N(t) : \zeta(J^t) \text{ is the set of gen. supp. const. for CSP}(N)\}$$

In contrast to the uniformly supported case, these two sets are equal in probability, not with probability 1, when conditioned on $Z^N(t)$ because maximal sets of latent support constraints are not unique.

Lemma 13.21. Fix any $N \geq n$ and suppose assumption C13.3 holds. Then for any $J^t \subseteq \{1,\ldots,N\}$ with $1 \leq t \leq n$

1 $x_t^* = x_{t+1}^* = \cdots = x_N^*$ for all $\zeta \in Z^N(t)$ where x_k^* is the optimal solution of the resulting CSP(k) after $N-k$ *non-support* constraints are removed.
2 $\mathbb{P}^N\left(Z^N(J^t)\middle|Z^N(t)\right) = \mathbb{P}^N\left(Y^N(J^t)\middle|Z^N(t)\right)$.
3 We have

$$\mathbb{P}^N\left(Z^N(J^t)\middle|Z^N(t)\right) = \mathbb{P}^N\left(Y^N(J^t)\middle|Z^N(t)\right) \leq \int_0^1 (1-v)^{N-s} dF^t(v|Z^t(t)) \tag{13.106}$$

Proof Suppose $\zeta \in Z^N(J^t)$. Then J^t is the unique set of generalized support constraints for CSP(N) with a unique decomposition $J^t = I^s \cup L^{t-s}$ of support constraints and latent support constraints. The set $J^{tc} := \{i \notin J^t\}$ may contain other latent support constraints for CSP(N) but no support constraints. If we remove a constraint from J^{tc}, the resulting optimal solution $x^*_{N-1} = x^*_N$. If $N-1=t$ then $x^*_{J^t} = x^*_{N-1} = x^*_N$ since optimal solutions are unique and hence $x^*_{J^t}$ satisfies $X_{\zeta^i}, i \notin J^t$. If $N-1 > t$ then J^t remains the set of generalized support constraints for CSP($N-1$), i.e., J^t remains the set of support constraints for CSP($\zeta(J^t)$), and no constraint $\zeta^i, i \notin I^s$, becomes a support constraint for CSP($N-1$); see Exercise 13.19. Moreover Exercise 13.19 shows that $x^*_{J^t} = x^*_t = x^*_{t+1} = \cdots = x^*_N$.[10] In particular $x^*_{J^t} \in X_{\zeta^i}, i \in J^{tc}$. Therefore $\zeta \in Y^N(J^t)$ and hence $\mathbb{P}^N(Z^N(J^t)|Z^N(t)) \le \mathbb{P}^N(Y^N(J^t)|Z^N(t))$.

Conversely suppose $\zeta \in Y^N(J^t)$. We will show that, in probability, J^t is the set of generalized support constraint for CSP(N) (with respect to ζ), by showing (i) $J^t \supseteq I^s$ where $I^s := I^s(\zeta)$ is the unique set of support constraints; and (ii) J^t is the set of support constraints for CSP($\zeta(J^t)$) in probability. Clearly $x^*_{J^t} \in X_{\zeta^i}, i \in J^t$, since $x^*_{J^t}$ is optimal for CSP($\zeta(J^t)$). Moreover $x^*_{J^t} \in X_{\zeta^i}, i \notin J^t$ since $\zeta \in Y^N(J^t)$. Therefore $x^*_{J^t}$ is feasible, and hence optimal, for the resulting scenario programs CSP($N-k$) after k constraints in J^{tc} are removed, $k = 1, \ldots, N-t$. By uniqueness of optimal solutions, we must have $x^*_{J^t} = x^*_t = x^*_{t+1} = \cdots = x^*_N$. This implies that none of the constraints in J^{tc} can be support constraints for CSP($N-k$), and in particular $J^t \supseteq I^s$.

If J^t is a set of support constraints for CSP(J^t) with respect to ζ but $J^t \setminus I^s$ is not the (unique) maximum set L^{t-s} of latent support constraints, then we can exchange $J^t \setminus I^s$ for L^{t-s} (relabeling constraints) so that J^t becomes the set of generalized support constraints for CSP(N). This amounts to replacing ζ by a different ζ' that has the same probability for which $\zeta'(J^t)$ is the set of generalized support constraints for CSP(N). Suppose for the sake of contradiction that J^t is not a set of support constraints for CSP(J^t). Write $J^t =: J_1 \cup J_2$ where J_1 is the set of support constraints and J_2 the set of non-support constraints for CSP(J^t). Removing constraints in J_2 yields the scenario program CSP($\zeta(J_1)$) whose optimal solution satisfies $x^*_{J_1} = x^*_{J^t}$. Since $\zeta \in Z^N(t)$, there is a set J_2' of $|J_2|$ constraints in J^{tc} such that if we add them back to J_1, then $J_1 \cup J_2'$ is a set of t support constraints for CSP($\zeta(J_1 \cup J_2')$). Since J_2' are support constraints, removing them from CSP($\zeta(J_1 \cup J_2')$) results in CSP($\zeta(J_1)$) with optimal solutions $x^*_{J_1} \neq x^*_{J^t}$. This is a contradiction since optimal solutions are unique. Hence J^t is equal, in probability, to the set of generalized support constraints for CSP(N). This shows $\mathbb{P}^N(Z^N(J^t)|Z^N(t)) \ge \mathbb{P}^N(Y^N(J^t)|Z^N(t))$, and completes the proof of $\mathbb{P}^N\left(Z^N(J^t)\middle|Z^N(t)\right) = \mathbb{P}^N\left(Y^N(J^t)\middle|Z^N(t)\right)$.

The argument above shows that $\zeta \in Z^N(J^t)$ implies $x^*_{J^t} = x^*_t = x^*_{t+1} = \cdots = x^*_N$ for any $Z^N(J^t)$. Hence $x^*_{J^t} = x^*_t = x^*_{t+1} = \cdots = x^*_N$ for any $\zeta \in Z^N(t)$.

[10] Recall that x^*_t is the optimal solution of the resulting CSP(t) after $N-t$ non-support constraints are removed, and $x^*_{J^t}$ is the optimal solution of CSP($\zeta(J^t)$) defined by constraints in J^t. They are equal for $\zeta \in Z^Y(J^t)$.

Finally for part 3, we first claim that

$$\begin{aligned}\mathbb{P}^N\left(Y^N(J^t)\,\middle|\,Z^N(t)\right) &= \mathbb{P}^N\left(x^*_{J^t}\in X_{\zeta^i}, i\notin J^t\,\middle|\,Z^N(t)\right)\\ &\le \mathbb{P}^N\left(\tilde{x}^*_t\in X_{\tilde{\zeta}^i}, i=t+1,\ldots,N\,\middle|\,Z^t(t)\right)\end{aligned}\tag{13.107}$$

in contrast to (13.99) for the uniformly supported case. The inequality follows for two reasons. First, given any $\zeta\in Y^N(J^t)$, the proof above shows that $\zeta(J^t)$ is a set of support constraints for $\mathrm{CSP}(\zeta(J^t))$ (even though the set $J^t\setminus I^s$ of maximal latent support constraints in J^t may not be the maximum set in the lexicographical order). Hence the two independent scenario programs $\mathrm{CSP}(\zeta(J^t))$ and $\mathrm{CSP}(\tilde{t})$ are both defined by t support constraints. If we treat their optimal solutions $x^*_{J^t}$ and $\tilde{x}^*_t$ respectively as maps from $Z^t(t)\to\mathbb{R}^n$, then these two maps are identical since optimal solutions are unique. Second, the inequality in (13.107) is equivalent to:

$$\frac{\mathbb{P}^N\left(x^*(\zeta(J^t))\in X_{\zeta^i}, i\notin J^t, \zeta\in Y^N(t)\right)}{\mathbb{P}^N\left(Z^N(t)\right)} \le \frac{\mathbb{P}^N\left(\tilde{x}^*(\tilde{\zeta})\in X_{\tilde{\zeta}^i}, i=t+1,\ldots,N, \tilde{\zeta}\in Z^t(t)\right)}{\mathbb{P}^t(Z^t(t))}$$

where we have written $x^*(\zeta(J^t)):=x^*_{J^t}$ and $\tilde{x}^*(\tilde{\zeta}):=\tilde{x}^*_t$ to emphasize their dependence on t support constraints $\zeta(J^t)$ and $\tilde{\zeta}$. This means on the numerators that

$$\begin{aligned}&\mathbb{P}^N\left(\zeta\in Y^N(t): x^*(\zeta(J^t))\in X_{\zeta^i}, i\notin J^t,\ \zeta(J^t)\text{ supp. const.}\right)\\ \le\ &\mathbb{P}^N\left(\zeta\in Z^N: x^*(\zeta(J^t))\in X_{\zeta^i}, i\notin J^t,\ \zeta(J^t)\text{ supp. const.}\right)\\ =\ &\mathbb{P}^N\left(\tilde{\zeta}\in Z^t(t): \tilde{x}^*(\tilde{\zeta})\in X_{\tilde{\zeta}^i}, i=t+1,\ldots,N\right)\end{aligned}$$

where the inequality follows since $Y^N(t)\subseteq Z^N$ if $\mathrm{CSP}(N)$ is not uniformly supported. Using (13.102b) the denominators satisfy, letting $[t]:=\{\zeta^1,\ldots,\zeta^t\}$,

$$\mathbb{P}^N\left(Z^N(t)\right) = \mathbb{P}^N\left(Z^N([t])\right) + \sum_{J^t\neq[t]}\mathbb{P}^t\left(Z^N(J^t)\right) \ge \mathbb{P}^t\left(Z^t(t)\right)$$

This proves (13.107).

The rest of the proof of part 3 follows the same argument as that of Lemma 13.19. Conditioned on a $\tilde{\zeta}\in Z^t(t)$, we have

$$\mathbb{P}^{N-t}\left(\tilde{x}^*_t\in X_{\tilde{\zeta}^i}, i=t+1,\ldots,N\,\middle|\,\tilde{\zeta}\in Z^t(t)\right) = \left(1-V\left(\tilde{x}^*_t\right)\right)^{N-t}$$

where $V\left(\tilde{x}^*_t\right)$ is defined in (13.105). Hence, conditioned on $Z^t(t)$ as opposed to a $\tilde{\zeta}\in Z^t(t)$, we have

$$\begin{aligned}\mathbb{P}^N\left(\tilde{x}^*_t\in X_{\tilde{\zeta}^i}, i=t+1,\ldots,N\,\middle|\,Z^t(t)\right) &= \int_{Z^t}\left(1-V\left(\tilde{x}^*_t\right)\right)^{N-t}\mathbb{P}^t\left(d\tilde{\zeta}^1,\ldots,d\tilde{\zeta}^t\,\middle|\,Z^t(t)\right)\\ &= \int_0^1(1-v)^{N-t}dF^t(v|Z^t(t))\end{aligned}\tag{13.108}$$

where the second equality follows from (13.105). Substituting this into (13.107) and using part 2 of the lemma proves (13.106). □

We now use Lemmas 13.20 and 13.21 to bound $\mathbb{P}^N\left(V(x_N^*) > \epsilon\right)$ for the general case when CSP(N) may not be uniformly supported.

Proof of Theorem 13.15: general case We will intersect the event $\left(V(x_N^*) > \epsilon\right)$ with the disjoint sets $Z^N(J^t)$ and the summands in (13.103) are

$$\begin{aligned}
\mathbb{P}^N\left(V(x_N^*) > \epsilon, Z^N(J^t)\middle|Z^N(t)\right) &= \mathbb{P}^N\left(V(x_{J^t}^*) > \epsilon, Y^N(J^t)\middle|Z^N(t)\right) \\
&\leq \mathbb{P}^N\left(V(\tilde{x}_t^*) > \epsilon, \tilde{x}_t^* \in X_{\tilde{\zeta}^i}, i = t+1,\ldots,N\middle|Z^t(t)\right) \\
&= \int_{\{V(\tilde{x}_t^*)>\epsilon\}} \left(1 - V\left(\tilde{x}_t^*\right)\right)^{N-t} \mathbb{P}^t\left(d\tilde{\zeta}^1,\ldots,d\tilde{\zeta}^t\middle|Z^t(t)\right) \\
&= \int_\epsilon^1 (1-v)^{N-t} dF^t(v|Z^t(t))
\end{aligned}$$

where the first equality follows because, conditioned on $Z^N(t)$, $x_N^* = x_{J^t}^*$ and $Z^N(I^t) = Y^N(I^t)$ in conditional probability from Lemma 13.21, the inequality follows from (13.107), and the last equality follows from (13.105).

Substituting this into (13.103) we have

$$\begin{aligned}
\mathbb{P}^N\left(V(x_N^*) > \epsilon\right) &= \sum_{t=0}^n \mathbb{P}^N(Z^N(t)) \sum_{J^t} \mathbb{P}^N\left(V(x_N^*) > \epsilon, Z^N(J^t)\middle|Z^N(t)\right) \\
&\leq \sum_{t=0}^n \mathbb{P}^N(Z^N(t)) \binom{N}{t} \int_\epsilon^1 (1-v)^{N-t} d(v^t)
\end{aligned}$$

where we have used $F^t(v) = v^t$ from Lemmas 13.20 and the fact that there are $\binom{N}{t}$ many J^t. It is shown in (13.101) that

$$\binom{N}{t} \int_\epsilon^1 (1-v)^{N-t} d(v^t) = \sum_{i=0}^{t-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i} \leq \sum_{i=0}^{t^{\max}-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}$$

where $1 \leq t^{\max} \leq n$ is an upper bound on the number of generalized support constraints for almost all $\zeta \in Z^N$. We therefore have

$$\mathbb{P}^N\left(V(x_N^*) > \epsilon\right) \leq \sum_{i=0}^{t^{\max}-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i} \tag{13.109}$$

since $\sum_{t=0}^n \mathbb{P}^N(Z^N(t)) = 1$. This implies the bound in Theorem 13.15 for the general case. □

13.3.5 Sample complexity

Theorems 13.14 and 13.15 translate into sample complexity results for CSP(N), making use of the Markov's inequality and the Chernoff bound. They provide thresholds

for N that guarantee sufficiently small violation probability $V(x_N^*)$, in expectation or probability (they are proved in Exercise 13.21).

Corollary 13.22 (Sample complexity)**.** Fix any $N \geq n$ and suppose assumption C13.3 holds. For any $\epsilon \in (0,1)$ and any $\beta \in (0,1)$:

1 $E^N\left(V(x_N^*)\right) \leq \beta$ if $N \geq (n/\beta) - 1$.
2 $\mathbb{P}^N\left(V(x_N^*) > \epsilon\right) \leq \beta$ if $N \geq N(\epsilon,\beta)$ where
   1 $N(\epsilon,\beta) := (n/\epsilon\beta) - 1$;
   2 or

$$N(\epsilon,\beta) := \min\left\{N : \sum_{i=0}^{n-1}\binom{N}{i}\epsilon^i(1-\epsilon)^{N-i} \leq \beta\right\} \tag{13.110}$$

   3 or

$$N(\epsilon,\beta) := \min\left\{N : (N-(n-1))\ln\frac{N-(n-1)}{N(1-\epsilon)} + (n-1)\ln\frac{n-1}{N\epsilon} \geq \ln\frac{1}{\beta}\right\}$$

Example 13.10. Numerical example to compare the thresholds for N in Corollary 13.22. □

13.3.6 Optimality guarantee

In Chapter 13.3.5 we use the violation probability bound of Theorem 13.15 to derive the sample complexity of CSP(N) (13.84). If $N \geq N(\epsilon,\beta)$ in (13.110) then its optimal solution x_N^* is feasible for CCP(ϵ) (13.83) with probability at least $1-\beta$, according to Corollary 13.22. In this subsection we show that the same $N(\epsilon,\beta)$ in (13.110) also guarantees that the optimal value $c^*_{\text{CSP}}(N)$ of CSP(N) is close to the optimal value c^*_{RCP} of the robust program RCP (13.82) and the optimal value $c^*_{\text{CCP}}(\epsilon)$ of the chance constrained program CCP(ϵ) (13.83) with high probability.

The feasibility of x_N^* for CCP(ϵ) with high probability connects $c^*_{\text{CSP}}(N)$ to $c^*_{\text{CCP}}(\epsilon)$, provided $N \geq N(\epsilon,\beta)$. Unless the violation probability $V(x_N^*) = 0$, x_N^* is however infeasible for RCP. The key to connecting $c^*_{\text{CSP}}(N)$ to c^*_{RCP} is that if x is feasible for CSP(ϵ) then it is feasible for a perturbed robust program defined as follows: for $v \in \mathbb{R}^m$,

$$\text{RCP}(v): \quad c^*_{\text{RCP}}(v) := \min_{x\in X\subseteq\mathbb{R}^n} c^{\mathsf{T}}x \quad \text{s.t.} \quad \bar{h}(x) := \sup_{\zeta\in Z} h(x,\zeta) \leq v \tag{13.111}$$

where $c \in \mathbb{R}^n$, $\zeta \in \mathbb{R}^k$ is an uncertain parameter taking value in the uncertainty set $Z \subseteq \mathbb{R}^k$, $v \in \mathbb{R}^m$ is a perturbation vector, $h : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^m$ is a convex (and hence continuous) function in x for every $\zeta \in Z$, and X is a nonempty closed convex set. Since $h(x,\zeta)$ is convex in x for every $\zeta \in Z$, $\bar{h}(x)$ is a convex function. The unperturbed robust program RCP (13.82) is (13.111) with $v = 0$. While CCP(ϵ) relaxes RCP by requiring constraint satisfaction only probabilitistically, RCP(v) relaxes RCP by allowing a certain amount v of violation. To relate the feasibility of CCP(ϵ) and RCP(v) we need the following definition.

Definition 13.6. 1 The *probability of worst-case constraints* is the function $p : X \times \mathbb{R}^m_+ \to [0,1]$ defined as:

$$p(x,b) \;:=\; \mathbb{P}\left(\left\{\zeta \in Z : \exists i := i(\zeta) \;\text{ s.t. }\; \bar{h}_i(x) - h_i(x,\zeta) < b_i\right\}\right)$$

where $\bar{h}(x) := \sup_{\zeta' \in Z} h(x,\zeta')$.

2 A *perturbation bound* with respect to p is the function $\bar{v} : [0,1] \to \mathbb{R}^m_+$ defined as:

$$\bar{v}(\epsilon) \;:=\; \sup\left\{b \in \mathbb{R}^m_+ : \inf_{x \in X} p(x,b) \le \epsilon\right\}$$

where the supremum here is taken componentwise of vectors b.

The motivation for Definition 13.6 is that $\bar{v}(\epsilon)$ connects RCP($\bar{v}(\epsilon)$) to CCP(ϵ), as follows. For each $x \in X$, ζ violates the constraint $h(x,\zeta) \le 0$ if and only if $\bar{h}_i(x) - h_i(x,\zeta) < \bar{h}_i(x)$ for at least one i and therefore $p(x,\bar{h}(x))$ is the violation probability $V(x)$ defined in (13.85a). This means that the chance constraint $V(x) \le \epsilon$ in CCP(ϵ) is equivalent to $p(x,\bar{h}(x)) \le \epsilon$. Hence $V(x) \le \epsilon$ implies $\bar{h}(x) \le \bar{v}(\epsilon)$, componentwise by definition of $\bar{v}(\epsilon)$. This is summarized in the following lemma. It implies that $\bar{v}(\epsilon)$ defines the tightest perturbation vector $v \in \mathbb{R}^m_+$ such that the feasible set of CCP(ϵ) is an inner approximation of the feasible set of RCP(v). We emphasize that, like the violation probability $V(x)$, $p(x,b)$ and hence the perturbation bound $\bar{v}(\epsilon)$, depend on the constraint function h, the uncertainty set Z and the probability measure $\mathbb{P}$.

Lemma 13.23. [148] If x is feasible for the chance constrained program CCP(ϵ) (13.83), then it is feasible for the perturbed robust program RCP($\bar{v}(\epsilon)$) (13.111).

The scenario program CSP(N) (13.84) is a relaxation of the robust program RCP (13.82) and is an approximation of the chance constrained program CCP(ϵ) (13.83). Let x^*_N be the random optimal solution of CSP(N) ensured by C13.3. If $N \ge N(\epsilon,\beta)$ defined in (13.110) then we have

$$c^*_{\text{RCP}}(\bar{v}(\epsilon)) \;\le\; c^*_{\text{CCP}}(\epsilon) \;\lessapprox\; c^{\mathsf{T}} x^*_N \;=\; c^*_{\text{CSP}}(N) \;\le\; c^*_{\text{RCP}} \tag{13.112}$$

where the first inequality follows from Lemma 13.23, $\lessapprox$ means "smaller or equal to with probability at least $1-\beta$" and it follows from Corollary 13.22 since $N \ge N(\epsilon,\beta)$, and the last inequality follows since CSP(N) is a relaxation of RCP. In particular the optimal values of the chance constrained and convex scenario programs lie between those of the robust program and its perturbed counterpart with high probability.

To quantify how close $c^*_{\text{CSP}}(N)$ is to c^*_{RCP} and to $c^*_{\text{CCP}}(\epsilon)$, we will relate the optimal values $c^*_{\text{RCP}}(v)$ and $c^*_{\text{RCP}}(0)$ by establishing, using the envelop theorem, sufficient conditions under which $c^*_{\text{RCP}}(v)$ is Lipschitz continuous. Let the Lagrangian and the dual function of the perturbed robust program (13.111) be: for $v \in \mathbb{R}^m$,

$$L(x,\mu;v) \;:=\; c^{\mathsf{T}} x + \mu^{\mathsf{T}}(\bar{h}(x) - v), \qquad x \in X \subseteq \mathbb{R}^n,\ \mu \in \mathbb{R}^m \tag{13.113a}$$

$$d(\mu;v) \;:=\; \inf_{x \in X \subseteq \mathbb{R}^n} L(x,\mu;v), \qquad \mu \in \mathbb{R}^m_+ \tag{13.113b}$$

For each perturbation vector v, let $(x(v), \mu(v))$ denote a primal-dual optimal solution of (13.111). We make the following assumptions on the perturbed robust program (13.111):

C13.4 For all $\epsilon \in [0,1]$ the perturbation bound $\bar{v}(\epsilon)$ in Definition 13.6 takes value in a compact and convex set $V \subseteq \mathbb{R}^m_+$.
C13.5 For each $v \in V \subseteq \mathbb{R}^m_+$:
1 There exists a unique primal-dual optimal solution $(x(v), \mu(v))$ and it is continuous at v.
2 Strong duality holds at $(x(v), \mu(v))$.
C13.6 [Slater condition]: There exists $\bar{x} \in X$ such that $h(\bar{x}) < v^{\min}$ where $v_i^{\min} := \min\{v_i : v \in V\}$ is the minimum element of V.

Define

$$L_{\text{RCP}} \;:=\; \frac{c^{\mathsf{T}}\bar{x} - \min_{x\in X} c^{\mathsf{T}}x}{\min_i \left(v_i^{\min} - \bar{h}_i(\bar{x})\right)} \;\geq\; 0 \tag{13.114}$$

where $v_i^{\min} := \min\{v_i : v \in V\}$ and $\bar{h}(x) := \sup_{\zeta\in Z} h(x,\zeta)$. The numerator in L_{RCP} is the cost of the Slater point $\bar{x}$ from a lower bound of the optimal cost and the denominator is the smallest gap of $\bar{x}$ from the feasibility boundary.

Lemma 13.24. Consider the perturbed robust program (13.111) and suppose assumptions C13.3–C13.6 hold. Then $c^*_{\text{RCP}}(v)$ is a Lipschitz continuous function on $V \subseteq \mathbb{R}^m_+$, i.e., for all $v_1, v_2 \in V$,

$$\left\|c^*_{\text{RCP}}(v_1) - c^*_{\text{RCP}}(v_2)\right\| \;\leq\; L_{\text{RCP}}\,\|v_1 - v_2\|$$

where $\|\cdot\|$ can either be the Euclidean norm or the ℓ_1 norm and L_{RCP} is defined in (13.114).

Proof For any $v \in V$, assumption C13.5 and the Saddle Point Theorem 8.19 implies that the primal-dual optimal solution $(x(v), \mu(v))$ is a saddle point of (13.113a):

$$L(x(v), \mu; v) \;\leq\; L(x(v), \mu(v); v) \;\leq\; L(x, \mu(v); v), \qquad x \in X,\ \mu \in \mathbb{R}^m_+$$

Clearly $\nabla_v L(x, \mu; v) = -\mu$ is a continuous function on $X \times \mathbb{R}^m_+ \times V$. This, together with assumption C13.5(a), allows us to apply the Saddle-point Envelop Theorem 8.19 which states that $c^*_{\text{RCP}}(v)$ is continuously differentiable and [11]

$$\nabla_v c^*_{\text{RCP}}(v) \;=\; \nabla_v L(x(v), \mu(v); v) \;=\; -\mu(v)$$

Fix any v_1, v_2 in V. The mean value theorem gives $c^*_{\text{RCP}}(v_1) - c^*_{\text{RCP}}(v_2) = \mu^{\mathsf{T}}(u)(v_1 - v_2)$ for some u between v_1 and v_2 ($u \in V$ because V is convex). Hence, by Cauchy-Schwarz inequality,

$$\left\|c^*_{\text{RCP}}(v_1) - c^*_{\text{RCP}}(v_2)\right\| \;\leq\; \|\mu(u)\|\|v_1 - v_2\| \tag{13.115}$$

[11] To be precise, assumption C13.5 should be defined for all $v \in V^\circ$ for some open set containing the compact set V so that $\nabla_v c^*_{\text{RCP}}(v)$ is well defined on the boundary of V.

where the norm $\|\cdot\|$ can either be the Euclidean norm or the ℓ_1 norm. We now bound $\|\mu(v)\|$ over $v \in V$. Fix any $v \in V$. Since $\mu(v)$ attains the optimal value of the perturbed robust program (13.111), strong duality implies:

$$c^*_{\mathrm{CRP}}(v) \;=\; d(\mu(v);v) \;\le\; c^{\mathsf{T}}\bar{x} + \mu^{\mathsf{T}}(v)\left(\bar{h}(\bar{x}) - v\right) \;\le\; c^{\mathsf{T}}\bar{x} + \max_i (h_i(\bar{x}) - v_i) \sum_i \mu_i(v)$$

where the first inequality follows from (13.113b) and the last inequality follows since $\mu(v) \ge 0$. Hence, noting that $\bar{h}(\bar{x}) - v < 0$ by the Slater condition C13.6,

$$\sum_i \mu_i(v) \;\le\; \frac{c^{\mathsf{T}}\bar{x} - c^*_{\mathrm{RCP}}(v)}{\min_i \left(v_i - \bar{h}_i(\bar{x})\right)} \;\le\; \frac{c^{\mathsf{T}}\bar{x} - \min_{x\in X} c^{\mathsf{T}}x}{\min_i \left(v_i - \bar{h}_i(\bar{x})\right)}$$

Since $\mu(v) \ge 0$ we have

$$\|\mu(v)\|_2 \;\le\; \|\mu(v)\|_1 \;\le\; \frac{c^{\mathsf{T}}\bar{x} - \min_{x\in X} c^{\mathsf{T}}x}{\min_i \left(v_i - \bar{h}_i(\bar{x})\right)}$$

Maximizing both sides over the compact set V yields $\sup_{v\in V} \|\mu(v)\| \le L_{\mathrm{RCP}}$. Substituting into (13.115) proves the lemma. □

The next result from [148] uses (13.112) and Lemma 13.24 to quantify how close $c^*_{\mathrm{CSP}}(N)$ is to c^*_{RCP} and to $c^*_{\mathrm{CCP}}(\epsilon)$.

Theorem 13.25 (Optimality guarantees [148])**.** Consider the robust program RCP (13.82), the chance constrained program CCP(ϵ) (13.83), and the convex scenario program CSP(N) (13.84). Suppose assumptions C13.3–C13.6 hold. Given any $\epsilon \in [0,1]$, any $\beta \in [0,1]$ and any $N \ge N(\epsilon,\beta)$ in (13.110), we have

$$\mathbb{P}^N\left(c^*_{\mathrm{RCP}} - c^*_{\mathrm{CSP}}(N) \in [0, C(\epsilon)]\right) \;\ge\; 1-\beta \tag{13.116a}$$

$$\mathbb{P}^N\left(c^*_{\mathrm{CSP}}(N) - c^*_{\mathrm{CCP}}(\epsilon) \in [0, C(\epsilon)]\right) \;\ge\; 1-\beta \tag{13.116b}$$

where

$$C(\epsilon) \;:=\; \min\left\{L_{\mathrm{RCP}}\|\bar{v}(\epsilon)\|_2,\; \max_{x\in X} c^{\mathsf{T}}x - \min_{x\in X} c^{\mathsf{T}}x\right\}$$

L_{RCP} is defined in (13.114) and the perturbation bound $\bar{v} : [0,1] \to \mathbb{R}^m_+$ in Definition 13.6.

Proof The inequalities in (13.112) imply that $c^*_{\mathrm{RCP}} - c^*_{\mathrm{CSP}}(N) \in [0, C_1]$ with probability 1 and $c^*_{\mathrm{CSP}}(N) - c^*_{\mathrm{CCP}}(\epsilon) \in [0, C_1]$ with probability at least $1-\beta$ where $C_1 := \max_{x\in X} c^{\mathsf{T}}x - \min_{x\in X} c^{\mathsf{T}}x$. We are hence left with showing that, with probability at least $1-\beta$, $c^*_{\mathrm{RCP}} - c^*_{\mathrm{CSP}}(N) \le L_{\mathrm{RCP}}\|\bar{v}(\epsilon)\|_2$ and $c^*_{\mathrm{CSP}}(N) - c^*_{\mathrm{CCP}}(\epsilon) \le L_{\mathrm{RCP}}\|\bar{v}(\epsilon)\|_2$.

From (13.112) we have, with probability at least $1-\beta$,

$$c^*_{\mathrm{CSP}}(N) \;\ge\; c^*_{\mathrm{RCP}}(\bar{v}(\epsilon)) \;\ge\; c^*_{\mathrm{RCP}}(0) - L_{\mathrm{RCP}}\,\|\bar{v}(\epsilon)\|_2$$

where the last inequality follows from Lemma 13.24. Hence $c^*_{\mathrm{RCP}} - c^*_{\mathrm{CSP}}(N) \le$

$L_{\text{RCP}}\|\bar{v}(\epsilon)\|_2$ with probability at least $1-\beta$. Furthermore (13.112) implies that, with probability at least $1-\beta$,

$$c^*_{\text{CSP}}(N) - c^*_{\text{CCP}}(\epsilon) \;\le\; c^*_{\text{RCP}} - c^*_{\text{RCP}}(\bar{v}(\epsilon)) \;\le\; L_{\text{RCP}}\|\bar{v}(\epsilon)\|_2$$

where the last inequality follows from Lemma 13.24. □

13.4 Two-stage optimization with recourse

Consider the situation where decisions are made in two stages under uncertainty indexed by ω in a sample space Ω. The first-stage decision x needs to be made before ω is realized and the second-stage decision $y(\omega)$ is made after ω is realized as a function of ω. The first-stage decision x is made taking into account of the effect of uncertainty, e.g., by minimizing not just a first-stage cost in x, but also the expected second-stage cost incurred by $y(\omega)$ given a first-stage decision x. This can be formulated as a two-stage stochastic program with recourse. In this section we study the structure of feasible regions associated with such a problem, the optimal value of the second-stage decision, and the optimality condition and strong duality of the overall problem. As we will see two-stage optimization generally involves extended real-valued functions that will require the use of nonsmooth techniques studied Chapter 12.

13.4.1 Stochastic linear program with fixed recourse

Consider the following two-stage stochastic program with recourse where the second-stage problem is a linear program:

$$\min_{x\in\mathbb{R}^{n_1}} \quad f(x) + E_\zeta\left(\min_{y(\omega)\in\mathbb{R}^{n_2}} q^\mathsf{T}(\omega)y(\omega)\right) \tag{13.117a}$$

$$\text{s.t.} \quad Ax = b,\; x \in K \tag{13.117b}$$

$$T(\omega)x + Wy(\omega) = h(\omega),\; y(\omega) \ge 0, \qquad \forall \omega \in \Omega \tag{13.117c}$$

where

- For the first-stage problem, the real-valued cost function $f : \mathbb{R}^{n_1} \to \mathbb{R}$ is convex (and hence continuous over $\mathbb{R}^{n_1}$), $A \in \mathbb{R}^{m_1\times n_1}$, $b \in \mathbb{R}^{m_1}$, $K \subseteq \mathbb{R}^{n_1}$ is a closed convex cone. For instance $K := \mathbb{R}^{n_1}_+$, the nonnegative quadrant (closed in $\mathbb{R}^{n_1}$). It is important that the first-stage quantities (f, A, b, K) are certain.
- For each sample $\omega \in \Omega$ the second-stage problem is a linear program in $y(\omega)$, with the cost vector $q(\omega) \in \mathbb{R}^{n_2}$, and the constraint parameters $T(\omega) \in \mathbb{R}^{m_2\times n_1}$, $W \in \mathbb{R}^{m_2\times n_2}$, and $h(\omega) \in \mathbb{R}^{m_2}$. The second-stage decision $y(\omega)$ is called a *recourse action* (or *corrective action*). These quantities, except W, are random, dependent on ω. The second-stage problem is generally semi-infinite and intractable when

Ω is an infinite set. The constraint $y(\omega) \geq 0$ does not lose generality because if $y(\omega)$ is allowed to take value in $\mathbb{R}^{n_2}$, it can be replaced by $z_1(\omega) - z_2(\omega)$ where $z_1(\omega) \geq 0$ and $z_2(\omega) \geq 0$ are two nonnegative variables.

- The matrix W is called a *recourse matrix*. It is assumed to be deterministic, i.e., independent of ω, in (13.117c). Problems with deterministic W are said to have *fixed recourse*. In general $W(\omega)$ can also depend on ω. Stochastic programs with random recourse are much more complicated (see Lemma 13.26 and the discussion that follows). We will only deal with problems with fixed recourse.
- The random variable $\zeta := \zeta(\omega)$ is a function of ω and is the column vector

$$\zeta \;:=\; \zeta(\omega) := (q(\omega),\ h(\omega), T_i^{\mathsf{T}}(\omega), i = 1, \ldots, m_2)$$

where $T_i(\omega)$ is the ith row of $T(\omega)$. The size of ζ is $k := n_2 + m_2 + m_2 n_1$. Denote the set of possible values of ζ by $Z := \{\zeta(\omega) \in \mathbb{R}^k : \omega \in \Omega\}$. The expectation E_ζ in (13.117a) is taken with respect to ζ.

To understand the structure of the stochastic program (13.117), re-write it in terms of the solution of the second-stage problem. Given a first-stage decision x and a realization of the random vector $\zeta \in Z$ define the extended real-valued functions $\tilde{Q} : \mathbb{R}^{n_1} \times \mathbb{R}^k \to [-\infty, \infty]$ and $Q : \mathbb{R}^{n_1} \to [-\infty, \infty]$ as:

$$\tilde{Q}(x, \zeta) := \min_{y(\omega) \geq 0} q^{\mathsf{T}}(\omega) y(\omega) \quad \text{s.t.} \quad W y(\omega) = h(\omega) - T(\omega) x \tag{13.118a}$$

$$Q(x) := E_\zeta \tilde{Q}(x, \zeta) \tag{13.118b}$$

In particular, $\tilde{Q}(x, \zeta)$ is defined to be ∞ if the second-stage problem (13.118a) is infeasible for the given x, and $-\infty$ if it is feasible and unbounded below. The case of $\tilde{Q}(x, \zeta) = \infty$ can be a reasonable model of a practical situation (e.g. a generation schedule in the first stage leads to insufficient supply when outages occur in the second stage), but $\tilde{Q}(x, \zeta) = -\infty$ means that the objective can be infinitely improved in the second stage and usually indicates an improper model. We thus usually assume $\tilde{Q}(x, \zeta) > -\infty$ on the domain of interest. The optimal value $\tilde{Q}(x, \zeta)$ of the second-stage problem (13.118a) is called the *second-stage value function* and $Q(x)$ the *second-stage expected value function* or the *recourse function*. Both are extended real-valued functions studied in Chapter 12.2.1.

The stochastic program (13.117) is then equivalent to the following problem:

$$f^* \;:=\; \min_{x \in \mathbb{R}^{n_1}} f(x) + Q(x) \quad \text{s.t.} \quad Ax = b,\ x \in K \tag{13.118c}$$

where the cost function is extended real-valued even though f is real-valued. Comparing the conic program (12.65) studied in Chapter 12.8.4 with (13.118), it is clear that the difficulty of stochastic program (13.118) lies in the structural and computational properties of $Q(x)$. Even though the second-stage problem (13.118a) is a linear program in $y(\omega)$, the recourse function $Q(x)$ is generally not a linear function of x and therefore (13.118) is generally not a linear program. We will show below that, for the problem (13.117) with fixed recourse, if ζ has finite second moment, then $Q(x)$ is

a convex function and (13.118) is indeed a conic program studied in Chapter 12.8.4. Conditions for strong duality and KKT optimality of (13.118) can therefore be derived from Theorem 12.31 (although the computation of $Q(x)$ and its subdifferential is generally difficult). The fact that the second-stage problem (13.118a) is a linear program is important in deriving these results. We therefore sometimes refer to (13.117) as a stochastic linear program.

Tractability.

We start with the feasibility of (13.118) and some basic properties of the recourse function $Q(x)$. We then present the optimality condition and strong duality for the problem when it is convex.

Let $C_1 := \{x \in \mathbb{R}^{n_1} : Ax = b, x \in K\}$. The first-stage decision x is feasible if $x \in C_1$ and if x has a feasible second-stage completion so that (13.118c) is well defined. There are two interpretations of feasible second-stage completion, expressed by the following two definitions:

$$C_2 := \operatorname{dom}(Q) := \{x \in \mathbb{R}^{n_1} : Q(x) < \infty\} \tag{13.119a}$$

$$C_2' := \bigcap_{\text{a.e.}\,\zeta \in Z} \{x \in \mathbb{R}^{n_1} : \tilde{Q}(x,\zeta) < \infty\} \tag{13.119b}$$

The set C_2 consists of x for which the expected $\tilde{Q}(x,\zeta)$ is finite. The set C_2' consists of x for which the second-stage problem is always feasible for almost every (a.e.) $\zeta \in Z$, i.e., for a.e. $\omega \in \Omega$, there exists an $y(\omega) \geq 0$ that satisfies $Wy(\omega) = h(\omega) - T(\omega)x$. If ζ can take only finitely many values, then $C_2 = C_2'$, as the next example shows.

Example 13.11 (Generator scheduling)**.** Consider the scheduling of two independent generators with the same capacity a. A slow but cheap generator must be scheduled in advance of a random demand $\zeta(\omega) > 0$ at a generation level $x \in [0,a]$ and unit cost c_1. A fast but expensive generator can be scheduled after the random demand $\zeta(\omega)$ is realized at a generation level $y(\omega) := y(\zeta(\omega)) \in [0,a]$ and unit cost $c_2 > c_1$. Our goal is to choose $(x, y(w))$ to meet demand $\zeta(\omega)$ at the minimum total expected cost:

$$f^* := \min_{x \in \mathbb{R}} c_1 x + Q(x) \quad \text{s.t.} \quad 0 \leq x \leq a \tag{13.120a}$$

where $Q(x) := E_\zeta \tilde{Q}(x,\zeta)$ and

$$\tilde{Q}(x,\zeta) := \min_{0 \leq y(\omega) \leq a} c_2 y(\omega) \quad \text{s.t.} \quad x + y(\omega) = \zeta(\omega) \tag{13.120b}$$

Given the first-stage decision x and the realized demand $\zeta(\omega)$, the second-stage decision is $y(\omega) := y(\zeta(\omega)) = \zeta(\omega) - x$ if this generation level lies in $[0,a]$; otherwise, the second-stage problem is infeasible and $\tilde{Q}(x,\zeta) = \infty$. This means that the first-stage decision x must satisfy $\zeta(\omega) - a \leq x \leq \zeta(\omega)$ in order that $\tilde{Q}(x,\zeta) = c_2 y(\omega) < \infty$.

Suppose $\zeta(\omega) = a + \epsilon$ with probability p and $\zeta(\omega) = a - \epsilon$ with probability $1 - p$.

Then

$$y(a+\epsilon) = \begin{cases} a+\epsilon-x & \text{if} \quad x \geq \epsilon \\ \text{infeasible} & \text{if} \quad x < \epsilon \end{cases}, \quad \tilde{Q} = \begin{cases} c_2(a+\epsilon-x) & \text{if} \quad x \geq \epsilon \\ \infty & \text{if} \quad x < \epsilon \end{cases}$$

$$y(a-\epsilon) = \begin{cases} a-\epsilon-x & \text{if} \quad x \leq a-\epsilon \\ \text{infeasible} & \text{if} \quad x > a-\epsilon \end{cases}, \quad \tilde{Q} = \begin{cases} c_2(a-\epsilon-x) & \text{if} \quad x \leq a-\epsilon \\ \infty & \text{if} \quad x > a-\epsilon \end{cases}$$

Therefore when $\zeta(\omega) = a+\epsilon$, which happens with probability p, $\tilde{Q}(x,\zeta) = \infty$ if $x < \epsilon$. When $\zeta(\omega) = a-\epsilon$, which happens with probability $1-p$, $\tilde{Q}(x,\zeta) = \infty$ if $x > a-\epsilon$. Hence

$$C_2' := \bigcap_{\zeta} \{x : \tilde{Q}(x,\zeta) < \infty\} = \{x : x \geq \epsilon\} \bigcap \{x : x \leq a-\epsilon\}$$

Moreover if $x < \epsilon$ or $x > a-\epsilon$ then $Q(x) = E_\zeta \tilde{Q}(x,\zeta) = \infty$, i.e.,

$$C_2 := \text{dom}(Q) := \{x : \epsilon \leq x \leq a-\epsilon\}$$

Hence $C_2' = C_2$.

We also have $C_2 \subseteq C_1 := \{x : 0 \leq x \leq a\}$. On C_2,

$$Q(x) = pc_2(a+\epsilon-x) + (1-p)c_2(a-\epsilon-x) = c_2(a+\epsilon(2p-1)) - c_2 x$$

Then (13.120) is:

$$f^* := \min_{x \in \mathbb{R}} (c_1 - c_2)x + c_2(a+\epsilon(2p-1)) \quad \text{s.t.} \quad \epsilon \leq x \leq a-\epsilon$$

Since $c_2 > c_1$, the optimal $x^* = a-\epsilon$ and $f^* = c_1(a-\epsilon) + 2c_2\epsilon p$, i.e., the cheap generator should always produces at the lower level $a-\epsilon$ of the random demand and the expensive generator will pick up the slack (2ϵ with probability p). □

If ζ is a continuous random variable, however, C_2 and C_2' may be different, e.g., when the problem has random rather than fixed recourse or when $E_\zeta \zeta^2 = \infty$ (see Exercise 13.22). The following result provides a sufficient condition for the equivalence of these two interpretations ($C_2 = C_2'$) for the case of fixed recourse.[12]

Lemma 13.26. [143, Theorems 4 and 5, p.111] Consider the stochastic program (13.117) or its equivalent (13.118) with fixed recourse, i.e., W is independent of ω. Suppose ζ has finite second moment. Then

1. $C_2 = C_2' = \text{dom}(Q)$.
2. C_2 is closed and convex.
3. C_2 is polyhedral, i.e., defined by a finite set of linear inequalities, provided
   - $T(\omega) = T$ is fixed; or
   - $T(\omega)$ and $h(\omega)$ are independent and the support of the distribution of $T(\omega)$ is polyhedral.

[12] In general we assume all functions have the necessary properties that allow us to mostly ignore issues with measurability and well-posedness of $Q(x)$ for general distributions. See e.g. [142, Chapter 2.1.3], [143] for discussions on these issues.

We now give an intuition on why a finite second moment is sufficient for $C_2 = C_2'$. The argument also shows the importance of the second-stage problem (13.118a) being a linear program. Suppose the optimal value $\tilde{Q}(x,\zeta)$ is finite. Suppose also for simplicity that $K = \mathbb{R}^{n_1}$. Then an optimal $y^*(\omega)$ of the linear program exists that is an extreme point (vertex) of the feasible set. Such a point is called an optimal basic feasible solution. Rewrite the constraint in (13.118a) as an inequality constraint

$$\tilde{W}y(\omega) \;:=\; \begin{bmatrix} W \\ -W \\ \mathbb{I}_{n_2} \end{bmatrix} y(\omega) \;\geq\; \begin{bmatrix} h(\omega) - T(\omega)x \\ -(h(\omega) - T(\omega)x) \\ 0 \end{bmatrix} \;=:\; d(\omega)$$

where $\mathbb{I}_{n_2}$ is the identity matrix of size n_2. Then an optimal basic feasible solution $y^*(\omega)$ takes the form given in (8.60):

$$y^*(\omega) \;=\; \tilde{W}_{I^*}^{-1}\, d_{I^*}(\omega)$$

where W_{I^*} is a $n_2 \times n_2$ nonsingular submatrix of $\tilde{W}_{I^*}$ and $d_{I^*}(\omega)$ is the corresponding n-subvector of $d(\omega)$ that depend on $y^*(\omega)$. The second-stage value function is

$$\tilde{Q}(x,\zeta) = q^{\mathsf{T}}(\omega)\, y^*(\omega) \;=\; q^{\mathsf{T}}(\omega)\, \tilde{W}_{I^*}^{-1}\, d_{I^*}(\omega)$$

Hence $\tilde{Q}(x,\zeta)$ is a quadratic function in ζ and the finite second moment of ζ implies that $Q(x) := E_\zeta \tilde{Q}(x,\zeta)$ is bounded. If, on the other hand, $W(\omega)$ and hence $\tilde{W}(\omega)$ depend on ω, then $Q(x)$ depends on higher moments of ζ. The assumption of fixed recourse and finite second moments is only sufficient; see [149] for more general sufficient conditions, including for the case where $W(\omega)$ is not fixed.

In view of Lemma 13.26 we will consider stochastic program (13.117) with fixed recourse and assume ζ has finite second moment. Then we will not need to differentiate between $C_2 := \mathrm{dom}(Q)$ and its alternative C_2'. A stochastic program is said to have a *relatively complete recourse* if $C_1 \subseteq \mathrm{dom}(Q)$, i.e., an x that satisfies the first-stage constraint always has a feasible second-stage completion for a.e. $\zeta \in Z$. It is said to have a *complete recourse* if $\{Wy : y \geq 0\} = \mathbb{R}^{m_2}$ regardless of the first-stage decision x, i.e., the positive cone spanned by the columns of W equals $\mathbb{R}^{m_2}$. This means that there is a second-stage completion for any x (not necessarily in C_1) and a.e. ζ. A stochastic program that has a complete recourse has a relatively complete recourse, but the converse may not hold.

The following result implies that the deterministic equivalent (13.118) is a convex and differentiable problem.

Lemma 13.27 (Recourse function $Q(x)$)**.** [143, Theorems 6, p.112] Consider problem (13.118) with fixed recourse, i.e., W is independent of ω. Suppose ζ has finite second moment. Then

1 The recourse function $Q(x)$ is convex and Lipschitz on $\mathrm{dom}(Q) := \{x \in \mathbb{R}^{n_1} : Q(x) < \infty\}$.

2 If the distribution function of ζ is absolutely continuous, then $Q(x)$ is differentiable in the relative interior $\mathrm{ri}(\mathrm{dom}(Q))$ of $\mathrm{dom}(Q)$.
3 Suppose ζ takes finitely many values a.s. Then
- $\mathrm{dom}(Q)$ is closed, convex, and polyhedral.
- $Q(x)$ is piecewise linear and convex on $\mathrm{dom}(Q)$.

Note that $Q(x)$ is convex even for problems with random recourse and without the finite moment assumption; see Lemma 13.29 below (proved in Exercises 13.25).

Example 13.12 ($\partial Q(x)$ and $E_\zeta \partial_x \tilde{Q}(x,\zeta)$)**.** Consider the second-stage linear program (13.118a) with fixed recourse, specified by: $y(\omega) \in \mathbb{R}^2$, $W = [1\ 1]$, $T \in \mathbb{R}^{1\times n}$ is fixed, $h(\omega) \in \mathbb{R}$ is a uniform random variable over $[1,2]$,

$$q_1(\omega) = \begin{cases} 1 & \text{with probability } 1-\alpha \\ -1 & \text{with probability } \alpha \end{cases}, \quad q_2(\omega) = 0 \quad \text{with probability 1}$$

and h and q are independent random variables. The random vector $\zeta := \zeta(\omega) := (q(\omega), h(\omega)) \in \mathbb{R}^3$. For each $\omega \in \Omega$,

$$\tilde{Q}(x,\zeta) := \min_{y\geq 0} \ q_1 y_1 \quad \text{s.t.} \quad y_1 + y_2 = h - Tx \tag{13.121}$$

1 Solve the linear program (13.121) explicitly to obtain the extended real-valued function $Q(x) := E_\zeta \tilde{Q}(x,\zeta)$.
2 Show that the effective domain $\mathrm{dom}(Q) = \{x \in \mathbb{R}^{n_1} : Tx \leq 1\}$ and $\partial Q(\bar{x}) = \alpha T^\mathsf{T} + N_{\mathrm{dom}(Q)}(\bar{x})$ for $\bar{x} \in \mathrm{dom}(Q)$ where $N_X(\bar{x})$ denotes the normal cone of X at $\bar{x} \in X$.
3 For each ζ, derive the extended real-valued function $\tilde{Q}(x,\zeta)$ and $\partial_x \tilde{Q}(\bar{x},\zeta)$ for $\bar{x} \in \mathrm{dom}(Q(\cdot,\zeta))$. (The effective domain of $\tilde{Q}(\cdot,\zeta)$ depends on ζ and is generally different from $\mathrm{dom}(Q)$.)
4 Show that $\partial Q(\bar{x}) = E_\zeta\left(\partial_x \tilde{Q}(\bar{x},\zeta)\right) + N_{\mathrm{dom}(Q)}(\bar{x})$ for $\bar{x} \in \mathrm{dom}(Q)$.

Solution. The distribution function for h is $F_h(\eta) = \mathbb{P}_h(h \leq \eta) = \min\{\max\{\eta - 1, 0\}, 1\}$. From the figure there are two cases:

1 $Tx > 1$: When $1 < Tx \leq 2$, $\tilde{Q}(x,\zeta) = \infty$, i.e., (13.121) is infeasible, with probability $Tx - 1$. When $Tx > 2$ then $\tilde{Q}(x,\zeta) = \infty$ with probability 1. Therefore $Q(x) = \infty$ when $Tx > 1$.
2 $Tx \leq 1$: In this case $\tilde{Q}(x,\zeta) < \infty$ for all ζ. The optimal solution y^* and optimal value of (13.121) are

$$y^* = \begin{cases} (0, h-Tx) & \text{if} \quad q_1 = 1 \\ (h-Tx, 0) & \text{if} \quad q_1 = -1 \end{cases}$$

$$\tilde{Q}(x,\zeta) = q_1 y_1^* = \begin{cases} 0 & \text{if} \quad q_1 = 1 \\ Tx - h & \text{if} \quad q_1 = -1 \end{cases}$$

Hence, in this case, $Q(x) = E_{h|q_1=-1}(Tx - h | q_1 = -1)\,\mathbb{P}_{q_1}(q_1 = -1) = \alpha(Tx - E_h(h))$ where the last equality follows from the independence of h and q. Here $E_h(h) = 1.5$.

Therefore $\mathrm{dom}(Q) = \{x \in \mathbb{R}^{n_1} : Tx \le 1\}$ and the extended real-valued function $Q : \mathbb{R}^{n_1} \to (-\infty, \infty]$ is:

$$Q(x) \;=\; \alpha\,(Tx - E_h(h)) + \delta_{\mathrm{dom}(Q)}(x)$$

where the indicator function is $\delta_X(x) = 0$ if $x \in X$ and ∞ if $x \notin X$. From Table 12.2, the subdifferential of an indicator function is its normal cone, i.e., $\partial\delta_X(\bar{x}) = N_X(\bar{x})$ for any $\bar{x} \in X$. Hence, for all $\bar{x} \in \mathrm{dom}(Q)$,

$$\partial Q(\bar{x}) \;=\; \alpha T^{\mathsf{T}} + N_{\mathrm{dom}(Q)}(\bar{x}) \tag{13.122}$$

In particular if $\bar{x} \in \mathrm{ri}(\mathrm{dom}(Q))$ then $N_{\mathrm{dom}(Q)}(\bar{x}) = \{0\}$ and $\partial Q(\bar{x}) = \{\nabla Q(\bar{x})\} = \left\{\alpha T^{\mathsf{T}}\right\}$.

We now derive, for each fixed $\zeta = (q, h)$, the effective domain $\mathrm{dom}(\tilde{Q}(\cdot, \zeta)) \subseteq \mathbb{R}^{n_1}$ and the proper extended real-valued function $\tilde{Q}(\cdot, \zeta)$ on $\mathbb{R}^{n_1}$. As discussed above, $\tilde{Q}(x, \zeta) = \infty$ if $Tx > 2$ or if $Tx \in (1, 2]$ but $h < Tx$. Otherwise $\tilde{Q}(x, \zeta)$ is real-valued. Specifically, let $C(h) := \{x \in \mathbb{R}^{n_1} : Tx \le h\}$; note that $C(h)$ is a random set depending on $h \in [1, 2]$. Then, given a $\zeta = (q, h)$, for $x \in \mathbb{R}^{n_1}$,

$$\tilde{Q}(x, q_1, h) \;=\; \begin{cases} \delta_{C(h)}(x) & \text{if} \quad q_1 = 1 \\ \delta_{C(h)}(x) + Tx - h & \text{if} \quad q_1 = -1 \end{cases}$$

Hence, for each ζ,

$$\partial_x \tilde{Q}(\bar{x}, q_1, h) \;=\; \begin{cases} N_{C(h)}(\bar{x}) & \text{if} \quad q_1 = 1 \\ N_{C(h)}(\bar{x}) + T^{\mathsf{T}} & \text{if} \quad q_1 = -1 \end{cases}, \qquad \bar{x} \in C(h)$$

We now evaluate $E_\zeta \partial_x \tilde{Q}(\bar{x}, \zeta)$ for $\bar{x}$ in the deterministic set $\mathrm{dom}(Q)$. Note that $\mathrm{dom}(Q) \subset C(h)$ with probability 1; in particular $\mathrm{dom}(Q) = C(h)$ only when $h = 1$ which happens with probability 0. Since q and h are independent we have

$$E_\zeta \partial_x \tilde{Q}(\bar{x}, \zeta) \;=\; (1-\alpha) E_h\left(N_{C(h)}(\bar{x})\right) + \alpha\left(E_h\left(N_{C(h)}(\bar{x})\right) + T^{\mathsf{T}}\right) \;=\; \alpha T^{\mathsf{T}} + E_h\left(N_{C(h)}(\bar{x})\right)$$

We claim that $E_h\left(N_{C(h)}(\bar{x})\right) = 0$. Note that $E_h\left(N_{C(h)}(\bar{x})\right) = \int_{1+}^{2} N_{C(h)}(\bar{x})\,dh$. Since $\bar{x} \in \mathrm{dom}(Q) \subset C(h)$ with probability 1, $\bar{x}$ is in the interior of $C(h)$ with probability for $h \in (1, 2]$. Therefore $E_h\left(N_{C(h)}(\bar{x})\right) = 0$ and $E_\zeta \partial_x \tilde{Q}(\bar{x}, \zeta) = \alpha T^{\mathsf{T}}$ for $\bar{x} \in \mathrm{dom}(Q)$, giving

$$\partial Q(\bar{x}) \;=\; E_\zeta \partial_x \tilde{Q}(\bar{x}, \zeta) + N_{\mathrm{dom}(Q)}(\bar{x}), \qquad \bar{x} \in \mathrm{dom}(Q)$$

from (13.122).

Finally for the polyhedral set $\mathrm{dom}(Q) = \{x \in \mathbb{R}^{n_1} : Tx \le 1\}$, Theorem 12.3 says that $N_{\mathrm{dom}(Q)}(\bar{x}) = \{\lambda T^{\mathsf{T}} \in \mathbb{R}^{n_1} : \lambda \in \mathbb{R}_+ \text{ s.t. } \lambda(T\bar{x} - 1) = 0\}$. Substituting into (13.122) we have

$$\partial Q(\bar{x}) \;=\; E_\zeta \partial_x \tilde{Q}(\bar{x}, \zeta) + N_{\mathrm{dom}(Q)}(\bar{x}) \;=\; \begin{cases} \left\{\alpha T^{\mathsf{T}}\right\} & \text{if} \quad T^{\mathsf{T}}\bar{x} < 1 \\ \left\{(\alpha + \lambda) T^{\mathsf{T}} : \lambda \ge 0\right\} & \text{if} \quad T^{\mathsf{T}}\bar{x} = 1 \end{cases}$$

□

KKT condition and duality.

When the problem (13.118) with fixed recourse has finite second moment, Lemma 13.27 implies that the extended real-valued recourse function $Q(x)$ is convex and hence always subdifferentiable in $\mathrm{ri}(\mathrm{dom}(Q))$, whether or not the distribution of ζ is absolutely continuous. This makes (13.118) a conic program (12.65) studied in Chapter 12.8.4. Recall the dual cone K^* of K in Definition 12.1:

$$K^* := \{\xi \in \mathbb{R}^{n_1} : \xi^\mathsf{T} x \geq 0 \ \ \forall x \in K\} \tag{13.123a}$$

Let the dual variables be $\lambda \in \mathbb{R}^{m_1}$ and $\mu \in K^* \subseteq \mathbb{R}^{n_1}$. Define the Lagrangian function of (13.118):

$$L(x,\lambda,\mu) := f(x) + Q(x) - \lambda^\mathsf{T}(Ax-b) - \mu^\mathsf{T} x, \quad x \in \mathbb{R}^{n_1},\ \lambda \in \mathbb{R}^{m_1},\ \mu \in K^*$$

The dual function is

$$d(\lambda,\mu) := \min_{x \in \mathbb{R}^{n_1}} L(x,\lambda,\mu) = \lambda^\mathsf{T} b + d_0(\lambda,\mu), \quad \lambda \in \mathbb{R}^{m_1},\ \mu \in K^* \tag{13.123b}$$

where

$$d_0(\lambda,\mu) := \min_{x \in \mathbb{R}^{n_1}} \left(f(x) + Q(x) - (A^\mathsf{T}\lambda + \mu)^\mathsf{T} x \right) \tag{13.123c}$$

The dual problem is:

$$d^* := \max_{\lambda \in \mathbb{R}^{m_1}, \mu \in K^*} \lambda^\mathsf{T} b + d_0(\lambda,\mu) \tag{13.123d}$$

We make the following assumptions:

C13.7: *Finite second moment and well posed* $Q(x)$. $E_\zeta \zeta^2 < \infty$ and $Q(x) \in (-\infty, \infty]$.
C13.8:
- $f : \mathbb{R}^{n_1} \to \mathbb{R}$ in (13.118) is convex over $\mathbb{R}^{n_1}$ and K is a closed convex cone.
- *Slater condition.* There exists $\bar{x} \in \mathrm{ri}(\mathrm{dom}(Q)) \cap \mathrm{ri}(K)$ such that $A\bar{x} = b$.

Assumption C13.7 and Lemma 13.27 imply that $Q(x)$ is convex on $\mathrm{dom}(Q)$ (hence subdifferentiable). Assumptions C13.7 and C13.8 imply that Q is proper. The properness and the convexity of Q on $\mathrm{dom}(Q)$, and the existence of $\bar{x} \in \mathrm{ri}(\mathrm{dom}(Q))$ imply that $\partial(f+Q)(x) = \partial f(x) + \partial Q(x)$ for all $x \in \mathrm{dom}(Q)$, according to Theorem 12.18. These properties, together with the Slater condition C13.8, allow us to apply Theorem 12.31 on conic program (or more precisely for an extended real-valued cost function, the Slater Theorem 12.27 and the generalized KKT Theorem 12.21) to the stochastic program (13.118), with the following implication.

Theorem 13.28 (Strong duality and KKT for stochastic LP)**.** Consider problem (13.118) with fixed recourse, i.e., W is independent of ω, and its dual (12.67). Suppose assumptions C13.7 and C13.8 hold. Then

1 *Strong duality and dual optimality*. If the optimal value f^* of (13.118) is finite then there exists a dual optimal solution $(\lambda^*, \mu^*) \in \mathbb{R}^{m_1} \times K^*$ that closes the duality gap, i.e., $f^* = d^* = d(\lambda^*, \mu^*)$.
2 *KKT characterization*. A feasible $x^* \in K$ with $Ax^* = b$ is optimal if and only if there exist subgradients $\xi^* \in \partial f(x^*)$ and $\psi^* \in \partial Q(x^*)$, a dual feasible $(\lambda^*, \mu^*) \in \mathbb{R}^{m_1} \times K^*$ such that

$$\xi^* + \psi^* \;=\; A^\mathsf{T}\lambda^* + \mu^*, \qquad\qquad \mu^{*\mathsf{T}} x^* = 0$$

In this case (x^*, λ^*, μ^*) is a saddle point that closes the duality gap and is primal-dual optimal. □

Example 13.13 (Linear program)**.** Consider problem (13.118) with fixed recourse and its dual (12.67). Suppose $f(x) := c^\mathsf{T}x$ and $K := \mathbb{R}^{n_1}_+$ the nonnegative quadrant. Then $K^* = K = \mathbb{R}^{n_1}_+$, $d_0(\lambda, \mu) = 0$ if $c = A^\mathsf{T}\lambda + \mu$ and $-\infty$ otherwise in which case the dual problem becomes:

$$d^* \quad := \quad \max_{\lambda \in \mathbb{R}^{m_1}, \mu \in \mathbb{R}^{n_1}_+} \lambda^\mathsf{T} b \quad \text{s.t.} \quad c = A^\mathsf{T}\lambda + \mu$$

Suppose Q is differentiable. Then the KKT condition becomes: $x^* \in \mathrm{dom}(Q)$ with $Ax^* = b$ and $x^* \geq 0$ is optimal if and only if there exists $(\lambda^*, \mu^*) \in \mathbb{R}^{m_1} \times \mathbb{R}^{n_1}_+$ such that

$$\nabla Q(x^*) \;=\; -c + A^\mathsf{T}\lambda^* + \mu^*, \qquad\qquad \mu^{*\mathsf{T}} x^* = 0$$

□

Problems with relative complete recourse.

When problem (13.118) has a relative complete recourse we can rewrite the KKT condition in Theorem 13.28 in terms of $E_\zeta \partial_x Q(x^*, \zeta)$ instead of $\partial Q(x^*)$. Then $\partial_x Q(x^*, \zeta)$ can be evaluated using envelop theorems studied in Chapter 8.3.6 (see Exercise 13.24). Write the stochastic program (13.118) as an unconstrained optimization:

$$\min_{x \in \mathbb{R}^{n_1}} \quad f(x) + Q(x) + \delta_{C_1}(x)$$

where $C_1 := \{x \in \mathbb{R}^{n_1} : Ax = b, x \in K\}$, $K \subseteq \mathbb{R}^{n_1}$ is a closed convex cone, and $\delta_{C_1}(x)$ is the indicator function of C_1. The generalized KKT Theorem 12.21 implies that a feasible $x^* \in C_1$ is optimal if and only if

$$0 \;\in\; \partial f(x^*) + \partial Q(x^*) + N_{C_1}(x^*) \tag{13.124}$$

The property $\partial Q(\bar{x}) = E_\zeta \partial_x \tilde{Q}(\bar{x}, \zeta) + N_{\mathrm{dom}(Q)}(\bar{x})$ in Example 13.12 holds more generally. Usually $\partial Q(\bar{x}) = \partial_x E_\zeta Q(\bar{x}, \zeta)$ is not the same as $E_\zeta \partial_x Q(\bar{x}, \zeta)$, i.e., one cannot generally interchange the order of expectation and subderivative. It is shown in [143, Theorem 11, p.117] [150, Proposition 2.11] however that if $\bar{x} \in C_1 \cap \mathrm{dom}(Q)$, i.e., if $\bar{x}$ is feasible for (13.118), then

$$\partial Q(\bar{x}) \;=\; E_\zeta \partial_x Q(\bar{x}, \zeta) + N_{\mathrm{dom}(Q)}(\bar{x})$$

If the stochastic program has a relatively complete recourse, then $C_1 \subseteq \mathrm{dom}(Q)$ and hence $N_{\mathrm{dom}(Q)}(\bar{x}) \subseteq N_{C_1}(\bar{x})$ for all feasible $\bar{x} \in C_1$. This and the fact that $N_{C_1}(\bar{x})$ and $N_{\mathrm{dom}(Q)}(\bar{x})$ are convex cones imply that $N_{C_1}(\bar{x}) + N_{\mathrm{dom}(Q)}(\bar{x}) = N_{C_1}(\bar{x})$. Substituting all this into (13.124) we have: $x^* \in C_1$ is optimal if and only if

$$0 \in \partial f(x^*) + E_\zeta \partial_x Q(x^*, \zeta) + N_{C_1}(x^*) \tag{13.125a}$$

Theorem 12.5 implies that $N_{C_1}(\bar{x}) = \{A^\mathsf{T}\lambda + \mu \in \mathbb{R}^{n_1} : \lambda \in \mathbb{R}^{m_1}, \mu \in K^\circ, \mu^\mathsf{T}\bar{x} = 0\}$ for $\bar{x} \in C_1$ where $K^\circ \subseteq \mathbb{R}^{n_1}$ is the polar cone of K (Exercise 12.13). Therefore, while (13.124) yields the KKT condition in Theorem 13.28, for problems with a relatively complete recourse, (13.125a) yields the equivalent KKT condition in terms of $E_\zeta \partial_x Q(x^*, \zeta)$: $x^* \in C_1$ is optimal if and only if there exists subgradients $\xi^* \in \partial f(x^*)$ and $\psi^* \in E_\zeta \partial_x Q(x^*, \zeta)$, a dual feasible $(\lambda^*, \mu^*) \in \mathbb{R}^{m_1} \times K^*$ such that

$$\xi^* + \psi^* = A^\mathsf{T}\lambda^* + \mu^*, \qquad \mu^{*\mathsf{T}} x^* = 0 \tag{13.125b}$$

It is not common in applicaitons, however, that an analytical expression for $Q(x)$ or $\tilde{Q}(x, \zeta)$ is available. When ζ is a continuous random variable, $Q(x)$ and its derivative generally need to be computed by numerical integration of $Q(x, \zeta)$ and its derivative. This limits the practical solution of stochastic linear programs to problems where the dimensionality of ζ is small. One approach is to approximate a continuous ζ by a discrete random variable.

Multi-stage extension.

The multi-stage extension of the stochastic program with fixed recourse (13.118) is:

$$f^* := \min_{x_0 \in \mathbb{R}^{n_0}} f(x_0) + Q_1(x_0) \quad \text{s.t.} \quad W_0 x_0 = h_0,\ x_0 \in K$$

where the initial decision x_0 is independent of ω and the value function $Q_1(x_0)$ is given by: for $t = 1, \ldots, \tau$,

$$\begin{aligned} Q_t(x_{t-1}(\omega)) &:= E_{\zeta_t(\omega)} \tilde{Q}_t(x_{t-1}(\omega), \zeta_t(\omega)) \\ \tilde{Q}_t(x_{t-1}(\omega), \zeta_t(\omega)) &:= \min_{x_t(\omega) \geq 0} \left(q_t^\mathsf{T}(\omega) x_t(\omega) + Q_{t+1}(x_t(\omega)) \right) \\ &\qquad \text{s.t.} \quad W_t x_t(\omega) = h_t(\omega) - T_t(\omega) x_{t-1}(\omega) \end{aligned}$$

where $Q_{\tau+1}(x) := 0$ at the last stage $t = \tau$. Hence the initial decision x_0 is made before the realization of ω. For each $t = 1, \ldots, \tau$, the stage-t decision $x_t(\omega)$ depends on stage-$(t-1)$ decision $x_{t-1}(\omega)$, the realized stage-t cost $q_t(\omega)$ and constraint parameters $(W_t, h_t(\omega), T_t(\omega))$, as well as the stage-$(t+1)$ value function $Q_{t+1}(x)$. The basic theory on the effective domains $\mathrm{dom}(Q_t)$, the value functions $Q_t(x)$, and optimality conditions can be extended from two-stage to multi-stage problems. Like dynamic programming, a multi-stage stochastic program with recourse can suffer from the curse of dimensionality as the number of stages grows; see [143, Chapter 10] for

computational methods for multi-stage stochastic programs that possess simplifying structures.

13.4.2 Stochastic nonlinear program with general recourse

Consider the stochastic nonlinear program:

$$\inf_{x\in\mathbb{R}^{n_1}} \quad f^1(x)+Q(x) \quad \text{s.t.} \quad A^1x=b^1,\ h^1(x)\le 0 \tag{13.126}$$

where the extended real-valued function $Q:\mathbb{R}^{n_1}\to[-\infty,\infty]$ is $Q(x):=E_\omega\,\tilde{Q}(x,\omega)$, ω takes value in a sample space Ω, and

$$\tilde{Q}(x,\omega) \ := \inf_{y(\omega)\in\mathbb{R}^{n_2}} f^2(x,y(\omega),\omega) \tag{13.127a}$$

$$\text{s.t.} \quad A^2(\omega)x+W(\omega)y(\omega)=b^2(\omega), h^2(x,y(\omega),\omega)\le 0 \tag{13.127b}$$

For first-stage functions, $f^1:\mathbb{R}^{n_1}\to\mathbb{R}$, $A^1\in\mathbb{R}^{m_1\times n_1}$, $b^1\in\mathbb{R}^{m_1}$, $h^1:\mathbb{R}^{n_1}\to\mathbb{R}^{l_1}$. For second-stage functions, $f^2:\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}\times\Omega\to\mathbb{R}$, $A^2(\omega)\in\mathbb{R}^{m_2\times n_1}$, $W(\omega)\in\mathbb{R}^{m_2\times n_2}$ and $b^2(\omega)\in\mathbb{R}^{m_2}$ for each $\omega\in\Omega$, and $h^2:\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}\times\Omega\to\mathbb{R}^{l_2}$. Compared with the stochastic linear program (13.118) the main difference is that the recourse problem (13.127) is generally not a linear program and that the recourse is generally not fixed, i.e., the second-stage functions (f^2,A^2,W,b^2,h^2) generally depend on ω. We ignore measurability issues, i.e., we assume all functions and sets we encounter are measurable. Furthermore we make the following assumptions:

C13.9: *Convexity.*

1 f^1 and h^1 are convex on $\mathbb{R}^{n_1}$.

2 For a.e. $\omega\in\Omega$, $f^2(\cdot,\cdot,\omega)$ and $h^2(\cdot,\cdot,\omega)$ are convex on $\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}$.

We next study properties of the recourse function $Q(x)$ and then optimality conditions. Under assumption C13.9, both $\tilde{Q}(x,\omega)$ and $Q(x)$ are closed convex functions in x (Exercises 13.25), even though their effective domains $\mathrm{dom}(Q(\cdot,\omega))$ and $\mathrm{dom}(Q)$ may not be closed sets (see Remark 12.3).

Lemma 13.29. Consider the stochastic nonlinear program with recourse (13.126)(13.127) and suppose C13.9 holds.

1 $\tilde{Q}(x,\omega)$ and $Q(x)$ are convex on $\mathbb{R}^{n_1}$ for a.e. $\omega\in\Omega$.
2 If for every $x_1\in\mathbb{R}^{n_1}$ the feasible region of the recourse problem (13.127) is bounded, then
   1 $\tilde{Q}(x,\omega)$ and $Q(x)$ are lower semicontinuous on $\mathbb{R}^{n_1}$ for a.e. $\omega\in\Omega$.
   2 $\tilde{Q}(x,\omega)$ and $Q(x)$ are closed functions on $\mathbb{R}^{n_1}$ for a.e. $\omega\in\Omega$.
3 The effective domain $\mathrm{dom}(Q):=\{x\in\mathbb{R}^{n_1}:Q(x)<\infty\}$ is a convex set.

Let $C_1 := \{x \in \mathbb{R}^{n_1} : A^1 x = b^1,\ h^1(x) \le 0\}$. The Weierstrass Theorem 12.22 in Chapter 12.6 implies the existence of primal optimal solution (Exercise 13.26) under the additional assumption:

C13.10: *Well posed* $Q(x)$. $Q(x) \in (-\infty, \infty]$.

Note that it is not necessary for the feasible set $C_1 \cap \mathrm{dom}(Q)$ of (13.126) to be closed.

Theorem 13.30 (Primal optimality)**.** Consider the stochastic nonlinear program with recourse (13.126) and suppose assumptions C13.9 and C13.10 hold. Suppose further that, for every $x_1 \in \mathbb{R}^{n_1}$, the feasible region of the recourse problem (13.127) is bounded. If C_1 is bounded and $C_1 \cap \mathrm{dom}(Q) \neq \emptyset$, then (13.126) has a finite optimal value and it is attained at some $x^* \in \mathbb{R}^{n_1}$.

The stochastic program with general recourse (13.126) can be written equivalently as:

$$f^* := \inf_{x \in \mathbb{R}^{n_1}} f^1(x) + Q(x) \quad \text{s.t.} \quad A^1 x = b^1,\ h^1(x) \le 0 \tag{13.128a}$$

where $Q(x) := E_\omega \tilde{Q}(x, \omega)$ as defined in (13.127). Lemma 13.29 implies that $\mathrm{dom}(Q)$ is a convex set (not necessarily closed) and $Q(x)$ is a convex function on $\mathbb{R}^{n_1}$ under Assumption C13.9, and hence (13.128a) is a convex problem. The Lagrangian is

$$L(x, \lambda, \mu) := f^1(x) + Q(x) + \lambda^\mathsf{T}(A^1 x - b^1) + \mu^\mathsf{T} h^1(x) \tag{13.128b}$$

The dual function is

$$d(\lambda, \mu) := \inf_{x \in \mathbb{R}^{n_1}} L(x, \lambda, \mu), \qquad \lambda \in \mathbb{R}^{m_1},\ \mu \in \mathbb{R}^{l_1} \tag{13.128c}$$

and the dual problem is

$$d^* := \sup_{\lambda, \mu \ge 0} d(\lambda, \mu) \tag{13.128d}$$

For strong duality and dual optimality we need the following additional assumption.

C13.11: *Slater condition.* There exists $\bar{x} \in \mathrm{ri}(\mathrm{dom}(f^1)) \cap \mathrm{ri}(\mathrm{dom}(Q))$ such that $A\bar{x} = b$, and $h^1(\bar{x}) < 0$.

Assumptions C13.10 and C13.11 imply that Q is proper. The properness and the convexity of Q on $\mathrm{dom}(Q)$ (from Lemma 13.29), and the existence of $\bar{x} \in \mathrm{ri}(\mathrm{dom}(Q))$ imply that $\partial(f^1 + Q)(x) = \partial f^1(x) + \partial Q(x)$ for all $x \in \mathrm{dom}(Q)$, according to Theorem 12.18. These properties, together with the Slater condition C13.11, allow us to apply the Slater Theorem 12.27 and the generalized KKT Theorem 12.21 to (13.128), with the following implication (cf. Exercise 12.21).

Theorem 13.31 (Strong duality and KKT for stochastic NLP)**.** Consider the stochastic program with general recourse and its dual (13.128). Suppose assumptions C13.9, C13.10 and C13.11 hold. Then

1 *Strong duality and dual optimality*. If the optimal value f^* of (13.128a) is finite then there exists a dual optimal solution $(\lambda^*, \mu^*) \in \mathbb{R}^{m_1} \times \mathbb{R}^{l_1}_+$ that closes the duality gap, i.e., $f^* = d^* = d(\lambda^*, \mu^*)$.

2 *KKT characterization*. A feasible $x^* \in C_1$ is optimal if and only if there exists a dual feasible $(\lambda^*, \mu^*) \in \mathbb{R}^{m_1} \times \mathbb{R}^{l_1}_+$ such that

$$0 \in \partial f^1(x^*) + \partial Q(x^*) + A^{1\mathsf{T}}\lambda^* + \sum_i \mu_i^* \partial h_i^1(x^*) \qquad \mu^{*\mathsf{T}} h^1(x^*) = 0$$

i.e., there exist subgradients $\xi^* \in \partial f(x^*)$ and $\psi^* \in \partial Q(x^*)$, $\theta_i^* \in \partial h_i^1(x^*)$, and a dual feasible $(\lambda^*, \mu^*) \in \mathbb{R}^{\bar{m}_1} \times \mathbb{R}^{m_1 - \bar{m}_1}_+$ such that

$$0 = \xi^* + \psi^* + A^{1\mathsf{T}}\lambda^* + \Theta^{*\mathsf{T}}\mu^*, \qquad \mu^{*\mathsf{T}} h^1(x^*) = 0$$

where the rows of the matrix Θ^* are θ_i. In this case (x^*, λ^*, μ^*) is a saddle point that closes the duality gap and is primal-dual optimal. □

As for stochastic linear programs, under appropriate conditions, we can express $\partial Q(x)$ in terms of the expectation over ω of $\partial Q(x, \omega)$, as

$$\partial Q(x) = E_\omega \partial_x Q(x, \omega) + N_{\mathrm{dom}(Q)}(x)$$

13.5 Example application: stochastic economic dispatch

In rest of this chapter we present power system examples to illustrate stochastic optimization ideas studied in Chapters 13.1–13.4.

We have studied in Chapter 6.4 the problem of optimally scheduling generations and demands and pricing electricity when there is no uncertainty. In this section we discuss how the nominal economic dispatch problem of Chapter 6.4 can be modified when uncertainty arises. Our main purpose is to illustrate various concepts of stochastic OPF in a concrete application.

Consider a power network modeled by the DC power flow model of Chapter 4.6.2. The network is represented by a connected graph $G = (\overline{N}, E)$ of $N+1$ nodes and $M := |E|$ lines where $\overline{N} := \{0\} \cup N$, $N := \{1, 2, \ldots, N\}$ and $E \subseteq \overline{N} \times \overline{N}$. Let C denote the $(N+1) \times M$ incidence matrix (defined in (4.11)). Each line $l := (j, k) \in E$ is parametrized by its susceptance $b_l > 0$. Let $B := \mathrm{diag}(b_l, l \in E) \succ 0$ be the diagonal matrix of line susceptances. Suppose at each bus j:

- There is possibly an uncontrollable generation $g_j \geq 0$ (e.g. photovoltaic) and an uncontrollable load $d_j \geq 0$. The net demand to the grid is $g_j - d_j$.
- There is a single dispatchable unit p_j taking value within its capacity limits $[p_j^{\min}, p_j^{\max}]$. It can be a generator ($p_j^{\min} \geq 0$), a controllable load ($p_j^{\max} \leq 0$), or a prosumer $p_j^{\min} \leq 0 \leq p_j^{\max}$. Let $f_j(p_i)$ denote the cost function of unit j, i.e.,

$f_j(p_j)$ models the generation cost at a generator bus with $p_j \geq 0$ and $-f_j(p_j)$ models the utility of consuming $-p_j \geq 0$ at a load bus.

Any of (g_j, d_j) and $p_j^{\min} = p_j^{\max}$ can be set to zero if they are not present at node i.

The Laplacian matrix L associated with G is defined to be

$$L \;:=\; CBC^{\mathsf{T}}$$

(See Chapter 4.6.1 for properties of L.) A net injection (vector) $(p+g-d)$ induces power flows P on lines given by

$$P \;=\; S^{\mathsf{T}}p \;:=\; BC^{\mathsf{T}}L^{\dagger}(p+g-d) \tag{13.129}$$

where $S := L^{\dagger}CB$ is called a *shift factor*. The expression (13.129) for P is valid if and only if $\mathbf{1}^{\mathsf{T}}(p+g-d) = 0$, i.e., if and only if supply and demand are balanced. The power flow P_{jk} on each line $j \to k \in E$ is directional (i.e, $P_{jk} < 0$ means power flows from buses k to j). There are line capacities $P_{jk}^{\min} < 0 < P_{jk}^{\max}$ in each direction and the line flows $P = S^{\mathsf{T}}p$ induced by p must lie within these limits.

13.5.1 Nominal ED

We have studied the following nominal economic dispatch in Chapter 6.4 that minimizes aggregate production cost subject to capacity limits, power balance, and line limits, when (g, d) are known:

$$\min_{P^{\min} \leq p \leq P^{\max}} \quad \sum_{j \in \overline{N}} f_j(p_j) \tag{13.130a}$$

$$\text{s.t.} \qquad \mathbf{1}^{\mathsf{T}}(p+g-d) \;=\; 0 \qquad [\gamma] \tag{13.130b}$$

$$P^{\min} \;\leq\; S^{\mathsf{T}}(p+g-d) \;\leq\; P^{\max} \qquad [\kappa^-, \kappa^+] \tag{13.130c}$$

with associated Lagrange multipliers $(\gamma, \kappa^-, \kappa^+)$ with $(\kappa^-, \kappa^+) \geq 0$. The *locational marginal price* (LMP) or *nodal price* is the following vector:

$$\lambda \;:=\; \lambda(\gamma, \kappa) \;=:\; \gamma\mathbf{1} + S\kappa \;:=\; \gamma\mathbf{1} + \left(L^{\dagger}CB\right)\kappa \;\in\; \mathbb{R}^{N+1} \tag{13.131}$$

where $\kappa := \kappa^- - \kappa^+$. The Slater Theorem 8.17 of Chapter 8.3.4 implies that if the cost functions f_j are convex and the economic dispatch (6.22) has a finite optimal value, then there exist optimal Lagrange multipliers $(\gamma^*, \kappa^{-*}, \kappa^{+*})$ and hence an LMP λ^* such that a dispatch p^* is optimal for (13.131) if and only if p^* is primal feasible, $(\kappa^{-*}, \kappa^{+*}) \geq 0$, and $(p^*, \gamma^*, \kappa^{-*}, \kappa^{+*})$ satisfies stationarity:

$$f_j'(p_j^*) \;\begin{cases} = \lambda_j^* & \text{if} \quad p_j^{\min} < p_j^* < p_i^{\max} \\ > \lambda_j^* & \text{only if} \quad p_j^* = p_j^{\min} \\ < \lambda_j^* & \text{only if} \quad p_j^* = p_j^{\max} \end{cases} \tag{13.132a}$$

and complementary slackness:

$$(\kappa^{-*})^{\mathsf{T}}\left(P^{\min} - S^{\mathsf{T}}(p^* + g - d)\right) = 0, \quad (\kappa^{+*})^{\mathsf{T}}\left(S^{\mathsf{T}}(p^* + g - d) - P^{\max}\right) = 0 \quad (13.132b)$$

13.5.2 Robust ED

Suppose the uncontrollable generations and demands (g_j, d_j) are uncertain. For simplicity we take $f_j(p_j) := c_j p_j$ so that the economic dispatch is a linear program. To formulate robust economic dispatch we first relax the power balance equality constraint (13.130b) into an inequality constraint:

$$f_{\min} := \min_{p^{\min} \le p \le p^{\max}} \; c^{\mathsf{T}} p \quad (13.133a)$$

$$\text{s.t.} \quad b^{\min} \le \mathbf{1}^{\mathsf{T}}(p + g - d) \le b^{\max} \quad [\gamma^-, \gamma^+] \quad (13.133b)$$

$$P^{\min} \le S^{\mathsf{T}}(p + g - d) \le P^{\max} \quad [\kappa^-, \kappa^+] \quad (13.133c)$$

with associated Lagrange multipliers $(\gamma^-, \gamma^+, \kappa^-, \kappa^+)$ with $(\gamma^-, \gamma^+) \ge 0$ and $(\kappa^-, \kappa^+) \ge 0$. We assume $b^{\min} < 0 < b^{\max}$ and $P^{\min} < 0 < P^{\max}$. The rationale is that the dispatch decisions and LMP (p^*, λ^*) are made in advance, e.g., 5 or 15 minutes before delivery, before (g, d) are realized. At delivery time when (g, d) are realized, as long as the power imbalance $\mathbf{1}^{\mathsf{T}}(p + g - d)$ over the entire network is small enough in magnitude, it can be met by some reserve generation and demand response in some manner. (In Chapter 6.4.4, we will optimize the scheduling of reservers using two-stage optimization with recourse.) Let $\gamma := \gamma^- - \gamma^+$ and recall $\kappa := \kappa^- - \kappa^+$. Then, as for the nominal ED (13.130), a primal feasible p^* and a dual feasible $(\gamma^{-*}, \gamma^{+*}, \kappa^{-*}, \kappa^{+*})$ are optimal if and only if they satisfy (13.132) with LMP $\lambda^* := \gamma^* \mathbf{1} + S\kappa^*$, as in (13.131) but with $\gamma^* := \gamma^{-*} - \gamma^{+*}$.

Suppose the uncertain generations and loads (g_i, d_i) take values in $G_i \times D_i := [0, g_i^{\max}] \times [0, d_i^{\max}]$ and let $G \times D := (\prod_i G_i) \times (\prod_i D_i)$. The robust counterpart of the relaxed economic dispatch (13.133) chooses an optimal dispatch p^* so that power can be balanced in the worst-case realization of (g, d):

$$f^*_{\text{rED}} := \min_{p^{\min} \le p \le p^{\max}} \; c^{\mathsf{T}} p \quad (13.134a)$$

$$\text{s.t.} \quad b^{\min} \le \mathbf{1}^{\mathsf{T}}(p + g - d) \le b^{\max}, \quad \forall (g, d) \in G \times D \quad (13.134b)$$

$$P^{\min} \le S^{\mathsf{T}}(p + g - d) \le P^{\max}, \quad \forall (g, d) \in G \times D \quad (13.134c)$$

We now show that this semi-infinite problem is equivalent to a finite linear program. The subproblems (13.8) corresponding to the power balance constraint (13.134b) are:

$$\min_{(g,d) \in G \times D} \mathbf{1}^{\mathsf{T}}(g - d) = -\mathbf{1}^{\mathsf{T}} d^{\max}, \qquad \max_{(g,d) \in G \times D} \mathbf{1}^{\mathsf{T}}(g - d) = \mathbf{1}^{\mathsf{T}} g^{\max}$$

Therefore the semi-infinite constraint (13.134b) has the finite reformulation:

$$b^{\min} + \mathbf{1}^\mathsf{T} d^{\max} \;\le\; \mathbf{1}^\mathsf{T} p \;\le\; b^{\max} - \mathbf{1}^\mathsf{T} g^{\max}$$

which is feasible only if $\mathbf{1}^\mathsf{T}(d^{\max} + g^{\max}) \le b^{\max} - b^{\min}$. This constraint says that the dispatch must be able to meet the largest possible demand but also allow the largest possible generation, which can be too conservative.

Denote by s_{jl} the (j,l) entry of $S := L^\dagger CB$ and let $s_l := (s_{jl}, j \in \overline{N})$ denote the lth column of S. Then we have for the lth constraint in (13.134c):

$$\min_{(g,d)\in G\times D} \; s_l^\mathsf{T}(g-d) \;=\; -\left(t_l^-\right)^\mathsf{T}\mathbf{1}, \qquad \max_{(g,d)\in G\times D} \; s_l^\mathsf{T}(g-d) \;=\; \left(t_l^+\right)^\mathsf{T}\mathbf{1}$$

where $t_l^- := t_l^-(s_l)$ and $t_l^+ := t_l^+(s_l)$ are row vectors in $\mathbb{R}_+^{N+1}$ that depend on s_l:

$$t_{lj}^- := \begin{cases} |s_{jl}| d_j^{\max} & \text{if} \quad s_{jl} \ge 0 \\ |s_{jl}| g_j^{\max} & \text{if} \quad s_{jl} \le 0 \end{cases}, \qquad t_{li}^+ := \begin{cases} |s_{jl}| g_j^{\max} & \text{if} \quad s_{jl} \ge 0 \\ |s_{jl}| d_j^{\max} & \text{if} \quad s_{jl} \le 0 \end{cases}$$

Recall that s_{jl} is the marginal increase in line flow ΔP_l for additional unit of injection Δp_j at bus j. Therefore, for upper line limit t_l^+, when node $s_{jl} \ge 0$, the worst-case uncertainty on line l (in terms of pushing the line flow towards $P_l^{\max}$) is $g_j^{\max}$; otherwise the worst-case uncertainty is $d_j^{\max}$. The worst-case realization of (g,d) can be different for different constraints l and the robust formulation requires that the line flow P_l on any line l due to any realization (g_k, d_k) must stay within its line limit $(P_l^{\min}, P_l^{\max})$. Let T^- and T^+ be matrices whose lth rows are the row vectors t_l^- and t_l^+ respectively. Then the semi-infinite constraint (13.134c) has the finite reformulation:

$$P^{\min} + T^-\mathbf{1} \;\le\; S^\mathsf{T} p \;\le\; P^{\max} - T^+\mathbf{1}$$

Therefore the semi-infinite robust program (13.134) can be reformulated as a linear program:

$$f_{\text{rED}}^* := \min_{p^{\min}\le p\le p^{\max}} \; c^\mathsf{T} p \tag{13.135a}$$

$$\text{s.t.} \quad b^{\min} + d^{\max\mathsf{T}}\mathbf{1} \;\le\; \mathbf{1}^\mathsf{T} p \;\le\; b^{\max} - g^{\max\mathsf{T}}\mathbf{1} \quad [\gamma^-, \gamma^+] \tag{13.135b}$$

$$P^{\min} + T^-\mathbf{1} \;\le\; S^\mathsf{T} p \;\le\; P^{\max} - T^+\mathbf{1} \quad [\kappa^-, \kappa^+] \tag{13.135c}$$

with associated Lagrange multipliers $(\gamma^-, \gamma^+, \kappa^-, \kappa^+)$ with $(\gamma^-, \gamma^+) \ge 0$ and $(\kappa^-, \kappa^+) \ge 0$. As for the relaxed economic dispatch (13.133), a primal feasible p^* and a dual feasible $(\gamma^{-*}, \gamma^{+*}, \kappa^{-*}, \kappa^{+*})$ are optimal if and only if the stationarity and complementary slackness conditions (13.132) hold with LMPs defined as $\lambda^* := \gamma^*\mathbf{1} + S\kappa^*$ where $\gamma^* := \gamma^{-*} - \gamma^{+*}$ and $\kappa^* := \kappa^{-*} - \kappa^{+*}$.

The lower and upper limits on power imbalance and line flows are however tighter in the robust program (13.135) than those in (13.133). The tightening accommodates the worst-case uncertainty and can be too conservative.

13.5.3 Chance constrained ED

The constraints (13.133b)(13.133c) are

$$\begin{aligned} \mathbf{1}^\mathsf{T}(g-d) &\geq b^{\min} - \mathbf{1}^\mathsf{T} p, \qquad & S^\mathsf{T}(g-d) &\geq P^{\min} - S^\mathsf{T} p \\ \mathbf{1}^\mathsf{T}(g-d) &\leq b^{\max} - \mathbf{1}^\mathsf{T} p, \qquad & S^\mathsf{T}(g-d) &\leq P^{\max} - S^\mathsf{T} p \end{aligned}$$

Define the random (column) vector taking value in $\mathbb{R}^{M+1}$:

$$\zeta := \left(\mathbf{1}^\mathsf{T}(g-d),\, S^\mathsf{T}(g-d)\right) = \begin{bmatrix} \mathbf{1} & S \end{bmatrix}^\mathsf{T} (g-d) \tag{13.136a}$$

Let $F_\zeta(z)$ denote the distribution function of ζ and assume it is continuous. Let $h_1 : \mathbb{R}^{N+1} \to \mathbb{R}^{M+1}$ and $h_2 : \mathbb{R}^{N+1} \to \mathbb{R}^{M+1}$ be the affine functions:

$$h_1(p) := \left(b^{\min} - \mathbf{1}^\mathsf{T} p,\, P^{\min} - S^\mathsf{T} p\right) \tag{13.136b}$$

$$h_2(p) := \left(b^{\max} - \mathbf{1}^\mathsf{T} p,\, P^{\max} - S^\mathsf{T} p\right) \tag{13.136c}$$

Then the chance constrained formulation (13.49) of the relaxed economic dispatch (13.133) is:

$$f^*_{\text{ccED}} := \min_{p^{\min} \leq p \leq p^{\max}} \sum_{j \in \overline{N}} f_j(p_j) \quad \text{s.t.} \quad F_\zeta(h_2(p)) - F_\zeta(h_1(p)) \geq 1 - \epsilon \tag{13.136d}$$

corresponding to the chance constraint $\mathbb{P}(h_1(p) \leq \zeta \leq h_2(p)) \geq 1 - \epsilon$. Even if $F_\zeta(h_j(p))$ are concave functions in p (which will be the case if $F_\zeta(z)$ is concave in z since F_ζ is nondecreasing and nonnegative), their difference $F_\zeta(h_2(p)) - F_\zeta(h_1(p))$ may not be concave in p, and hence the chance constrained economic dispatch (13.136) is generally a nonconvex problem.

13.5.4 Scenario-based ED

Suppose $\left(\zeta^1, \ldots, \zeta^K\right)$ with $K \geq N+1$ are iid samples according to the distribution function F_z. Then the scenario program corresponding to (13.136) is:

$$f^*_{\text{sED}} := \min_{p^{\min} \leq p \leq p^{\max}} \sum_{j \in \overline{N}} f_j(p_j) \quad \text{s.t.} \quad h_1(p) \leq \zeta^k \leq h_2(p), \;\; k = 1, \ldots, K \tag{13.137}$$

Suppose the cost is linear, i.e., $f_j(x_j) = c_j x_j$ and that the minimum cost f^*_{sED} in (13.137) is finite for every realization $\zeta := \left(\zeta^1, \ldots, \zeta^K\right)$. Then (13.137) is a linear program for every ζ since $h_i(p)$ are affine functions and Theorem 8.23 on linear program optimality implies that both primal and dual optimal values are attained. Let p^*_K denote an optimal solution of the randomized problem (13.137). It violates the chance constraint in (13.136d) with a (random) probability $V(p^*_K)$ with mean $E^N\left(V(p^*_K)\right) \leq (N+1)/(K+1)$ according to Theorem 13.14. Moreover Theorem 13.15

implies that the tail probability of $V(p_K^*)$ is bounded by a Binomial tail:

$$\mathbb{P}^K\left(V\left(p_K^*\right) > \epsilon\right) \;\le\; \sum_{i=0}^{N} \binom{K}{i} \epsilon^i (1-\epsilon)^{K-i}$$

For any $\beta > 0$ we can choose the number K of samples greater than the threshold $K(\epsilon, \beta)$ given in (13.110) to guarantee that the $\mathbb{P}^K\left(V\left(p_K^*\right) > \epsilon\right) \le \beta$. Moreover such a K will ensure that, with probability at least $1-\beta$, the optimal value f_{sED}^* of the scenario program is close to the optimal values f_{rED}^* and f_{ccED}^* according to Theorem 13.25.

Let

$$\tilde{S} \;:=\; \begin{bmatrix} \mathbf{1} & S \end{bmatrix}, \qquad \tilde{P}^{\min} \;:=\; \begin{bmatrix} b^{\min} \\ P^{\min} \end{bmatrix}, \qquad \tilde{P}^{\max} \;:=\; \begin{bmatrix} b^{\max} \\ P^{\max} \end{bmatrix}$$

For each realization of the K samples $(\zeta^1, \ldots, \zeta^K)$, the scenario program (13.137) is a convex program and a special case of robust ED (13.134) with a finite set of uncertain values for (g, d):

$$f_{\text{sED}}^* \;:=\; \min_{p^{\min} \le p \le p^{\max}} \sum_{j \in \overline{N}} f_j(p_j) \tag{13.138a}$$

$$\text{s.t.} \qquad \tilde{P}^{\min} - \min_k \zeta^k \;\le\; \tilde{S}^{\mathsf{T}} p \;\le\; \tilde{P}^{\max} - \max_k \zeta^k \tag{13.138b}$$

Therefore for the (randomized) program (13.138), LMP λ_K^* can be defined in the same way as that for (13.133), but with possibly a tighter constraint. A primal feasible p_K^* and a dual feasible $(\gamma_K^{-*}, \gamma_K^{+*}, \kappa_K^{-*}, \kappa_K^{+*})$ are optimal if and only if the stationarity and complementary slackness conditions (13.132) hold.

13.5.5 Special case: no congestion

We illustrate the impact of uncertainty on the prices, optimal dispatch and cost using the following special case:

1. Omit line limits, i.e., the line capacities are large enough not to pose any constraint. This problem is traditionally called the economic dispatch.
2. All units are generators with cost functions $f_j(p_j) := p_j^2/(2\eta_j)$ over $[0, \infty]$ where $\eta_j > 0$. We assume no production limits, i.e., the generators' capacities are large so that their generation levels will be constrained by their quadratically increasing costs rather than capacity limits.

Let $\zeta := \mathbf{1}^{\mathsf{T}}(g - d)$ denote the total uncontrollable excess generation and we assume it takes value in a compact set Z. Then the nominal economic dispatch (13.133) given a realization $\zeta^0 \in Z$, its robust counterpart (13.135), and the scenario-based ED (13.137) are all convex quadratic programs of the form (we assume $\underline{b} > 0$):

$$f^*(\underline{b}) \;:=\; \min_{p \ge 0} \quad \sum_j f_j(p_j) \quad \text{s.t.} \quad \underline{b} \;\le\; \mathbf{1}^{\mathsf{T}} p \;\le\; \overline{b} \qquad [\gamma^-, \gamma^+] \tag{13.139a}$$

with respectively

$$f^*_{\min}: \quad 0 < \underline{b} := b^{\min} - \zeta^0, \qquad \overline{b} := b^{\max} - \zeta^0 \tag{13.139b}$$

$$f^*_{\text{rED}}: \quad 0 < \underline{b} := b^{\min} - \min_{\zeta \in Z} \zeta, \qquad \overline{b} := b^{\max} - \max_{\zeta \in Z} \zeta \tag{13.139c}$$

$$f^*_{\text{sED}}: \quad 0 < \underline{b} := b^{\min} - \min_{k} \zeta^k, \qquad \overline{b} := b^{\max} - \max_{k} \zeta^k \tag{13.139d}$$

where the scenario-based ED is a randomized program defined by K independent random samples of the total uncontrollable excess generation $\zeta^1, \ldots, \zeta^K$.

We now analyze the LMP $\gamma^* := \gamma^{-*} - \gamma^{+*}$ and optimal dispatch programs in p^* for (13.139) and compare their optimal values $f^*_{\min}, f^*_{\text{rED}}, f^*_{\text{sED}}(K)$. Since the marginal costs $f_j'(p_j) = p_j/\eta_j > 0$ for $p_j > 0$ for all j, $\gamma^* = f_j'(p_j^*) > 0$ at optimality and the lower bound of the power balance constraint is tight. Given any $\gamma > 0$, $p_j(\gamma) := f_j'^{-1}(\gamma) = \eta_j \gamma$ is the amount that is incentive compatible for unit j to produce. At optimality, power balance becomes $\underline{b} = \gamma^* \sum_{j=0}^N \eta_j$, and hence

$$\gamma^* = \frac{\underline{b}}{\sum_i \eta_i}, \qquad p_j^* = p_j(\gamma^*) = \frac{\eta_j}{\sum_i \eta_i}\underline{b}, \qquad f^*(\underline{b}) = \frac{\underline{b}^2}{2\sum_i \eta_i} \tag{13.140}$$

Hence the optimal cost f^* depends only on the lower limit $\underline{b}$. We can interpret η_j as a participation factor: generator j produces a share of the minimum excess demand $\underline{b}$ proportional to its η_j. Define the deterministic quantity ζ_Z and the random variable ζ_K as:

$$\zeta_Z := \min_{z \in Z} \zeta, \qquad \zeta_K := \min_k \zeta^k$$

i.e., ζ_Z represents the worst-case demand ($-\zeta_Z > 0$ is the largest in Z) and ζ_K represents the worst-case demand among the K random samples. Applying (13.140) to (13.139), the differences in LMPs, optimal dispatches and optimal costs under robust and scenario-based ED, in comparison with the nominal ED (13.133) if the realization of (g, d) were known in advance, are respectively

$$\gamma^*_{\text{rED}} - \gamma^*_{\min} = \frac{\zeta^0 - \zeta_Z}{\sum_i \eta_i} \geq 0, \qquad \gamma^*_{\text{sED}} - \gamma^*_{\min} = \frac{\zeta^0 - \zeta_K}{\sum_i \eta_i}$$

$$p^*_{\text{rED},j} - p^*_{\min,j} = \frac{\eta_j(\zeta^0 - \zeta_Z)}{\sum_i \eta_i} \geq 0, \qquad p^*_{\text{sED},j} - p^*_{\min,j} = \frac{\eta_j(\zeta^0 - \zeta_K)}{\sum_i \eta_i}$$

and the differences in the optimal costs are:

$$f^*_{\text{rED}} - f^*_{\min} = \frac{1}{2\sum_j \eta_j}\left(\zeta^0 - \zeta_Z\right)\left(2b^{\min} - \zeta^0 - \zeta_Z\right) \geq 0$$

$$f^*_{\text{sED}} - f^*_{\min} = \frac{1}{2\sum_j \eta_j}\left(\zeta^0 - \zeta_K\right)\left(2b^{\min} - \zeta^0 - \zeta_K\right)$$

Since the worst-case demand is always higher, i.e., $\zeta_0 \geq \zeta_Z$ a.s., robust ED always produces a larger LMP, dispatches more power and incurs a higher optimal cost than the nominal ED (13.133). This may not be the case with scenario-based ED since it is a randomized program. If $\zeta^0 < \zeta_K$ (i.e., actual excess generation is less than the

scenario minimum), then $f^*_{\text{sED}} < f^*_{\min}$ though the scenario-based dispatch will not meet the actual supply and will rely on reserves. On the other hand

$$f^*_{\text{rED}} \;\geq\; \max\left\{f^*_{\min}, f^*_{\text{sED}}\right\} \qquad \text{a.s.}$$

Suppose ζ^0 is also drawn from the same distribution as the K random samples in scenario-based ED. Then the expected optimality gaps are, from (13.139) and (13.140),

$$f^*_{\text{rED}} - Ef^*_{\min} \;=\; \frac{1}{2\sum_j \eta_j}\left(2b^{\min}(E\zeta - \zeta_Z) - \left(E(\zeta^2) - \zeta_Z^2\right)\right) \;\geq\; 0$$

$$Ef^*_{\text{sED}} - Ef^*_{\min} \;=\; \frac{1}{2\sum_j \eta_j}\left(2b^{\min}(E\zeta - \zeta_K) - \left(E(\zeta^2) - E(\zeta_K^2)\right)\right) \;\geq\; 0$$

where EX denotes the expectation of the random variable X.

Finally consider the security constrained economic dispatch (6.40) and assume $f_{kj}(p_j) := p_j^2/(2\eta_j)$ for all scenarios k. Assume $\zeta := \mathbf{1}^{\mathsf{T}}(g-d)$ can take only finitely many values $\zeta_1, \ldots, \zeta_K$. (Note that k here indexes the K different deterministic values ζ can take, not random samples in scenario-based ED.) Then we have the deterministic two-stage program with recourse (reserves $(r^{\min}, r^{\max})$ play no role because we have assumed generators have no capacity limits):

$$\begin{aligned} f^*_{\text{scED}} \;:=\; \min_{\substack{p, r^{\min}, r^{\max} \\ (r_k, k\geq 1)}} \quad & \sum_k w_k \sum_j f_{kj}(p_j + r_{kj}) \;:=\; \sum_k w_k \sum_j \frac{1}{2\eta_j}\left(p_j + r_{kj}\right)^2 \\ \text{s.t.} \quad & \mathbf{1}^{\mathsf{T}}(p + r_k) \;=\; -\zeta_k & [\gamma_k] \\ & h_k\left(r^{\min}, r^{\max}\right) := \sum_j h_{kj}\left(r_j^{\min}, r_j^{\max}\right) \;\geq\; 0 & [\mu_k] \end{aligned}$$

where we recall that $\zeta_k := \mathbf{1}^{\mathsf{T}}(g_k - d_k)$ is the total uncontrollable excess generation. The optimal scenario-dependent LMP γ^*_k/w_k, generations, and cost are respectively

$$\frac{\gamma^*_k}{w_k} \;=\; \frac{-\zeta_k}{\sum_i \eta_i}, \qquad p^*_j + r^*_{kj} \;=\; -\frac{\eta_j \zeta_k}{\sum_i \eta_i}, \qquad f^*_{\text{scED}} \;=\; \frac{1}{2\sum_i \eta_i}\sum_k w_k \zeta_k^2$$

If one knew scenario k will be materialized, we assume here that one solves the economic dispatch (13.133) with $b^{\min} = b^{\max} := 0$ to produce the LMP and optimal dispatch and, incurs an optimal cost, from (13.140),

$$\gamma^*_{\min} \;=\; \frac{-\zeta_k}{\sum_i \eta_i}, \qquad p^*_{\min,j} \;=\; -\frac{\eta_j \zeta_k}{\sum_i \eta_i}, \qquad f^*_{\min} \;=\; f^*(-\zeta_k) \;=\; \frac{\zeta_k^2}{2\sum_j \eta_j}$$

Hence, without the reliability requirement (6.39c) (nor startup or ramping constraints), reserves play no role and the two-stage optimization with recourse will be the same as a single-stage decision after observing the realization of (g, d) because the actual generations $p^*_j + r^*_{kj} = -\eta_j \zeta_k / \sum_i \eta_i$ in the security constrained ED can always exactly meet the realized excess load $-\zeta_k > 0$. Hence the expected optimality gap of security

constrained ED under these (unrealistic) assumptions is:

$$f^*_{\text{scED}} - E f^*_{\min} \;=\; \frac{1}{2\sum_i \eta_i}\left(\sum_k w_k \zeta_k^2 - E(\zeta_k^2)\right) \;=\; 0$$

13.6 Example application: security constrained unit commitment

We have formulated in Chapter 6.4.5 security constrained unit commitment as a two-stage stochastic linear program with fixed recourse (studied in Chapter 13.4.1) where the second-stage cost is the expected cost. The first-stage variable $u := (u_j(t)\ \forall j \forall t)$ are binary commitment decisions for all units j in all periods t. The second-stage variable $x(t) := (p(t), r^{\min}(t), r^{\max}(t), r_k(t), k \geq 1)$ in period t are dispatch and reserve amounts for all units in t. Let $x := (x(t) \forall t)$. The uncertainty ω takes a finite number values indexed by $k = 1, \ldots, K$. It takes the form (assuming the dispatch costs and the reserve requirement functions h_{tk} are linear functions; see (6.47)):

$$\min_u \quad f(u) + E_\omega \tilde{Q}^*(u, \omega) \quad \text{s.t.} \quad Au \leq b \tag{13.141a}$$

where, given the first-stage decision u and uncertainty ω, the second-stage problem is the linear program:

$$\tilde{Q}^*(u, \omega) \;:=\; \min_x \; q^{\mathsf{T}}(\omega) x \quad \text{s.t.} \quad T(\omega)u + Wx(\omega) \leq h(\omega) \tag{13.141b}$$

In this section we present an alternative formulation from [119] that combines the idea of two-stage optimization with recourse with robust optimization where the second-stage cost is not the expected cost, but the worst-case cost.

13.6.1 Two-stage adaptive robust formulation

Suppose the uncertain parameter is the uncontrollable net demand $\zeta(t) := d(t) - g(t)$ in period t that takes continuous values in the uncertainty set:

$$Z^t \;:=\; \left\{ \zeta(t) \in \mathbb{R}^{N+1} : \sum_j \frac{|\zeta_j(t) - \bar{\zeta}_j(t)|}{\hat{\zeta}_j(t)} \leq \Delta^t,\ |\zeta_j(t) - \bar{\zeta}_j(t)| \leq \hat{\zeta}_j(t)\, \forall j \right\}$$

where $\bar{\zeta}_j(t)$ and $\hat{\zeta}_j(t)$ are the forecast net demand and the maximum forecast error respectively for $t = 1, \ldots, T$. Let $\zeta := (\zeta(t) \forall t) \in Z := Z^1 \times \cdots \times Z^T$. The first-stage variable $u := (u_j(t)\ \forall j \forall t) \in \{0,1\}^{(N+1)T}$ are binary commitment decisions for all units j in all periods t. The second-stage variable $x(t, \zeta) := (p(t), r^{\min}(t), r^{\max}(t), r(t, \zeta))$ in period t are dispatch and reserve amounts for all units in t in response to $\zeta(t) \in Z^t$. Let $x(\zeta) := (x(t, \zeta)\ \forall t) \in \mathbb{R}^{4(N+1)T}$ be the second-stage responses to ζ for all periods. The

two-stage robust adaptive formulation of security constrained unit commitment takes the form (cf. (13.141)):

$$\min_{u\in\{0,1\}^{(N+1)T}} \; f(u) + \max_{\zeta\in Z} \tilde{Q}^*(u,\zeta) \qquad \text{s.t.} \qquad Au \le b \tag{13.142}$$

where, given a commitment decision $u \in \{0,1\}^{(N+1)T}$ and a net demand $\zeta \in Z$, the second-stage problem is:

$$\tilde{Q}^*(u,\zeta) \;:=\; \min_{x\in\mathbb{R}^{4(N+1)T}} \; q^\mathsf{T}(\zeta)x \quad \text{s.t.} \quad T(\zeta)u + Wx(\zeta) \le h(\zeta) \tag{13.143}$$

i.e., the optimal commitment decision u^* is chosen to minimize the worst-case optimal dispatch cost where the worst case is over all possible uncontrollable net demands $\zeta \in Z$.

Suppose for each (u,ζ) the linear program (13.143) is feasible so that strong duality holds and $\tilde{Q}^*$ is finite (Theorem 8.23). The dual problem is

$$\tilde{Q}^*(u,\zeta) \;=\; \max_{\mu\ge 0} \; (T(\zeta)u - h(\zeta))^\mathsf{T}\mu \quad \text{s.t.} \quad W^\mathsf{T}\mu + q(\zeta) = 0$$

Therefore the problem $\max_{\zeta\in Z} \tilde{Q}^*(u,\zeta)$ in (13.142) becomes:

$$\max_{\zeta\in Z} \tilde{Q}^*(u,\zeta) \;=\; \max_{\zeta\in Z,\mu\ge 0} \; (T(\zeta)u - h(\zeta))^\mathsf{T}\mu \quad \text{s.t.} \quad W^\mathsf{T}\mu + q(\zeta) = 0 \tag{13.144}$$

This problem is generally intractable if $(T(\zeta), h(\zeta), q(\zeta))$ depend on ζ (even with fixed recourse W). For the security constrained unit commitment problem (6.47), suppose the matrix $T(\zeta)$ and the cost vector $q(\zeta)$ in (13.144) are independent of ζ, i.e., $T(\zeta) = T$ and $q(\zeta) = q$. This will be the case if the cost functions $f_t(p(t) + r(t,\zeta),\zeta)$ and the reserve requirement functions $h_t(r^{\min}(t), r^{\max}(t), \zeta)$ of the security constrained real-time dispatch problem f^* in (6.47) are of the form:

$$f_t(y,\zeta) \;:=\; q_t^\mathsf{T} y, \qquad h_t(\underline{r},\bar{r},\zeta) \;:=\; \underline{h}_t^\mathsf{T}\underline{r} + \bar{h}_t^\mathsf{T}\bar{r}$$

i.e., the coefficients $q_t, \underline{h}_t, \bar{h}_t$ are independent of the uncertain parameter ζ. The uncertain parameter ζ enters only into (6.47g)(6.47h) in a way that the cost coefficient $h(\zeta) = \zeta$ in (13.144). Therefore the problem (13.144) becomes:

$$Q^*(u) \;:=\; \max_{\zeta\in Z} \tilde{Q}^*(u,\zeta) \;=\; \max_{\zeta\in Z,\mu\ge 0} \; (Tu - \zeta)^\mathsf{T}\mu \quad \text{s.t.} \quad W^\mathsf{T}\mu + q = 0 \tag{13.145}$$

In particular the feasible set is a fixed polyhedron independent of ζ. The only nonlinearity is the bilinear term $\zeta^\mathsf{T}\mu$ in the objective function. Bilinear programs are generally NP-hard.

13.6.2 Solution

In the following we present the solution method from [119] for the two-stage optimization (13.142)(13.145). It is a two-level algorithm where the outer level uses Benders decomposition for solving (13.142) for the commitment decision u using cuts

generated from the inner level. The inner level solves the bilinear program (13.145) approximately.

Outer algorithm: Benders decomposition

Step 0 Choose any feasible first-stage solution u_0 to (13.142). Solve $Q^*(u_0)$ in (13.145) to get an initial solution (ζ_1, μ_1). Set $C^{\text{lb}} := -\infty$, $C^{\text{ub}} := \infty$ and $k := 1$.

Step 1 Solved the mixed integer program

$$\min_{u,\alpha} \; f(u) + \alpha \quad \text{s.t.} \quad Ax \le b,\; \alpha \ge (Tu - \zeta_l)^{\mathsf{T}} \mu_l,\; l \le k \tag{13.146}$$

This step solves (13.142) with Q^* in (13.145) approximated by α. Let the optimum be (u_k, α_k) and the minimum value $C^{\text{lb}} := f(u_k) + \alpha_k$.

Step 2 Solve a linearized version of the inner problem $Q^*(u_k)$ in (13.145) and denote its optimal solution by (ζ_{k+1}, μ_{k+1}). Let the maximum value be $C^{\text{ub}} := f(u_k) + Q^*(u_k)$ (see below).

Step 3 If $C^{\text{ub}} - C^{\text{lb}} < \epsilon$, stop and return u_k; otherwise, set $k := k + 1$ and goto Step 1.

Inner algorithm: bilinear program $Q^*(u)$

Step 0 Choose an initial $\zeta_1 \in Z$. Set $O^{\text{lb}} := -\infty$, $O^{\text{ub}} := \infty$ and $j := 1$.

Step 1 Solved the (dual) linear program $\tilde{Q}^*(u_k, \zeta_j) := \max_{\mu \ge 0} (Tu_k - \zeta_j)^{\mathsf{T}} \mu$ s.t. $W^{\mathsf{T}} \mu + q = 0$ in (13.145). Let the optimum be μ_j and the linearization of the bilinear term $\zeta^{\mathsf{T}} \mu$ around (ζ_j, μ_j) be

$$L_j(\zeta, \mu) \;:=\; \zeta_j^{\mathsf{T}} \mu_j + (\mu - \mu_j)^{\mathsf{T}} \zeta_j + (\zeta - \zeta_j)^{\mathsf{T}} \mu_j$$

Let $O^{\text{lb}} := (Tu_k - \zeta_j)^{\mathsf{T}} \mu_j$.

Step 2 If $O^{\text{ub}} - O^{\text{lb}} < \delta$, stop and return (ζ_j, μ_j); otherwise, set $j := j + 1$ and goto Step 3.

Step 3 Solve the linearized version of $Q^*(u_k)$ in (13.145):

$$O^{\text{ub}} \;:=\; \max_{\zeta \in Z, \mu \ge 0, \beta} \; u_k^{\mathsf{T}} T \mu - \beta \quad \text{s.t.} \quad W^{\mathsf{T}} \mu + q = 0,\; \beta \ge L_j(\zeta, \mu)$$

Let the optimum be (ζ_{j+1}, μ_{j+1}). Goto Step 1.

Note that the Benders cut added to the Outer algorithm in Step 1 are valid, i.e., $Q^*(u) \ge (Tu - \zeta_j)^{\mathsf{T}} \mu_j$ for all u because from (13.145)

$$Q^*(u) \;:=\; \max_{\zeta \in Z} \; \tilde{Q}^*(u, \zeta) \;\ge\; \tilde{Q}^*(u, \zeta_j) \;\ge\; (Tu - \zeta_j)^{\mathsf{T}} \mu_j, \qquad \forall u$$

where the last inequality follows because μ_j in Step 1 of the Inner algorithm is feasible for the dual linear program, i.e., $W^{\mathsf{T}} \mu_j + q = 0$.

13.7 Bibliographical notes

13.8 Problems

Chapter 13.1

Exercise 13.1 (Representation)**.** 1 Explain why $X_1 := \{x \in \mathbb{R}^n : \|x\|_1 \le 1\}$ is specified by n^2 inequalities. Show that it is the same as $X_2 := \{x \in \mathbb{R}^n : \sum_{i=1}^n y_i \le 1, -y_i \le x_i \le y_1, i = 1, \ldots, n\}$, a set specified by $2n$ variables and $2n+1$ inequalities.

2 Explain why $X_1 := \{x \in \mathbb{R}^n : \|x\|_\infty \le 1\}$ is specified by $2n$ inequalities. Show that it is the same as $X_2 := \{x \in \mathbb{R}^n : y_i \le 1, -y_i \le x_i \le y_1, i = 1, \ldots, n\}$, a set specified by $2n$ variables and $3n$ inequalities.

Exercise 13.2 (Closed and convex Z)**.** Consider the robust optimization (13.2) reproduced here:

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t.} \quad h(x, \zeta) \le 0, \ \forall \zeta \in Z \tag{13.147}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a cost function and $Z \in \mathbb{R}^L$ is an uncertainty set. Suppose, for every $x \in \mathbb{R}^n$, $h(x, \cdot)$ is convex and continuous on Z. Show that we can assume without loss of generality that Z is closed and convex. (Hint: Show that if x is a feasible solution for (13.147) then it remains feasible when Z is extended to its closure $\mathrm{cl}(Z)$ or convex hull $\mathrm{conv}(Z)$.)

The next problem shows how to formulate the robust counterpart of a nominal problem that involves equality constraints.

Exercise 13.3 (Robust SOCP relaxation of OPF)**.** Consider the second-order cone relaxation of optimal power flow (OPF) in (11.5). Recall the $(N+1) \times N$ incidence matrix C of a radial network and let $C^+ := \max\{C, 0\}$, $C^- := \min\{C, 0\}$. Let $y_1 := (p, q) \in \mathbb{R}^{2(N+1)}$ denote injections that are assumed controllable and $y_2 := (v, \ell, P, Q) \in \mathbb{R}^{4N+1}$ the resulting states. Let $r := \mathrm{diag}(r_{jk}, (j,k) \in E)$ and $x := \mathrm{diag}(x_{jk}, (j,k) \in E)$ denote the given diagonal matrices of line resistances and inductances.

1 Show that the SOCP relaxation of OPF in (11.5) takes the form:

$$\min_{y_1, y_2} \ c_1^\mathsf{T} y_1 + c_2^\mathsf{T} y_2 \tag{13.148a}$$

$$\text{s.t.} \quad A_0 y_1 + B_0 y_2 = 0, \quad B_{jk} y_2 \in K_{\text{soc}}, \ (j,k) \in E \tag{13.148b}$$

$$y_1^{\min} \le y_1 \le y_1^{\max}, \quad v^{\min} \le v \le v^{\max}, \quad \ell \le \ell^{\max} \tag{13.148c}$$

for some $(4N+1) \times (4N+1)$ matrix B_{jk} for every line $(j,k) \in E$, where $K_{\text{soc}} :=$

$\{(u,t) \in \mathbb{R}^4 : \|u\|_2 \le t\}$ is the standard second-order cone, and

$$A_0 := \begin{bmatrix} \mathbb{I}_{N+1} & 0_{N+1} \\ 0_{N+1} & \mathbb{I}_{N+1} \\ 0_N & 0_N \end{bmatrix}, \quad B_0 := \begin{bmatrix} 0_{N+1} & C^- r & -C & 0_{(N+1)\times N} \\ 0_{N+1} & C^- x & 0_{(N+1)\times N} & -C \\ C^\top & r^2 + x^2 & -2r & -2x \end{bmatrix} \tag{13.148d}$$

with $\mathbb{I}_m$ being the identity matrix of size m, and 0_m, $0_{m\times n}$ being respectively the $m \times m$ and $m \times n$ zero matrices.

2 Suppose the line resistances $r + \Delta r$ and inductances $x + \Delta x$ have uncertain perturbations of $\Delta r := \text{diag}(\Delta r_{jk}, (j,k) \in E)$ and $\Delta x := \text{diag}(\Delta x_{jk}, (j,k) \in E)$ respectively. Let the uncertain parameter $\zeta := (\Delta r, \Delta x)$ that takes value in some uncertainty set Z_ζ. Show that the robust counterpart of (13.148) is:

$$\min_{y_1, t} \; t \quad \text{s.t.} \quad c_1^\top y_1 + c_2^\top y_2 \le t, \; B_{jk} y_2 \in K_{\text{soc}}, \; (j,k) \in E, \; \forall y_2 \in Z(y_1)$$
$$y_1^{\min} \le y_1 \le y_1^{\max}, \; v^{\min} \le v \le v^{\max}, \; \ell \le \ell^{\max}, \; \forall y_2 \in Z(y_1)$$

where (derive $\Delta B(\zeta)$)

$$Z(y_1) \; := \; \{y_2 \in \mathbb{R}^{4N+1} : A_0 y_1 + (B_0 + \Delta B(\zeta)) y_2 = 0, \; \forall \zeta \in Z_\zeta\}$$

i.e., the uncertainty set Z_ζ has been embedded in the new uncertainty set $Z(y_1)$. Is the robust counterpart tractable?

Exercise 13.4 (Robust LP: $Z(x)$)**.** 1 Prove part 1 of Theorem 13.1. Show that if $Z(x) := \{\zeta \in \mathbb{R}^k : \|\zeta\|_\infty \le h(x)\}$ depends on x then the semi-infinite linear program (13.10) is equivalent to:

$$\min_{(x,y) \in \mathbb{R}^{n+k}} \; c^\top x \; \text{ s.t. } \; h(x) \sum_{l=1}^{L} y_l \; \le \; -(a_0^\top x - b_0), \quad -y_l \le a_l^\top x - b_l \le y_l, \; l = 1, \ldots, k$$

which may not be convex.

2 Prove part 2 of Theorem 13.1. Show that if $Z(x) := \{\zeta \in \mathbb{R}^k \; : \; \|\zeta\|_2 \le r(x)\}$ depends on x then the semi-infinite linear program (13.10) is equivalent to:

$$\min_{x \in \mathbb{R}^n} \; c^\top x \; \text{ s.t. } \; a_0^\top x + r(x) \sqrt{\sum_l \left(a_l^\top x - b_l\right)^2} \; \le \; b_0$$

which may not be convex.

Exercise 13.5. Recall the second order cone $K_{\text{soc}} := \{(\zeta, u) \in \mathbb{R}^{k+1} : \|\zeta\|_2 \le u\}$ and the affine set $H := \{(\zeta, u) \in \mathbb{R}^{k+1} : u = r\}$ for a given $r > 0$. Derive a tractable reformulation of the robust linear program (13.10) with the uncertainty set $Z := K_{\text{soc}} \cap H$, by adapting the proof of part 3 of Theorem 13.1. Compare your result with part 2 of Theorem 13.1.

Exercise 13.6 (Robust LP)**.** Show that the robust LP:

$$\min_{x\in\mathbb{R}^n} \quad c^\mathsf{T} x \quad \text{s.t.} \quad (a_i+u_i)^\mathsf{T} x \le b_i, \quad \forall \|u_i\|_2 \le \rho,\ i=1,\dots,m$$

where $a_i, u_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$ is equivalent to the deterministic second-order cone program:

$$\min_{x\in\mathbb{R}^n} \quad c^\mathsf{T} x \quad \text{s.t.} \quad a_i^\mathsf{T} x + \rho\|x\|_2 \le b_i,\ \ i=1,\dots,m$$

Exercise 13.7 (Robust SOCP)**.** Derive $\hat{A}(x) \in \mathbb{R}^{m\times k}$, $\hat{b}(x) \in \mathbb{R}^m$, $\hat{\alpha}(x) \in \mathbb{R}^k$, $\hat{\beta}(x) \in \mathbb{R}$ such that $x \in \mathbb{R}^n$ is feasible for the robust SOCP (13.16) if and only if

$$\left\|\hat{A}(x)\zeta + \hat{b}(x)\right\|_2 \le \hat{\alpha}^\mathsf{T}(x)\zeta + \hat{\beta}(x), \qquad \forall \zeta \in Z$$

Exercise 13.8 (Robust SOCP)**.** [151, Proposition 6.2.1] Prove Theorem 13.2, assuming the problem (13.22) is feasible and bounded.

Exercise 13.9. Prove (13.30): for any $a_1 \in \mathbb{R}^m$ and $a_2 \in \mathbb{R}^n$ we have

$$-\rho\,\|a_1\|_2\,\|a_2\|_2 \;=\; \min_{X\in\mathbb{R}^{m\times n}:\,\|X\|_2\le\rho} a_1^\mathsf{T} X a_2$$

where the spectral norm $\|X\|_2 := \sup_{\|v\|_2\le 1} \|Xv\|_2 = \sigma_{\max}(X)$ is the largest singular value of X.

Chapter 13.2.

Exercise 13.10 (Concavity)**.** Let $h : \mathbb{R}^n \to \mathbb{R}^m$ and $f : \mathbb{R}^m \to \mathbb{R}$ be real-valued functions (so their effective domains are $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively). Show that If f is concave nondecreasing and h is concave then $f(h(x))$ is concave in x.

Exercise 13.11 (α-concavity)**.** Prove Lemma 13.6.

Exercise 13.12 (Chance-constrained program)**.** Consider the dual problem (13.49b):

$$d^* \;:=\; \sup_{\mu\ge 0} d(\mu) \;:=\; \sup_{\mu\ge 0} d_X(\mu) \;+\; d_Z(\mu) \tag{13.149a}$$

where $X \subseteq \mathbb{R}^n$, $Z_p \subseteq \mathbb{R}^m$ and

$$d_X(\mu) \;:=\; \inf_{x\in X} \Big(c(x) - \mu^\mathsf{T} h(x)\Big), \qquad d_Z(\mu) \;:=\; \inf_{z\in Z_p} \mu^\mathsf{T} z \tag{13.149b}$$

Suppose X and Z_p are nonempty, convex and compact and denote the sets of minimizers in (13.149b) by

$$X(\mu) := \{x \in X : d_X(\mu) = c(x) - \mu^\mathsf{T} h(x)\}, \qquad Z(\mu) := \{z \in Z_p : d_Z(\mu) = \mu^\mathsf{T} z\}$$

1 Show that $X(\mu)$ and $Z(\mu)$ are nonempty, convex and compact. Hence $d(\mu)$ is a real-valued concave function.
2 Show that $\partial d_X(\mu) = \text{conv}\,(-h(x) : x \in X(\mu))$ for $\mu \in \mathbb{R}^m_+$.
3 Show that $\partial d_Z(\mu) = Z(\mu)$ for $\mu \in \mathbb{R}^m_+$.
4 Show that $\partial d(\mu) = \text{conv}\,(-h(x) : x \in X(\mu)) + Z(\mu)$ for $\mu \in \mathbb{R}^m_+$.

(Hint: For part 1 use Corollary 12.23 or Theorem 12.26 of Chapter 12.6. For parts 2 and 3 use Theorem 12.19 of Chapter 12.3.3. For part 4 use Theorem 12.18 of Chapter 12.3.3.)

Exercise 13.13 (Chance-constrained program)**.** Consider the dual problem and condition in Exercise 13.12. Suppose, in addition, that conditions C13.1 and C13.2 of Theorem 13.8 are satisfied, so that the set of dual optimal solutions μ^* is nonempty, convex and closed.

1 Show that $\mu^* \geq 0$ is optimal for (13.149) if and only if there exists (x^*, z^*) such that

$$x^* \in X(\mu^*), \quad z^* \in Z(\mu^*), \quad z^* - h(x^*) \in N_{\mathbb{R}^m_+}(\mu^*) \tag{13.150}$$

where $N_Y(y)$ denotes the normal cone of Y at $y \in Y$.
2 Conclude that (13.150) is equivalent to (the saddle point characterization (13.51) in Theorem 13.8)

$$x^* \in X(\mu^*), \quad z^* \in Z(\mu^*), \quad h(x^*) \geq z^*, \quad (\mu^*)^\mathsf{T}(h(x^*) - z^*) = 0$$

(Hint: For part 1 apply Theorem 12.21 to (13.149). Part 2 follows from Theorem 12.3.)

Exercise 13.14 (Log moment-generating function $\log E(e^{\lambda Y})$)**.** Show that $\psi_Y(\lambda) := \ln E(e^{\lambda Y})$ is convex in $\lambda \in \mathbb{R}$. (Hint: Use Hölder's inequality: $E|XY| \leq (E(|X|^p))^{1/p} (E(|Y|^q)^{1/q}$ for any random variables X, Y and any $p, q \in [1, \infty]$ with $1/p + 1/q = 1$.)

Exercise 13.15 (Chernoff bound: Binomial distribution)**.** Consider the Binomial random variable $Y \in \{0, \ldots, n\}$ with parameter (n, p), i.e., $\mathbb{P}(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}$ for $k \in \{0, \ldots, n\}$. Show that

1 The moment-generating function of Y is $Ee^{\lambda Y} = \left(pe^{\lambda} + 1 - p\right)^n$.
2 For any $a \in (0,1)$

$$\mathbb{P}(Y \geq na) \;\leq\; \exp\left(-n\left(a \log \frac{a}{p} + (1-a)\log \frac{1-a}{1-p}\right)\right)$$

This bound is used to bound the Binomial tail in Theorem 13.15.

Exercise 13.16 (Concentration inequality: $E(\max_i X_i)$)**.** Derive the concentration inequality (13.66) for the maximum of a finite number of sub-Gaussian random variables. (Hint: Apply Jensen's inequality to $e^{\lambda E(\max_i X_i)}$.)

Exercise 13.17 (Importance sampling)**.** We wish to estimate the tail probability $\mathbb{P}_Z(Z \geq t)$ where Z is a standard Gaussian random variable with distribution $F_Z := N(0,1)$.

1 Given N iid samples $(z_1, \ldots, z_N)$ under distribution $F_Z := N(0,1)$, what is a simple way to estimate $\mathbb{P}_Z(Z \geq t)$?
2 Suppose $t > 0$ is large so that $\mathbb{P}_Z(Z \geq t)$ is small and it will take many samples to have a reliable estimate. Suppose we obtain n iid samples $(y_i, i = 1, \ldots, n)$ from the distribution $F_Y := N(t,1)$ under which $\mathbb{P}_Y(Y \geq t) = 1/2$. Explain how to estimate $\mathbb{P}_Z(Z \geq t)$.

Chapter 13.3.

Exercise 13.18 (Violation probability)**.** Consider the following scenario program CSP(N) with N iid constraints:

$$\min_{x \in \mathbb{R}} \; -x \qquad \text{s.t.} \qquad x \leq \zeta^i, \; i = 1, \ldots, N$$

where each ζ^i takes value in $[0,1]$ with uniform distribution. Is the problem uniformly supported? Derive the distribution of the violation probability $V(x_N^*)$.

Exercise 13.19 (Latent and generalized support constraints)**.** 1 Show that if CSP(N) with $N \geq n$ has at most $s^{\max}$ support constraints then CSP(k) has at most $s^{\max}$ support constraints for all $k \geq N$.
2 Suppose CSP(N) with $N \geq n$ is uniformly supported with $1 \leq s \leq n$ support constraints. Suppose $N \geq s+2$ if $s < n$ (not fully supported case). Fix any $\zeta \in Z^N = Z^N(s)$ with I^s as its (unique) set of support constraints.

(a) Let $I^{sc} := \{i : i \notin I^s\}$ be the set of non-support constraints in ζ. Consider CSP$(N-1)$ obtained from CSP(N) by removing a constraint from I^{sc}. Show that I^s remains support constraints for CSP$(N-1)$ and $I^{sc} \setminus \{j\}$ remains non-support constraints for CSP$(N-1)$.

(b) Let CSP$(N-k)$ be the resulting scenario program by removing k constraints from I^{sc}, $1 \le k \le N-s$. Show that $x_s^* = x_{s+1}^* = \cdots = x_N^*$.

(c) Show that ζ has no latent support constraint.

3 Suppose CSP(N) is not uniformly supported and let J^t be a set of generalized support constraints for CSP(N). Fix a $\zeta \in Z^N(J^t)$.

(a) Let $J^{tc} := \{i : i \notin J^t\}$ be constraints in ζ that are not generalized support constraints (for any $0 \le t \le n$). Consider CSP$(N-1)$ obtained from CSP(N) by removing a constraint from J^{tc}. Show that J^t remains the unique set of generalized support constraints for CSP$(N-1)$ and $B \setminus \{j\}$ contains no support constraint for CSP$(N-1)$.

(b) Let CSP$(N-k)$ be the resulting scenario program by removing k constraints from J^{tc}, $1 \le k \le N-t$. Show that $x_t^* = x_{t+1}^* = \cdots = x_N^*$.

Exercise 13.20 (Latent support constraints)**.** 1 We are given three points $a, b, c \in \mathbb{R}^2$ on a plane. Each of the $N \ge 4$ iid random variables $(\zeta^1, \ldots, \zeta^N)$ is equal to a, b, c with nonzero probabilities p_a, p_b, p_c respectively with $p_a + p_b + p_c := 1$. Given $\zeta := (\zeta^1, \ldots, \zeta^N) \in Z^N$, CSP$(N)$ determines the smallest circle, specified by $x := (x_1, x_2, x_3) \in \mathbb{R}^3$, going through all N points $(\zeta^1, \ldots, \zeta^N)$.

(a) Derive the sets $Z^N(J^t)$ where J^t contain a single support constraint ζ^1 and maximum sets of latent constraints.

(b) Give a $\zeta \in Z^N(\{1\})$ with respect to which L_1 and L_2 being sets of latent support constraints imply that their union $L_1 \cup L_2$ is a set of latent support constraints.

2 Give an example where L_1 and L_2 are sets of latent support constraints wrt a certain $\zeta \in Z^N(I^s)$, but not their union $L_1 \cup L_2$. (Hint: Modify part 1 by adding a forth point $d \in \mathbb{R}^2$ to a, b, c.)

Exercise 13.21 (Sample complexity)**.** Prove Corollary 13.22.

Chapter 13.4.

Exercise 13.22 (Stochastic LP: C_2 and C_2')**.** In general, $C_2 \ne C_2'$ in (13.119).

1 For stochastic linear program with fixed recourse, provide an example where $C_2 \subsetneq C_2'$. (Hint: $E_\zeta \zeta^2 = \infty$; see Lemma 13.26.)

2 For stochastic linear program with random recourse, provide an example where $C_2' \subsetneq C_2$.

Exercise 13.23 (Stochastic LP: C_2 and C_2')**.** Give an example random variable that has finite first moment but infinite second moment.

Exercise 13.24 ($\partial_x \tilde{Q}(x,\zeta)$)**.** Fix any ζ and recall from (13.118a) (omitting ω or ζ in notation)

$$\tilde{Q}(x,\zeta) := \min_{y \geq 0} q^\mathsf{T} y \quad \text{s.t.} \quad Wy = h - Tx$$

Suppose, for each $x \in \mathbb{R}^{n_1}$,

1 There exists a unique primal-dual optimal solution $(y(x), \lambda(x), \mu(x))$ for $\tilde{Q}(x,\zeta)$; moreover it is continuous at x.
2 Strong duality holds at $(y(x), \lambda(x), \mu(x))$.

Show that $\tilde{Q}(x,\zeta)$ is continuously differentiable and $\nabla_x \tilde{Q}(x,\zeta) = T^\mathsf{T}\lambda(x)$. (Hint: Use envelop theorem (Chapter 8.3.6). See [142, Proposition 2.2, p.28] on subdifferentiability of $\tilde{Q}(\bar{x},\zeta)$ when $(y(x), \lambda(x), \mu(x))$ is not unique and continuous in x.)

Exercise 13.25. [143, Theorem 34 and 35; p.157][Stochastic nonlinear program] Prove Lemma 13.29.

Exercise 13.26. [143, Theorem 39; p.158][Stochastic nonlinear program] Prove Theorem 13.30.

Chapter 13.5

Exercise 13.27. Derive (13.140) when all generators have a common and finite capacity $0 < p^{\max} < \infty$.

Part III

Unbalanced three-phase networks

14 Component models, I: devices

Single-phase models are a good approximation of the reality for many transmission network applications where lines are symmetric and loads are balanced. In that case, a similarity transformation produces three networks in a sequence coordinate, called zero, positive, and negative-sequence networks, that are decoupled. Each network can be analyzed using a single-phase model studied in previous chapters. These sequence networks are coupled when lines are not transposed or equally spaced, e.g., as in distribution systems, or when loads are unbalanced or nonlinear, e.g., AC furnaces, high-speed trains, power electronics, or single or two-phase laterals in distribution networks. In that case single-phase analysis can produce incorrect power flow solutions. In this and next chapters we extend single-phase models to unbalanced three-phase models.

We first provide in Chapter 14.1 an overview of models for three-phase devices, lines and transformers, and how to use these component models to compose an overall network model. We summarize in Chapter 14.2 mathematical properties that underly the behavior of three-phase systems. Finally we derive in Chapter 14.3 the models of three-phase voltage sources, current sources, power sources, and impedances in Y and Δ configurations. In Chapter 15 we derive models for three-phase lines and transformers. We will use these component models in Chapters 16 and 17 to construct network models and study unbalanced three-phase analysis.

14.1 Overview

Figure 14.1 shows a simple example of a three-phase system with three components, two devices connected by a line. For example the single-terminal device on the left can model a three-phase generator and the other single-terminal device can be a three-phase load. Each terminal has three wires (or ports or conductors) indexed by its phases a, b, c, and possibly a neutral wire indexed by n. Internally, it can be in Y or Δ configuration, and the Y configuration may have a neutral wire that may be grounded. A three-phase line has two terminals, each terminal with three or four wires, and it connects two single-terminal devices, one at each end of the line. The line may model a transmission or distribution line or a transformer. The distribution line can be

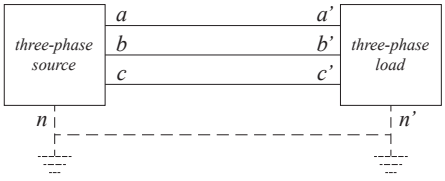
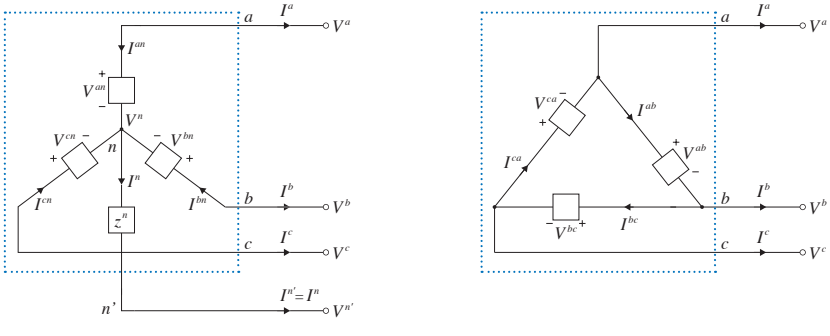


Figure 14.1 A simple model of a three-phase system consisting of a source connected through a line to a load.

underground or overhead with a neutral wire that may be grounded in regular spacing along the line.

The basic idea in modeling a three-phase component is to explicitly separate its model into an *internal model* that specifies the characteristics of the constituent single-phase components in terms of internal variables, and a *conversion rule* that maps its internal variables to its terminal variables. The internal model depends only on the type of components (non-ideal voltage sources, ZIP loads, or different single-phase transformer models) regardless of their configurations. The conversion rule depends only on their configurations regardless of the type of components. They determines an *external model* which is a relation between the terminal variables, obtained by eliminating the internal variables from the set of equations describing the internal model and the conversion rule. We next describe this procedure in detail.

14.1.1 Internal and terminal variables



(a) Y configuration (b) Δ configuration

Figure 14.2 Internal and external variables associated with a single-terminal device in Y and Δ configurations.

The *internal variables* of a generic single-terminal device are shown in Figure 14.2 and defined as follows:

- $V^Y := \left(V^{an}, V^{bn}, V^{cn}\right) \in \mathbb{C}^3$, $I^Y := \left(I^{an}, I^{bn}, I^{cn}\right) \in \mathbb{C}^3$, $s^Y := \left(s^{an}, s^{bn}, s^{cn}\right) \in \mathbb{C}^3$, $\left(V^n, I^n, s^n\right) \in \mathbb{C}^3$: line-to-neutral voltages, currents, and power across the single-phase devices in Y configuration, as well as the voltage, current, and power across the neutral impedance z^n, respectively. By definition $s^{an} := V^{an}\left(I^{an}\right)^{\mathsf{H}}$ is the power across the phase-a device, etc. The neutral voltage V^n, with respect to a common reference point, is generally nonzero. A Y-configured device may or may not have a neutral line which may or may not be grounded and the grounding impedance z^n may or may not be zero. When present, the current on the neutral line is denoted by I^n in the direction coming out of the device. The Kirchhoff current law dictates that $I^n = \sum_\phi I^{\phi n}$. The internal power across the neutral impedance is $s^n := \left(V^n - V^{n'}\right)\overline{I}^n$ where $\overline{I}^n$ denotes the complex conjugate of I^n. The term $V^n\overline{I}^Y$, in contrast, is the vector power delivered across the neutral and the common reference point (e.g., the ground).
- $V^\Delta := \left(V^{ab}, V^{bc}, V^{ca}\right) \in \mathbb{C}^3$, $I^\Delta := \left(I^{ab}, I^{bc}, I^{ca}\right) \in \mathbb{C}^3$, $s^\Delta := \left(s^{ab}, s^{bc}, s^{ca}\right) \in \mathbb{C}^3$: line-to-line voltages, currents, and power across the single-phase devices respectively in Δ configuration. By definition $s^{ab} := V^{ab}\left(I^{ab}\right)^{\mathsf{H}}$ is the power across the phase-a device, etc.

Note that the direction of the internal power s^{an} or s^{ab} across a single-phase device is defined in the direction of the current across the device. The neutral line, when present, is often assumed grounded, i.e., $V^{n'} = 0$, and the voltage reference point is the ground. In this case $s^n = V^n I^{n\mathsf{H}}$.

The *terminal variables* of the single-terminal device in Figure 14.2 are defined as follows:

- $V := \left(V^a, V^b, V^c\right) \in \mathbb{C}^3$, $I := \left(I^a, I^b, I^c\right) \in \mathbb{C}^3$, $s := \left(s^a, s^b, s^c\right) \in \mathbb{C}^3$, $\left(V^{n'}, I^{n'}, s^{n'}\right) \in \mathbb{C}^3$: terminal voltages, currents, and power respectively. The terminal voltage V is defined with respect to an arbitrary but common reference point, e.g., the ground. The terminal current I is defined in the direction coming out of the device, i.e., I is defined to be the current injection from the device to the rest of the network when it is connected to a bus bar, regardless of whether it generates or consumes power. By definition $s^a := V^a\left(I^a\right)^{\mathsf{H}}$ is the power across the terminal a and the common reference point. When there is a neutral wire its terminal voltage (with respect to the common reference point), current and power are denoted by $\left(V^{n'}, I^{n'}, s^{n'}\right)$ with $I^{n'} = I^n$ and $s^{n'} := V^{n'} I^{n'\mathsf{H}} = V^{n'} I^{n\mathsf{H}}$.

The internal and external variables of a three-phase device are summarized in Table 14.1.

	Voltage	Current	Power	Neutral line
Internal variable	$V^{Y/\Delta}$	$I^{Y/\Delta}$	$s^{Y/\Delta}$	(V^n, I^n, s^n)
External variable	V	I	s	$\left(V^{n'}, I^{n'}, s^{n'}\right)$

Table 14.1 Internal and external variables of single-terminal three-phase devices. The notation $x^{Y/\Delta}$ is a shorthand for the pair (x^Y, x^Δ).

14.1.2 Three-phase device models

An *internal model* of a three-phase device is a relation between the internal variables $\left(V^Y, I^Y, s^Y\right)$ or between $\left(V^\Delta, I^\Delta, s^\Delta\right)$. It describes the behavior of the single-phase devices, and does *not* depend on their Y or Δ configuration nor the absence or presence of a neutral line. For example the internal model of an ideal voltage source specified by its internal voltage $E^{Y/\Delta} \in \mathbb{C}^3$ is

$$V^{Y/\Delta} = E^{Y/\Delta}, \qquad s^{Y/\Delta} = \operatorname{diag}\left(E^{Y/\Delta}\left(I^{Y/\Delta}\right)^{\mathsf{H}}\right)$$

where the notation $x^{Y/\Delta}$ is a shorthand for the pair (x^Y, x^Δ). The internal model of an impedance specified by a complex matrix $z^{Y/\Delta} \in \mathbb{C}^{3\times 3}$ is

$$V^{Y/\Delta} = z^{Y/\Delta} I^{Y/\Delta}, \qquad s^{Y/\Delta} = \operatorname{diag}\left(V^{Y/\Delta}\left(I^{Y/\Delta}\right)^{\mathsf{H}}\right)$$

Denote the internal model of a general three-phase device by

$$f^{\text{int}}\left(V^{Y/\Delta}, I^{Y/\Delta}\right) = 0, \qquad s^{Y/\Delta} = \operatorname{diag}\left(V^{Y/\Delta} I^{Y/\Delta \mathsf{H}}\right) \tag{14.1}$$

The *external model* of a device is the relation between its terminal variables (V, I, s) and possibly $\left(V^{n'}, I^{n'}, s^{n'}\right)$. It describes the externally observable behavior of the device and depends on both the internal model of the single-phase devices and their configuration. How the Y or Δ configuration determines its external model is described by conversion rules that map internal variables to terminal variables. While the internal model depends only on the type of *single-phase* devices, the conversion rules depend only on the configuration, but not on the device type. This will be explained in detailed in Chapter 14.3. Denote the external model by

$$f^{\text{ext}}(V, I) = 0, \qquad s = \operatorname{diag}\left(V I^{\mathsf{H}}\right) \tag{14.2}$$

The importance of the external model is that devices interact over a network only through their terminal variables. The external model of each three-phase device imposes local constraints on its own terminal variables while network equations, to be studied in Chapters 16 and 17, impose global constraints on the terminal variables across devices.

Though not explicit, the functions in (14.1) and (14.2) may be augmented with the

internal and terminal variables (V^n, I^n, s^n) and $\left(V^{n'}, I^{n'}, s^{n'}\right)$ respectively associated with the neutral in a Y configuration. The functions f^{int} and f^{ext} are linear for voltage sources, current sources and impedances, but quadratic for power sources; see Chapter 14.3.

A three-phase device can therefore be modeled in two equivalent ways:

1 An internal model (14.1) that describes the relation between its internal variables $\left(V^{Y/\Delta}, I^{Y/\Delta}, s^{Y/\Delta}\right)$ and the conversion rules, (14.8) (14.9) (14.10) below, that map internal variables to external variables.
2 An external model (14.2) that describes the relation between its terminal variables. The external model is obtained by applying the conversion rules to the internal model (14.1) to eliminate the internal variables.

The first model is useful when the application under study needs to determine or optimize some of the internal variables such as the power $s_j^{Y/\Delta}$ generated or consumed by each of the single-phase devices connected at a bus j. Otherwise the external model (14.2) can be used if the application involves only the terminal variables.

Remark 14.1. One should be careful with the direction in which currents and powers are defined when relating internal and external powers (see Chapter14.3). For instance V^{an} is the voltage drop between terminal a and the neutral n and I^{an} is the current from a to n. The power s^{an} is therefore the power delivered to the device in the direction of the current I^{an}. If the device models a generator then the power it generates is $-s^{an} = V^{an}\left(-I^{an}\right)^{\mathsf{H}}$.

14.1.3 Three-phase line and transformer models

Let the terminals of a three-phase line or transformer be indexed by j and k. Let $V_j := \left(V_j^a, V_j^b, V_j^c\right) \in \mathbb{C}^3$ and $V_k := \left(V_k^a, V_k^b, V_k^c\right) \in \mathbb{C}^3$ denote the voltages at terminals j and k respectively with respect to an arbitrary but common reference point. Let $I_{jk} := \left(I_{jk}^a, I_{jk}^b, I_{jk}^c\right) \in \mathbb{C}^3$ denote the sending-end current from terminal j to terminal k along the line or transformer, and I_{kj} denote the sending-end current in the opposite direction. The external behavior of a three-phase line or transformer is described by a linear relation between $\left(V_j, V_k, I_{jk}, I_{kj}\right) \in \mathbb{C}^{12}$ of the form

$$g\left(V_j, V_k, I_{jk}, I_{kj}\right) = 0 \tag{14.3a}$$

where g is defined by 3×3 matrix parameters of the line (j,k).

Let $S_{jk} := \left(S_{jk}^a, S_{jk}^b, S_{jk}^c\right) \in \mathbb{C}^3$ denote the sending-end power from terminal j to terminal k along the line or transformer, and S_{kj} denote the sending-end power in the

opposite direction. For each phase $\phi = a, b, c$, $S_{jk}^{\phi} := V_j^{\phi} \left(I_{jk}^{\phi}\right)^{\mathsf{H}}$. In vector form this is

$$S_{jk} := \operatorname{diag}\left(V_j I_{jk}^{\mathsf{H}}\right), \qquad S_{kj} := \operatorname{diag}\left(V_k I_{kj}^{\mathsf{H}}\right) \tag{14.3b}$$

When there is a neutral wire between terminals j and k, their voltages are V_j^n and V_k^n. The current in the neutral wire is denoted by $\left(I_{jg}^n, I_{kg}^n\right)$ if the neutral is grounded or $\left(I_{jk}^n, I_{kj}^n\right)$ otherwise. The function g (14.3a) includes neutral voltages and currents and is defined by 4×4 matrix parameters of the line. The power flow equation (14.3b) is modified accordingly.

The equations (14.3) describe the end-to-end behavior of a three-phase line or transformer. We reiterate that they depend on the three-phase devices connected to its terminals only through their external variables.

14.1.4 Three-phase network models

A network of three-phase devices connected by three-phase lines and transformers can be composed from the component models (14.2) and (14.3) for these components through the flow balance equations that relate nodal current and power (s_j, I_j) to line currents and power $\left(I_{jk}, S_{jk}\right)$ connected to the same bus bar j:

$$I_j = \sum_{k:j\sim k} I_{jk}, \qquad \forall j \tag{14.4a}$$

$$s_j = \sum_{k:j\sim k} S_{jk}, \qquad \forall j \tag{14.4b}$$

Depending on the application, what information is available and what quantities are controllable, we can model the network in two ways:

1 *IV model:* We can model the network using the relation $f^{\text{ext}}(V, I) = 0$ in (14.2) and (14.3a) (14.4a) between bus voltage and current vectors (V, I). This model is linear. Once nodal voltages $V_j \in \mathbb{C}^3$ and currents $I_j \in \mathbb{C}^3$ are determined, nodal powers $s_j := \operatorname{diag}\left(V_j I_j^{\mathsf{H}}\right)$ can be calculated.
2 *sV model:* We can model the network using the device model (14.2) and the power flow equations (14.3b) (14.4b) between bus voltages and power injections (V, s). This model is generally nonlinear.

The linear *IV* model can always be used if the system contains no power sources. Otherwise either the *IV* model or the *sV* model can be used to describe the network but, since the device model (14.2) is nonlinear, the overall model will always be

nonlinear. Network models are studied in Chapter 16 for bus injection models and Chapter 17 for branch flow models.

In summary a complete network model consists of

1 (14.3) (14.4) + (14.1) and (14.8) (14.9) (14.10): involves the internal variables of three-phase devices.
2 (14.3) (14.4) + (14.2): does not involve internal variables of the three-phase devices.

14.1.5 Balanced operation

If the following conditions are satisfied throughout the network:

1 all lines have symmetric geometry;
2 zero total current: $i^a(t)+i^b(t)+i^c(t)=0$ at all times t;
3 zero total charge: $q^a(t)+q^b(t)+q^c(t)=0$ at all times t;

then the system is balanced and its phases are decoupled. This means that (14.2) reduces to

$$f^{\text{ext},\phi}(V^\phi, I^\phi) = 0, \qquad s^\phi = V^\phi I^{\phi\mathsf{H}}, \qquad \phi = a,b,c$$

and similarly for equations (14.3)(14.4). For example the line current I^a_{jk} in phase a depends only on voltages $\left(V^a_j, V^a_k\right)$ in phase a, but not on voltages in other phases. This allows per-phase analysis, as we have done in earlier chapters. These decoupling conditions can be satisfied if the terminal voltages of all three-phase sources are balanced (i.e., they have equal magnitudes and are separated by 120° in phase), all three-phase loads consist of identical impedances, and all three-phase lines has symmetric geometry (e.g. through transposition). In that case the magnetic coupling across phases can be modeled by self-impedance alone, i.e., a three-phase line behaves *as if* its mutual inductances and capacitances across phases are zero and self inductances and capacitances are equal in each phase, as shown in Chapter 2.1.4. A general formulation of per-phase analysis of a balanced network and its formal justification is provided in Chapter 16.3. The underlying mathematical property is explained in Corollary 1.3 and Theorem 14.2.

Otherwise, self-impedance alone is not sufficient to model the coupling across phases of a line and per-phase analysis becomes inaccurate. A unbalanced three-phase model is necessary for power flow analysis. The overview of such a model is illustrated in Figure 14.3.

Before deriving in detail the internal and external models of these components we first describe some mathematical tools that are important for our derivation.

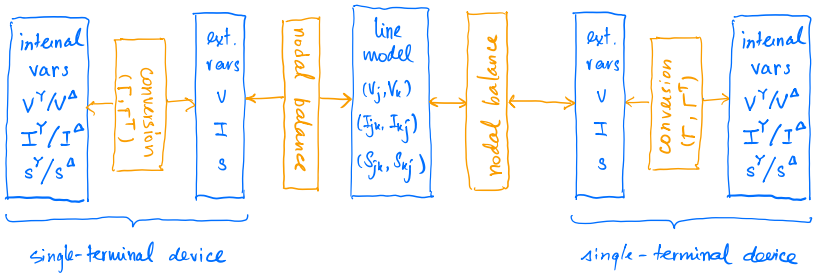


Figure 14.3 Overall network model of the system in Figure 14.1.

14.2 Mathematical properties of three-phase network

In this section we collect several mathematical properties that are used in the rest of this chapter, often without explicit references. These properties underlie much of the behavior of three-phase systems. Specifically we use the spectral properties of the conversion matrices Γ and Γ^{T} defined in Chapter 1.2.2 to derive in Chapter 14.2.1 their pseudo inverses. The eigenvectors of Γ are orthogonal and can serve as a basis of $\mathbb{C}^3$. In Chapter 14.2.2 we use this basis to transform voltages and currents to a sequence coordinate in which an unbalanced network may become decoupled.

14.2.1 Pseudo-inverses of $\Gamma, \Gamma^{\mathsf{T}}$.

The main characters of three-phase networks arise from the spectral properties of the conversion matrices Γ and Γ^{T}, defined in (1.12) of Chapter 1.2.2 and reproduced here:

$$\Gamma := \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{bmatrix}, \qquad \Gamma^{\mathsf{T}} := \begin{bmatrix} 1 & 0 & -1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \tag{14.5}$$

We have seen in Chapter 1.2.4 that these conversion matrices play an important role in relating the internal and external behaviors of a balanced three-phase system. In such a system, positive-sequence voltages and currents are in $\mathrm{span}(\alpha_+)$ and α_+ is an eigenvector of Γ and Γ^{T}. This means that the transformation of balanced voltages and currents under $\Gamma, \Gamma^{\mathsf{T}}$ reduces to a scaling of these variables by their eigenvalues $1-\alpha$ and $1-\alpha^2$ respectively (Corollary 1.3). The voltage and current at every point in a network can be written as linear combinations of transformed source voltages and source currents, transformed by (Γ, Γ^T) and line admittance matrices. Therefore if the source voltages and source currents are balanced positive-sequence sets and lines are identical and phase-decoupled, then the transformed voltages and currents remain in $\mathrm{span}(\alpha_+)$ and hence are balanced positive-sequence sets. This is the key property that enables balanced sources to induce balanced voltages and currents throughout a balanced network, allowing per-phase analysis of three-phase systems. A formal

statement and proof of this property for general three-phase networks is provided in Chapter 16.3.

For unbalanced systems where voltages and currents are not necessarily in span(α_+), Corollary 1.3 is not applicable and we need the concept of pseudo inverses of $\Gamma, \Gamma^{\mathsf{T}}$ in order to convert between terminal variables and line-to-line variables internal to a Δ configuration. Even though Γ and Γ^{T} are not invertible, their pseudo inverses $\Gamma^{\dagger}$ and $\Gamma^{\mathsf{T}\dagger}$ respectively always exist. The pseudo inverse $M^{\dagger}$ of a matrix $M \in \mathbb{C}^{n\times n}$ maps the null space of M^{H} to zero. The orthogonal complement of the null space of M^{H} is the range space of M. $M^{\dagger}$ restricted to the range space acts like an inverse of M in that it maps each vector v in the range space of M to the unique vector $u := M^{\dagger}v$ in the range space of M^{H}. The vector u is the one in $\mathbb{C}^n$ with the minimum norm such that $Mu = v$. See Appendix A.7 for more properties of pseudo-inverse. The facts relevant to us is summarized in the following lemma (from Theorem A.13, Theorem A.19 and Remark A.2.)

Lemma 14.1. Let $M \in \mathbb{C}^{n\times n}$ be a normal matrix, i.e., $MM^{\mathsf{H}} = M^{\mathsf{H}}M$.

1 *Unitary diagonalization*. There exists a unitary matrix $U \in \mathbb{C}^{n\times n}$ and a diagonal matrix $\Lambda \in \mathbb{C}^{n\times n}$ with

$$M = U\Lambda U^{\mathsf{H}} \;=\; \sum_{i=1}^{n} \lambda_i u_i u_i^{\mathsf{H}}$$

where
   1 Λ =diag$(\lambda_1, \ldots, \lambda_n)$ consists of the eigenvalues of A;
   2 the columns of U are the associated eigenvectors of A.

2 *Pseudo inverse*. The pseudo-inverse of M is given by $M^{\dagger} = U\Lambda^{\dagger}U^{\mathsf{H}}$ where $\Lambda^{\dagger} :=$ diag $\left(\lambda_1^{-1}, \ldots, \lambda_n^{-1}\right)$ with $\lambda_j^{-1} := 0$ if $\lambda_j = 0$.

3 Consider $Mx = b$. A solution x exists if and only if b is orthogonal to null$\left(M^{\mathsf{H}}\right)$ in which case

$$x = M^{\dagger} b \;+\; w, \qquad\qquad w \;\in\; \text{null}\,(M)$$

Moreover $M^{\dagger}b$ is the unique solution to $Mx = y$ with the minimum Euclidean norm $\|x\|_2 = \|M^{\dagger}b\|_2 + \|w\|_2$, $w \in \text{null}(M)$.

Theorem 1.2 shows that Γ and Γ^{T} are normal matrices and their spectral decompositions are

$$\Gamma \;=\; F\Lambda\overline{F}, \qquad\qquad \Gamma^{\mathsf{T}} \;=\; \overline{F}\Lambda F \tag{14.6a}$$

where Λ is a diagonal matrix and F is a unitary matrix defined in (1.18), reproduced here:

$$\Lambda \;:=\; \begin{bmatrix} 0 & & \\ & 1-\alpha & \\ & & 1-\alpha^2 \end{bmatrix}, \quad F \;:=\; \frac{1}{\sqrt{3}}\begin{bmatrix} \mathbf{1} & \alpha_+ & \alpha_- \end{bmatrix} \;:=\; \frac{1}{\sqrt{3}}\begin{bmatrix} 1 & 1 & 1 \\ 1 & \alpha & \alpha^2 \\ 1 & \alpha^2 & \alpha \end{bmatrix} \tag{14.6b}$$

with $\alpha := e^{-\mathrm{i}2\pi/3}$ and α_+ and α_- being the standard positive and negative sequence vectors respectively:

$$\alpha_+ := \begin{bmatrix} 1 \\ \alpha \\ \alpha^2 \end{bmatrix}, \qquad \alpha_- := \begin{bmatrix} 1 \\ \alpha^2 \\ \alpha \end{bmatrix}$$

Here $\overline{F}$ is the complex conjugate of F componentwise. Since F is symmetric (Theorem 1.2), Lemma 14.1 implies that the pseudo inverses of $\Gamma, \Gamma^{\mathsf{T}}$ are

$$\Gamma^{\dagger} = F\Lambda^{\dagger}\overline{F}, \qquad \Gamma^{\mathsf{T}\dagger} = \overline{F}\Lambda^{\dagger}F \tag{14.6c}$$

where $\Lambda^{\dagger} := \operatorname{diag}\left(0, (1-\alpha)^{-1}, (1-\alpha^2)^{-1}\right)$. This yields the following simple expressions for these pseudo inverses. The proof of the theorem is left as Exercise 14.1.

Theorem 14.2 (Pseudo inverses of $\Gamma, \Gamma^{\mathsf{T}}$)**.** 1 The null spaces of Γ and Γ^{T} are both $\operatorname{span}(1,1,1)$.

2 Their pseudo-inverses are

$$\Gamma^{\dagger} = \frac{1}{3}\Gamma^{\mathsf{T}}, \qquad \Gamma^{\mathsf{T}\dagger} = \frac{1}{3}\Gamma$$

3 Consider $\Gamma x = b$ where $b, x \in \mathbb{C}^3$. Solutions x exist if and only if $\mathbf{1}^{\mathsf{T}} b = 0$, in which case the solutions x are given by

$$x = \frac{1}{3}\Gamma^{\mathsf{T}} b \; + \; \gamma \mathbf{1}, \qquad \gamma \in \mathbb{C}$$

4 Consider $\Gamma^{\mathsf{T}} x = b$ where $b, x \in \mathbb{C}^3$. Solutions x exist if and only if $\mathbf{1}^{\mathsf{T}} b = 0$, in which case the solutions x are given by

$$x = \frac{1}{3}\Gamma b \; + \; \gamma \mathbf{1}, \qquad \gamma \in \mathbb{C}$$

5 $\Gamma\Gamma^{\dagger} = \Gamma^{\dagger}\Gamma = \frac{1}{3}\Gamma\Gamma^{\mathsf{T}} = \frac{1}{3}\Gamma^{\mathsf{T}}\Gamma = \mathbb{I} - \frac{1}{3}\mathbf{1}\mathbf{1}^{\mathsf{T}}$ where $\mathbb{I}$ is the identity matrix of size 3.

Recall that $\Gamma\Gamma^{\mathsf{T}} = \Gamma^{\mathsf{T}}\Gamma$ are complex symmetric Laplacian matrices of the graphs in Figure 1.9. This theorem underlies much of the materials in this chapter.

14.2.2 Similarity transformation and symmetrical components

Fortescue transformation.

Since Γ and Γ^{T} are normal matrices, they have orthonormal eigenvectors $(\mathbf{1}, \alpha_+, \alpha_-)$ which are the columns of F defined in (14.6b). We can therefore use F to define a similarity transformation (see Appendix A.4 for discussions on similarity transformation). This idea is due to Fortescue [152] and F is sometimes called a (normalized) Fortescue matrix. It simplifies the analysis of an unbalanced three-phase system when the network has a certain symmetry, as explained in Chapter 16.4.

Consider a vector x that may represent a voltage or current. Recall that F is unitary and complex symmetric (Theorem 1.2) and therefore its inverse is:

$$F^{-1} = F^{\mathsf{H}} = \overline{F} = \frac{1}{\sqrt{3}}\begin{bmatrix}\mathbf{1} & \overline{\alpha}_+ & \overline{\alpha}_-\end{bmatrix} = \frac{1}{\sqrt{3}}\begin{bmatrix}\mathbf{1}^{\mathsf{T}} \\ \alpha_-^{\mathsf{T}} \\ \alpha_+^{\mathsf{T}}\end{bmatrix} \tag{14.7}$$

(Note that $\overline{\alpha}_+ = \alpha_-$, $\overline{\alpha}_- = \alpha_+$; more properties of α are studied in Exercise 1.6). The matrix F defines the transformation:

$$x = F\tilde{x}, \qquad \tilde{x} := F^{-1}x = \overline{F}x$$

The vector $\tilde{x}$ is called the *sequence variable* of x. Its components

$$\tilde{x}_0 := \frac{1}{\sqrt{3}}\mathbf{1}^{\mathsf{H}}x, \qquad \tilde{x}_+ := \frac{1}{\sqrt{3}}\alpha_+^{\mathsf{H}}x, \qquad \tilde{x}_- := \frac{1}{\sqrt{3}}\alpha_-^{\mathsf{H}}x$$

are called the *zero-sequence, positive-sequence*, and *negative-sequence* components of x. They are also called *symmetrical components* of x. We will sometimes refer to x as a *phase variable* to differentiate it from the sequence variable $\tilde{x}$. The relation $x = F\tilde{x}$ expresses the phase variable in terms of its sequence components:

$$x = \frac{1}{\sqrt{3}}(\tilde{x}_0\mathbf{1} + \tilde{x}_+\alpha_+ + \tilde{x}_-\alpha_-) = \frac{1}{3}\left(\left(\mathbf{1}^{\mathsf{H}}x\right)\mathbf{1} + \left(\alpha_+^{\mathsf{H}}x\right)\alpha_+ + \left(\alpha_-^{\mathsf{H}}x\right)\alpha_-\right)$$

Sequence voltage, current, power.

Applying this similarity transformation to phase voltage V and current I, we obtain their sequence variables:

$$\tilde{V} = \overline{F}V, \qquad \tilde{I} = \overline{F}I,$$

The vector of power in the phase coordinate is $s := \operatorname{diag}\left(VI^{\mathsf{H}}\right)$ and that in the sequence coordinate is $\tilde{s} := \operatorname{diag}\left(\tilde{V}\tilde{I}^{\mathsf{H}}\right)$. They are related through the outer product of voltage and current in their respective coordinates according to:

$$\begin{aligned}\tilde{s} &:= \operatorname{diag}\left(\tilde{V}\tilde{I}^{\mathsf{H}}\right) = \operatorname{diag}\left(\overline{F}VI^{\mathsf{H}}\overline{F}^{\mathsf{H}}\right) = \operatorname{diag}\left(\overline{F}VI^{\mathsf{H}}F\right)\\ s &:= \operatorname{diag}\left(VI^{\mathsf{H}}\right) = \operatorname{diag}\left(F\tilde{V}\tilde{I}^{\mathsf{H}}F^{\mathsf{H}}\right) = \operatorname{diag}\left(F\tilde{V}\tilde{I}^{\mathsf{H}}\overline{F}\right)\end{aligned}$$

The total powers $\mathbf{1}^{\mathsf{T}}\tilde{s} = \mathbf{1}^{\mathsf{T}}s$ however are equal in both coordinates:

$$\mathbf{1}^{\mathsf{T}}\tilde{s} = \tilde{I}^{\mathsf{H}}\tilde{V} = \left(I^{\mathsf{H}}\overline{F}^{\mathsf{H}}\right)\left(\overline{F}V\right) = I^{\mathsf{H}}V = \mathbf{1}^{\mathsf{T}}s$$

since $\overline{F}^{\mathsf{H}}\overline{F} = F\overline{F} = \mathbb{I}$. This is sometimes referred to as power invariance property of the similarity transformation F. In Chapter 16.4 we will apply sequence variables to the external models of Chapter 14.3 to define sequence networks.

In Definition 1.1, we call x a balanced vector if its zero-sequence component $\tilde{x}_0 = 0$ and exactly one of $\tilde{x}_+$ and $\tilde{x}_-$ is nonzero. In particular a balanced positive-sequence vector is in $\operatorname{span}(\alpha_+)$. To simplify exposition in this chapter it is convenient to generalize the definition of balanced vector to include a zero-sequence component.

Definition 14.1 (Generalized balanced vector)**.** A vector $\hat{x} := (\hat{x}_1, \hat{x}_2, \hat{x}_3) \in \mathbb{C}^3$ is called a *generalized balanced vector* if $\hat{x} = x + \gamma \mathbf{1}$, for some $\gamma \in \mathbb{C}$, such that x is balanced according to Definition 1.1.

Hence a generalized balanced vector $\hat{x}$ may contain a nontrivial zero-sequence component $\tilde{x}_0$ and exactly one of $\tilde{x}_+$ and $\tilde{x}_-$. We will often refer to a generalized balanced vector $\hat{x}$ simply as *balanced* if there is no risk of confusion or if the differentiation is not important, even if $\gamma \neq 0$. The key property Corollary 1.3 for balanced networks holds for generalized balanced vectors, i.e., $\Gamma(x + \gamma \mathbf{1}) = (1-\alpha)x$ and $\Gamma^{\mathsf{T}}(x + \gamma \mathbf{1}) = (1-\alpha^2)x$ if x is a balanced positive-sequence vector.

Park transformation.

Besides Foretescue transformation F, several other similarity transformations have been proposed that have different advantages and disadvantages for steady-state fault analysis; see [153] that explains their relation. Park's transformation [154] is applicable not only to steady-state voltage and current phasors, but also to instantaneous voltages, currents, and flux linkages. It is originally proposed for analyzing synchronous machines and is defined by the following real orthonormal matrix (which is the normalized version of Park's original matrix; we follow [1]):

$$P := \sqrt{\frac{2}{3}} \begin{bmatrix} \frac{1}{\sqrt{2}} & \cos\theta & \sin\theta \\ \frac{1}{\sqrt{2}} & \cos(\theta - 120^\circ) & \sin(\theta - 120^\circ) \\ \frac{1}{\sqrt{2}} & \cos(\theta + 120^\circ) & \sin(\theta + 120^\circ) \end{bmatrix}$$

It can be verified that P is orthonormal so that $P^{-1} = P^{\mathsf{T}}$. The matrix can be used to transform instantaneous phase voltages, currents and flux linkages. For example, for instantaneous voltages we have

$$v = \begin{bmatrix} v^a \\ v^b \\ v^c \end{bmatrix} = \sqrt{\frac{2}{3}} \begin{bmatrix} \frac{1}{\sqrt{2}} & \cos\theta & \sin\theta \\ \frac{1}{\sqrt{2}} & \cos(\theta - 120^\circ) & \sin(\theta - 120^\circ) \\ \frac{1}{\sqrt{2}} & \cos(\theta + 120^\circ) & \sin(\theta + 120^\circ) \end{bmatrix} \begin{bmatrix} v^0 \\ v^d \\ v^q \end{bmatrix} = P\tilde{v}$$

$$\tilde{v} = \begin{bmatrix} v^0 \\ v^d \\ v^q \end{bmatrix} = \sqrt{\frac{2}{3}} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \cos\theta & \cos(\theta - 120^\circ) & \cos(\theta + 120^\circ) \\ \sin\theta & \sin(\theta - 120^\circ) & \sin(\theta + 120^\circ) \end{bmatrix} \begin{bmatrix} v^a \\ v^b \\ v^c \end{bmatrix} = P^{\mathsf{T}} v$$

The transformed coordinate is called the $0dq$, or zero-direct-quadrature, or rotor coordinate. The abc variables are stator-based quantities and the $0dq$ variables are rotor-based quantities. Similarly we can transform abc currents and flux linkages into the $0dq$ coordinate with $\tilde{i} = P^{\mathsf{T}} i$ and $\tilde{\lambda} = P^{\mathsf{T}} \lambda$. The model of a synchronous machine becomes simpler in the rotor coordinate. For example the inductance matrix L in the abc coordinate that relates currents and flux linkages, $\lambda = Li$, becomes diagonal in the rotor coordinate, i.e., $\tilde{\lambda} = \tilde{L}\tilde{i}$ for a diagonal $\tilde{L}$.

14.3 Three-phase device models

In this section we develop the external models (14.2)(14.3) of three-phase devices in terms of their internal specifications. The models of three-phase devices developed in Chapter 1.2 and the phase-decoupled line model of Chapter 2 are special cases of the models in this section.

We start by describing in Chapter 14.3.1 the conversion rules (14.8) and (14.9)(14.10) that maps internal variables $\left(V^{Y/\Delta}, I^{Y/\Delta}, s^{Y/\Delta}\right)$ to external variables (V, I, s) for devices in Y and Δ configurations respectively. These conversion rules depend only on the configuration and are applicable to any types of devices. In Chapters 14.3.3 and 14.3.4 we present the internal models of four types of devices in Y and Δ configuration respectively and apply the conversion rules to these internal models to derive their external models. In Chapter 14.3.5 we explain how to derive the Y equivalent of an ideal Δ-configured voltage or current source in an unbalanced setting.

14.3.1 Conversion rules

Conversion in Y configuration.

Consider a generic three-phase device in Y configuration with internal and terminal variables defined as in Figure 14.2(a). Its terminal voltage, current, and power (V, I, s) are related to its internal variables (V^Y, I^Y, s^Y) by:

$$V = V^Y + V^n \mathbf{1}, \qquad I = -I^Y, \qquad -\mathbf{1}^\mathsf{T} I = I^n, \qquad s = -\left(s^Y + V^n \overline{I}^Y\right) \tag{14.8}$$

where $\overline{I}^Y$ denotes the componentwise complex conjugate of the vector $I^Y \in \mathbb{C}^3$. The negative sign on the current and power conversions is due to the definition of $\left(I^Y, s^Y\right)$ as internal current and power delivered to the single-phase devices whereas (I, s) is defined as the terminal current and power injections out of the three-phase device; see Remark 14.1. The property $-\mathbf{1}^\mathsf{T} I = I^n$ follows from the KCL at the neutral.

Here $s^Y := \operatorname{diag}\left(V^Y I^{Y\mathsf{H}}\right)$ is the internal power delivered across the single-phase devices, or equivalently, $-s^Y$ is the power generated internally by these devices. The term $V^n \overline{I}^Y$ is the vector power delivered across the neutral and the common reference point (e.g., the ground). The terminal power $s := \operatorname{diag}\left(V I^\mathsf{H}\right)$ is power delivered from the device across the phase lines and the common reference point. Hence $-s^Y = s + V^n \overline{I}^Y$ says that the power generated by the device is equal to that delivered to the neutral impedance and the rest of the network. This follows from the conversion between voltages and currents:

$$s := \operatorname{diag}\left(V I^\mathsf{H}\right) = \operatorname{diag}\left(V^Y \left(-I^Y\right)^\mathsf{H}\right) + V^n \operatorname{diag}\left(\mathbf{1}\left(-I^Y\right)^\mathsf{H}\right) = -\left(s^Y + V^n \overline{I}^Y\right)$$

The conversion rule (14.8) holds whether or not there is a neutral line and whether

or not the neutral is grounded with zero or nonzero neutral impedance z^n. If there is not a neutral line then $I^n := 0$ and we have $\mathbf{1}^\mathsf{T} I = \mathbf{1}^\mathsf{T} I^Y = 0$. If the neutral is grounded, then I^n is the current from the neutral to the ground and $V^n = z^n I^n = -z^n \mathbf{1}^\mathsf{T} I$ whether or not $z^n = 0$. If the neutral is ungrounded but connected to the neutral of a 4-wire line, then I^n is the current on the neutral line leaving the neutral of the device. Its value will depend on network interaction; see Example 16.5 and Exercise 16.7.

Remark 14.2 (Neutral voltage V^n)**.** In general the neutral voltage V^n with respect to a common reference point is nonzero whether or not there is a neutral line and whether or not the neutral is grounded. If the neutral is grounded with zero neutral impedance and voltages are defined with respect to the ground, then $V^n = 0$, and hence $V = V^Y$ and $s = -s^Y$. It is important to explicitly include V^n in a network model because not every device in a network may be grounded or grounded with zero neutral impedance. □

Remark 14.3 (Total power)**.** The total terminal power is

$$\mathbf{1}^\mathsf{T} s \;=\; -\mathbf{1}^\mathsf{T} s^Y \;-\; V^n \left(\mathbf{1}^T \overline{I}^Y\right)$$

The first term $\mathbf{1}^\mathsf{T} s^Y$ on the right-hand side is the total power delivered across the single-phase devices. The expression says that the total terminal power injection is equal to the total power $-\mathbf{1}^\mathsf{T} s^Y$ generated internally net of power consumed by the neutral impedance.

If the neutral is ungrounded then $\mathbf{1}^T I^Y = 0$ by KCL and $\mathbf{1}^\mathsf{T} s = -\mathbf{1}^\mathsf{T} s^Y$. If the neutral is grounded (i.e., $V^{n'} = 0$) through an impedance then $V^n \left(\mathbf{1}^T \overline{I}^Y\right)$ is the power delivered to the neutral impedance. In general the internal power delivered to the neutral impedance is $s^n := \left(V^n - V^{n'}\right) \overline{I}^n$ □

Conversion in Δ configuration.

Consider a generic three-phase device in Δ configuration with internal and terminal variables defined as in Figure 14.2(b). We now apply Theorem 14.2 to convert between internal and external variables in Δ configuration.

Voltage and current conversion. The relation between terminal voltage and current (V, I) and internal voltage and current $\left(V^\Delta, I^\Delta\right)$ is:

$$\begin{bmatrix} V^{ab} \\ V^{bc} \\ V^{ca} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{bmatrix}}_{\Gamma} \begin{bmatrix} V^a \\ V^b \\ V^c \end{bmatrix}, \qquad \begin{bmatrix} I^a \\ I^b \\ I^c \end{bmatrix} = -\underbrace{\begin{bmatrix} 1 & 0 & -1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}}_{\Gamma^\mathsf{T}} \begin{bmatrix} I^{ab} \\ I^{bc} \\ I^{ca} \end{bmatrix}$$

or in vector form

$$V^\Delta \;=\; \Gamma V, \qquad I \;=\; -\Gamma^\mathsf{T} I^\Delta \tag{14.9a}$$

where $\Gamma, \Gamma^\mathsf{T}$ are given in (14.5). Given appropriate vectors V^Δ and I, solutions V and I^Δ to (14.9a) is provided by Theorem 14.2.

1 Given V^Δ, there is a solution V to (14.9a) if and only if V^Δ is orthogonal to $\mathbf{1}$, i.e.,

$$V^{ab} + V^{bc} + V^{ca} = 0$$

which expresses Kirchhoff's voltage law. In that case, there is a subspace of solutions V given by

$$V = \Gamma^\dagger V^\Delta + \gamma \mathbf{1} = \frac{1}{3}\Gamma^\mathsf{T} V^\Delta + \gamma \mathbf{1}, \qquad \gamma \in \mathbb{C} \tag{14.9b}$$

This amounts to an arbitrary reference voltage for V. The quantity $\gamma := \frac{1}{3}\mathbf{1}^T V$ is the (scaled) zero-sequence voltage of V. In most applications we are given a reference voltage (e.g., $V_0 := \alpha_+$ at the reference bus 0) which will fix the constant γ for every Δ-configured device (different devices may have different zero-sequence voltages γ).

2 Given I, there is a solution I^Δ to (14.9a) if and only if I is orthogonal to $\mathbf{1}$, i.e.,

$$I^a + I^b + I^c = 0$$

which expresses Kirchhoff's current law. In that case, there is a subspace of I^Δ that satisfy (14.9a), given by

$$I^\Delta = -\Gamma^{\mathsf{T}\dagger} I + \beta \mathbf{1} = -\frac{1}{3}\Gamma I + \beta \mathbf{1}, \qquad \beta \in \mathbb{C} \tag{14.9c}$$

where β specifies the amount of loop flow in I^Δ and does not affect the terminal current I since $\Gamma^\mathsf{T} I^\Delta = 0$. The quantity $\beta := \frac{1}{3}\mathbf{1}^T I^\Delta$ is the (scaled) zero-sequence current of I^Δ.

We make two remarks regarding the solutions (V, I^Δ). First the minimum-norm solution

$$V := \frac{1}{3}\Gamma^\mathsf{T} V^\Delta = \frac{1}{3}\begin{bmatrix} V^{ab} - V^{ca} \\ V^{bc} - V^{ab} \\ V^{ca} - V^{bc} \end{bmatrix}$$

sets $\gamma = 0$ such that $\mathbf{1}^\mathsf{T} V = 3\gamma = 0$. Note that this solution does not set one of (V^a, V^b, V^c) to zero. A consequence of the arbitrary reference voltage is that, given the internal voltage and current $\left(V^\Delta, I^\Delta\right)$ with $\mathbf{1}^\mathsf{T} V^\Delta = 0$ of a Δ-configured device, its terminal power vector s depends on the arbitrary constant γ (similar to the effect of the neutral voltage V^n on s for a Y-configured device); see Remark 14.4. To fix V to be the minimum-norm solution (14.9b) with $\gamma = 0$, it is important to include explicitly the condition $\mathbf{1}^\mathsf{T} V = 0$ together with $V^\Delta = \Gamma V$, i.e., the minimum-norm solution with $\gamma = 0$ is the unique solution to the system of equations:

$$V^\Delta = \Gamma V, \qquad \mathbf{1}^\mathsf{T} V = 0, \qquad \text{(given } V^\Delta \text{ that satisfies } \mathbf{1}^\mathsf{T} V^\Delta = 0)$$

Second the minimum-norm solution sets $\beta = 0$ and is

$$I^\Delta = -\frac{1}{3}\Gamma I = -\frac{1}{3}\begin{bmatrix} I^a - I^b \\ I^b - I^c \\ I^c - I^a \end{bmatrix}$$

It contains zero loop flow, i.e., $\mathbf{1}^{\mathsf{T}} I^{\Delta} = 3\beta = 0$. Analogous to the case above, a consequence of an arbitrary β is that, given the terminal voltage and current (V, I) of a Δ-configured device, its internal power vector s^{Δ} depends on the zero-sequence current β; see Remark 14.4. To fix I to be the minimum-norm solution (14.9c) with $\beta = 0$, it is important to include explicitly the condition $\mathbf{1}^{\mathsf{T}} I^{\Delta} = 0$ together with $I = -\Gamma^{\mathsf{T}} I^{\Delta}$, i.e., the minimum-norm solution with $\beta = 0$ is the unique solution to the system of equations:

$$I \;=\; -\Gamma^{\mathsf{T}} I^{\Delta}, \qquad \mathbf{1}^{\mathsf{T}} I^{\Delta} \;=\; 0 \qquad \text{(given } I \text{ that satisfies } \mathbf{1}^{\mathsf{T}} I = 0)$$

Power conversion. The terminal power injection from the device is $s := \operatorname{diag}\left(V I^{\mathsf{H}}\right)$ and the internal power delivered across the single-phase devices in the direction ab, bc, ca is $s^{\Delta} := \operatorname{diag}\left(V^{\Delta} I^{\Delta \mathsf{H}}\right)$. Unlike a Y-configured power source for which the terminal power s is related directly to the internal power s^{Y} (see (14.8)), for a Δ-configured power source, the relation between s and s^{Δ} is indirect through $\left(V^{\Delta}, I^{\Delta}\right)$, through (V, I), or through $\left(V, I^{\Delta}\right)$. We now derive these relations using the voltage and current conversion (14.9).

Specifically, given internal voltage and current $\left(V^{\Delta}, I^{\Delta}\right)$ with $\mathbf{1}^{\mathsf{T}} V^{\Delta} = 0$, the internal power is $s^{\Delta} := \operatorname{diag}\left(V^{\Delta} I^{\Delta \mathsf{H}}\right)$. To express the terminal power s in terms of $\left(V^{\Delta}, I^{\Delta}\right)$, we use (14.9a) (14.9b) to write the terminal voltage and current as

$$V \;=\; \Gamma^{\dagger} V^{\Delta} + \gamma \mathbf{1}, \quad \gamma \in \mathbb{C}, \qquad\qquad I \;=\; -\Gamma^{\mathsf{T}} I^{\Delta}$$

where different γ correspond to different reference voltages. Therefore

$$V I^{\mathsf{H}} = \left(\Gamma^{\dagger} V^{\Delta} + \gamma \mathbf{1}\right)\left(-\Gamma^{\mathsf{T}} I^{\Delta}\right)^{\mathsf{H}} \;=\; -\Gamma^{\dagger}\left(V^{\Delta} I^{\Delta \mathsf{H}}\right)\Gamma + \gamma\left(\mathbf{1}\, I^{\mathsf{H}}\right)$$

Hence the terminal power s can be expressed in terms of the internal voltage and current $\left(V^{\Delta}, I^{\Delta}\right)$ as

$$s \;:=\; \operatorname{diag}\left(V I^{\mathsf{H}}\right) \;=\; -\operatorname{diag}\left(\Gamma^{\dagger}\left(V^{\Delta} I^{\Delta \mathsf{H}}\right)\Gamma\right) + \gamma \bar{I}, \qquad \mathbf{1}^{\mathsf{T}} V^{\Delta} \;=\; 0 \tag{14.10a}$$

where $\bar{I}$ is the componentwise complex conjugate of the terminal current $I = -\Gamma^{\mathsf{T}} I^{\Delta}$ and $\gamma \in \mathbb{C}$ is determined by a reference voltage.

Example 14.1. Given internal voltage and current $\left(V^{\Delta}, I^{\Delta}\right)$ with $\mathbf{1}^{\mathsf{T}} V^{\Delta} = 0$, evaluate the terminal power $s := \operatorname{diag}\left(V I^{\mathsf{H}}\right)$ directly using the solution (14.9b) with $\gamma := 0$.

Solution. We have

$$I \;=\; -\Gamma^{\mathsf{T}} I^{\Delta} \;=\; -\begin{bmatrix} 1 & 0 & -1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}\begin{bmatrix} I^{ab} \\ I^{bc} \\ I^{ca} \end{bmatrix} \;=\; -\begin{bmatrix} I^{ab} - I^{ca} \\ I^{bc} - I^{ab} \\ I^{ca} - I^{bc} \end{bmatrix}$$

Combine with (14.9b) with $\gamma = 0$ to evaluate $\operatorname{diag}\left(V I^{\mathsf{H}}\right)$:

$$s \;:=\; -\frac{1}{3}\begin{bmatrix} \left(V^{ab} - V^{ca}\right)\left(I^{ab} - I^{ca}\right)^{\mathsf{H}} \\ \left(V^{bc} - V^{ab}\right)\left(I^{bc} - I^{ab}\right)^{\mathsf{H}} \\ \left(V^{ca} - V^{bc}\right)\left(I^{ca} - I^{ab}\right)^{\mathsf{H}} \end{bmatrix} \;=\; -\frac{1}{3}\left(\begin{bmatrix} s^{ab} + s^{ca} \\ s^{bc} + s^{ab} \\ s^{ca} + s^{bc} \end{bmatrix} + \begin{bmatrix} V^{ca} & 0 & V^{ab} \\ V^{bc} & V^{ab} & 0 \\ 0 & V^{ca} & V^{bc} \end{bmatrix}\begin{bmatrix} \bar{I}^{ab} \\ \bar{I}^{bc} \\ \bar{I}^{ca} \end{bmatrix}\right)$$

This is (14.10a) with $\gamma = 0$. □

We next relate s and s^Δ in terms of terminal voltage and current (V, I). Given (V, I) with $\mathbf{1}^\mathsf{T} I = 0$, $s := \operatorname{diag}\left(VI^\mathsf{H}\right)$. To express s^Δ in terms of (V, I), use (14.9a) (14.9c) to write the internal voltage and current as

$$V^\Delta \;=\; \Gamma V, \qquad\qquad I^\Delta \;=\; -\Gamma^{\mathsf{T}\dagger} I \,+\, \beta \mathbf{1}, \quad \beta \in \mathbb{C}$$

where different β correspond to different loop flows in the Δ configuration. Therefore

$$V^\Delta I^{\Delta\mathsf{H}} = -\Gamma\left(VI^\mathsf{H}\right)\Gamma^\dagger \,+\, \overline{\beta}\left(V^\Delta \mathbf{1}^\mathsf{T}\right)$$

Hence the internal power $s^\Delta := \operatorname{diag}\left(V^\Delta I^{\Delta\mathsf{H}}\right)$ can be expressed in terms of the terminal voltage and current (V, I) as

$$s^\Delta \;:=\; \operatorname{diag}\left(V^\Delta I^{\Delta\mathsf{H}}\right) \;=\; -\operatorname{diag}\left(\Gamma\left(VI^\mathsf{H}\right)\Gamma^\dagger\right) \,+\, \overline{\beta} V^\Delta, \qquad \mathbf{1}^\mathsf{T} I \;=\; 0 \tag{14.10b}$$

where $V^\Delta = \Gamma V$ and $\beta \in \mathbb{C}$ is determined by the amount of loop flow in I^Δ.

Even though (14.10a) and (14.10b) contain the zero-sequence voltage and current (γ, β), the total powers $\mathbf{1}^\mathsf{T} s$ and $\mathbf{1}^\mathsf{T} s^\Delta$ do not.

Remark 14.4 (Total powers)**.** 1 Given an internal voltage and current $\left(V^\Delta, I^\Delta\right)$, the terminal power vector s in (14.10a) does not depend on the zero-sequence current $\beta := \frac{1}{3}\mathbf{1}^\mathsf{T} I^\Delta$ but does depend on the zero-sequence voltage $\gamma := \frac{1}{3}\mathbf{1}^\mathsf{T} V$. Since $I = -\Gamma^\mathsf{T} I^\Delta$ and hence $\mathbf{1}^\mathsf{T} I = 0$, the total terminal power however is independent of γ:

$$\mathbf{1}^\mathsf{T} s \;=\; -\mathbf{1}^\mathsf{T} \operatorname{diag}\left(\Gamma^\dagger \left(V^\Delta I^{\Delta\mathsf{H}}\right)\Gamma\right)$$

This is the same as the effect of neutral voltage V^n on terminal power s and its aggregate $\mathbf{1}^\mathsf{T} s$ in Y configuration when the neutral is ungrounded so that $\mathbf{1}^\mathsf{T} I^Y = 0$ by KCL.

2 Analogously, from (14.10b), the internal power vector s^Δ depends on zero-sequence current β. Since $V^\Delta = \Gamma V$ and hence $\mathbf{1}^\mathsf{T} V^\Delta = 0$, the total internal power however is independent of the loop flow:

$$\mathbf{1}^\mathsf{T} s^\Delta = -\mathbf{1}^\mathsf{T} \operatorname{diag}\left(\Gamma\left(VI^\mathsf{H}\right)\Gamma^\dagger\right)$$

It can be shown that $\mathbf{1}^\mathsf{T} \operatorname{diag}\left(\Gamma\left(VI^\mathsf{H}\right)\Gamma^\dagger\right) = \mathbf{1}^\mathsf{T} \operatorname{diag}\left(VI^\mathsf{H}\right)$ (Exercise 14.6). Therefore the *total* internal and terminal powers are equal, i.e., $\mathbf{1}^\mathsf{T} s^\Delta = \mathbf{1}^\mathsf{T} s$. □

Finally we can relate s and s^Δ through the terminal voltage and internal current $\left(V, I^\Delta\right)$. Indeed both s and s^Δ can be expressed in terms of $\left(V, I^\Delta\right)$ using (14.9a):

$$s \;:=\; \operatorname{diag}\left(VI^\mathsf{H}\right) \;=\; -\operatorname{diag}\left(VI^{\Delta\mathsf{H}}\Gamma\right), \qquad s^\Delta \;:=\; \operatorname{diag}\left(V^\Delta I^{\Delta\mathsf{H}}\right) \;=\; \operatorname{diag}\left(\Gamma V I^{\Delta\mathsf{H}}\right) \tag{14.10c}$$

An important advantage of (14.10c) is that (V, I^Δ) contains implicitly both the zero-sequence voltage $\gamma := \frac{1}{3}\mathbf{1}^\mathsf{T} V$ and the zero-sequence current $\beta := \frac{1}{3}\mathbf{1}^\mathsf{T} I^\Delta$. This is often a more computationally convenient model than (14.10a) and (14.10b).

In summary:

- Given internal voltage and current (V^Δ, I^Δ) with $\mathbf{1}^\mathsf{T} V^\Delta = 0$, the terminal power s as a function of (V^Δ, I^Δ) is given by (14.10a).
- Given terminal voltage and current (V, I) with $\mathbf{1}^\mathsf{T} I = 0$, the internal power s^Δ as a function of (V, I) is given by (14.10b).
- Given terminal voltage and internal current (V, I^Δ), the terminal power s and the internal power s^Δ are given by (14.10c).

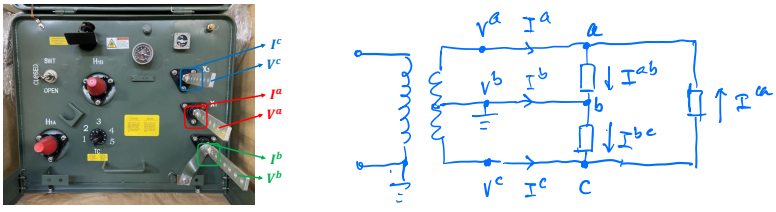
These expressions are used to derive the external model a constant-power source in Δ configuration; see Chapter 14.3.4.

Finally, note that unlike the relation $I = -\Gamma^\mathsf{T} I^\Delta$ which expresses KCL, it is *not* true that $s = -\Gamma^\mathsf{T} s^\Delta$. The relation between terminal power and internal power is given *only indirectly* by (14.10).

14.3.2 Case study: Riverside CA utility

In this subsection we present voltage and current measurements from a distribution transformer in a Southern California municipal utility grid. The case study makes concrete some of the concepts introduced in the previous sections. It also illustrates how unbalanced three-phase models can be used to analyze physical systems that are not necessarily three-phased, in this case a split-phase system modeled as a Δ-configured three-phase load with one terminal grounded.

Figure 14.4 shows a typical pad-mounted split-phase distribution transformer. The transformer in the Southern California grid supplies 8 houses in a residential area in Δ configuration. It is rated at 75 kVA, with 12 kV grounded-Y on the high-voltage side and single split-phase 240V/120V with grounded neutral on the low-voltage side as shown in the figure. We measure the voltage and current phasors $V := (V^a, V^b, V^c)$ and $I := (I^a, I^b, I^c)$ respectively at the low-voltage terminals of the transformer. Terminal b is grounded and used as the common reference point, i.e., $V^b := 0$. Note that the terminal current is defined here to be into the load which is in the opposite direction to what we usually use elsewhere in this chapter, corresponding to the direction in Figure 14.4. We assume that the line loss between the transformer and the load (8 houses) is negligible, and hence V and I are also the terminal voltage and terminal current respectively of the load in Δ configuration. We reiterate that even though we use unbalanced three-phase concepts to model the load, they are on a single (split-)phase on the low-voltage side of the transformer.



(a) Split-phase distribution transformer

(b) Equivalent circuit

Figure 14.4 Typical distribution transformer and the equivalent circuit of the Southern California system supplying 8 houses arranged in Δ configuration.

We illustrate in Figures 14.5 and 14.6 the behavior of the circuit using the noisy time series of (V, I) measured from the field on March 28 Thur, 2024.

1 *Voltage behavior*. The solid lines in Figures 14.5(a) and 14.5(b) show the magnitude and phase respectively of the terminal voltage V. We see from Figure 14.5(a) that the magnitudes $|V^a|$ and $|V^c|$ are roughly 120 V but their phase angles in Figure 14.5(b) are roughly 180° apart most of the time due to the split phase. Notice that the green solid line $|V^b|$ is zero in Figures 14.5(a) and there is no green solid line for voltage angle on line b. Instead the red solid line $\angle V^a = 0°$ in Figure 14.5(b). This is because voltage measurement $v^a(t)$ in the time domain is actually the voltage drop between terminal a and terminal b, which is grounded, and hence $v^b(t) := 0$. This means that, in the phasor domain $\angle V^a$ is arbitrary and it is set to be 0° in our calculation, i.e., $\angle V^a = 0$ is the reference for all voltage, current and power angles. Relative to the potential on the b terminal, $v^c(t)$ is approximately a half cycle off from $v^a(t)$ and $\angle V^c \approx -180°$ most of the time due to the split phase. (See also discussion below on voltage imbalance.)
2 *Current behavior*. The dash lines in Figures 14.5(a) and 14.5(b) show the magnitude and phase respectively of the terminal current I. There are three curves in each of the figures for phases a, b, c. As discussed above the angles $\angle I^\phi$ are relative to the reference $\angle V^a := 0$. The magnitudes of I^a and I^c are similar but their phases are approximately 180° apart most of the time due to the split phase. Both the magnitudes $|I^a|$, $|I^c|$ and their phases $\angle I^a$, $\angle I^c$ show prominently the effect of solar generation between roughly 8am to 5pm. In particular from Figure 14.5(b) during 9am–5pm the power factor angles $\angle V^\phi - \angle I^\phi \approx -180°$ for both phases a and c, resulting in negative real powers $\mathrm{Re}(s^\phi) = |V^\phi||I^\phi| \cos 180°$ during this period, i.e., real powers flow from the loads towards the transformer on phases a and c. The magnitude of I^b is much smaller in Figure 14.5(a) and its angle in Figure 14.5(b) fluctuates between 0° and ±180°, indicating that a relatively small amount of line b current flows back and forth between the transformer and the loads. This means that the current I^a on line a mostly returns as I^c on line c, and hence their angles are approximately 180° apart as noted above.
3 *Power behavior*. We can construct the behavior of the terminal power s from that

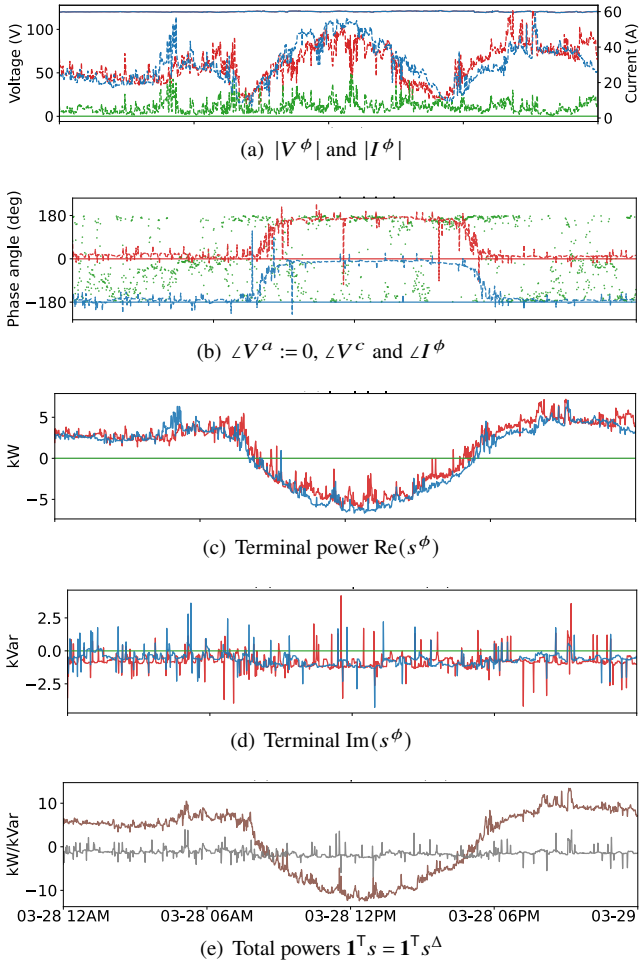
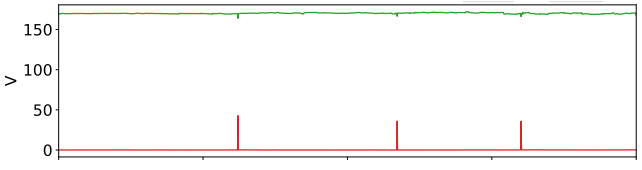


Figure 14.5 Voltage, current and power behavior. (a)(b)(c)(d) Solid lines: voltages, dashed lines: currents. Red: phase a, green: phase b, blue: phase c. (e) Brown: real (kW), grey: imaginary (kVar).

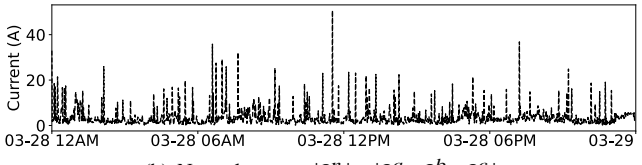
of V and I and confirm that in the measurement. As noted above, between 9am–5pm, the real powers $\mathrm{Re}(s^{\phi})$ on phases a and c are negative, shown as red and blue curves respectively in Figure 14.5(c), whereas they are positive and flow from the transformer to the loads outside this period. From Figure 14.5(d), the reactive powers $\mathrm{Im}(s^{\phi})$ are small most of the time. The green curve representing power on line b is zero because $V^{b} := 0$ by definition.

The internal (load) power s^{Δ}, from (14.10b), is $s^{\Delta} = \frac{1}{3}\mathrm{diag}\left(\Gamma\left(VI^{\mathsf{H}}\right)\Gamma^{\mathsf{T}}\right) + \overline{\beta}(\Gamma V)$ which cannot be computed from (V, I) because of the unknown loop flow parameter $\beta \in \mathbb{C}$. Even though s and s^{Δ} are generally different vectors, the total powers $\mathbf{1}^{\mathsf{T}}s$ and $\mathbf{1}^{\mathsf{T}}s^{\Delta}$ are equal as explained in Remark 14.4. They are illustrated

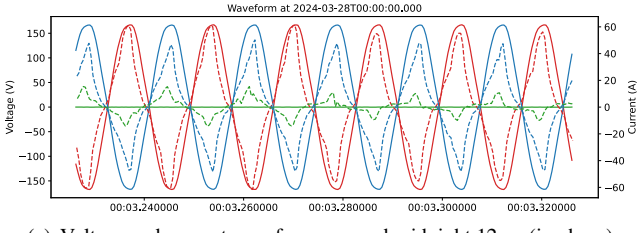
in 14.5(e) which are the sums of the curves in Figure 14.5(c) for the real part and those in Figure 14.5(d) for the imaginary part.



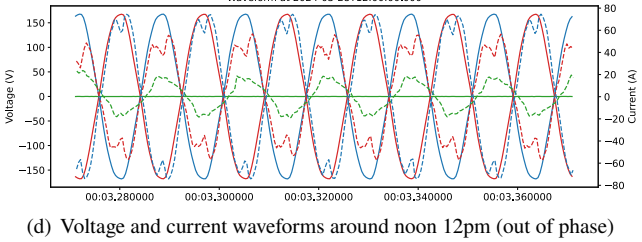
(a) Sequence voltage magntidues $|\tilde{V}_0|$ (red) and $|\tilde{V}_1|$ (green)



(b) Neutral current $|I^n| = |I^a + I^b + I^c|$



(c) Voltage and current waveforms around midnight 12am (in phase)



(d) Voltage and current waveforms around noon 12pm (out of phase)

Figure 14.6 Sequence voltages, neutral current, voltage and current waveforms. (a) Brown: real, grey: imaginary. (c)(d) Solid lines: voltages, dashed lines: currents. Red: phase a, green: phase b, blue: phase c.

4 *Voltage imbalance*. If we view our system as an unbalanced three phase system with grounded terminal b then the zero-sequence voltage $\gamma := \frac{1}{3}(V^a + V^b + V^c)$ can be treated as a measure of voltage imbalance. A more natural perspective is to view the split-phase system as a two-phase system with terminal phase voltages (V^a, V^c) and terminal phase currents (I^a, I^c), return current I^b and a neutral current I^n. We can decompose these voltages along an orthonormal basis for two-phase systems

to obtain the sequence voltages $\tilde{V}$:

$$\tilde{V} := \begin{bmatrix} \tilde{V}_0 \\ \tilde{V}_1 \end{bmatrix} := \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} V^a \\ V^c \end{bmatrix}$$

Note that $\tilde{V}_0$ can be viewed as a measure of voltage imbalance and is equal to $\frac{3}{\sqrt{2}}\gamma$ since $V^b := 0$. The magnitudes $|\tilde{V}_0|$ and $|\tilde{V}_1|$ are shown in Figures 14.6(a). Their normalized values averaged over the measurement period $t = 1, \ldots, T$ are:

$$|\tilde{V}_0| := \frac{1}{T} \sum_{t=1}^{T} \frac{|\tilde{V}_0(t)|}{\|\tilde{V}(t)\|_2} = 0.0010, \qquad |\tilde{V}_1| := \frac{1}{T} \sum_{t=1}^{T} \frac{|\tilde{V}_1(t)|}{\|\tilde{V}(t)\|_2} = 0.9999$$

5 *Neutral current*. From KCL we have $I^a + I^b + I^c = I^n$ where I^n is the neutral current from terminal b to the ground. Its magnitude $|I^n|$ is shown in Figure 14.6(b). It is small most of the time compared with $|I^b|$ on line b. Its magnitude relative to those of the phase currents averaged over the measurement period is

$$\text{average relative neutral current} := \frac{1}{T} \sum_{t=1}^{T} \frac{|I^a(t) + I^b(t) + I^c(t)|}{(|I^a(t)| + |I^b(t)| + |I^c(t)|)/3} = 0.1752$$

6 *Voltage and current waveforms*. Figure 14.6(c) shows the voltage (solid lines) and current (dashed lines) waveforms around midnight where the currents and voltages are roughly in phase, indicating that real power flows from the transformer to the loads. Figure 14.6(d) shows the voltage and current waveforms around noon where the currents and voltages are roughly out of phase, indicating that real power flows from the loads to the transformer.

14.3.3 Devices in Y configuration

In this subsection we first present parameters of a voltage source, current source, power source, and impedance in Y configuration. For each device we then specify its internal model. Finally we apply the conversion rule (14.8) to the internal model of each device to derive its external model.

Device specification.

The devices we study are shown in Figure 14.7.

1 *Voltage source* $\left(E^Y, z^Y, z^n\right)$. A voltage source is a single-terminal three or four-wire device. When the configuration is Y, as shown in Figure 14.7(a), it is specified by three parameters. Its internal voltage is fixed at $E^Y := (E^{an}, E^{bn}, E^{cn})$ and its series impedance matrix is $z^Y := \text{diag}\left(z^{an}, z^{bn}, z^{cn}\right)$. If there is a neutral wire then its impedance is a scalar z^n which may or may not be zero whether or not the neutral is grounded. An ideal voltage source is one with $z^Y = 0$ and $z^n = 0$. A voltage source can serve as a Thévenin equivalent circuit of a synchronous generator for which

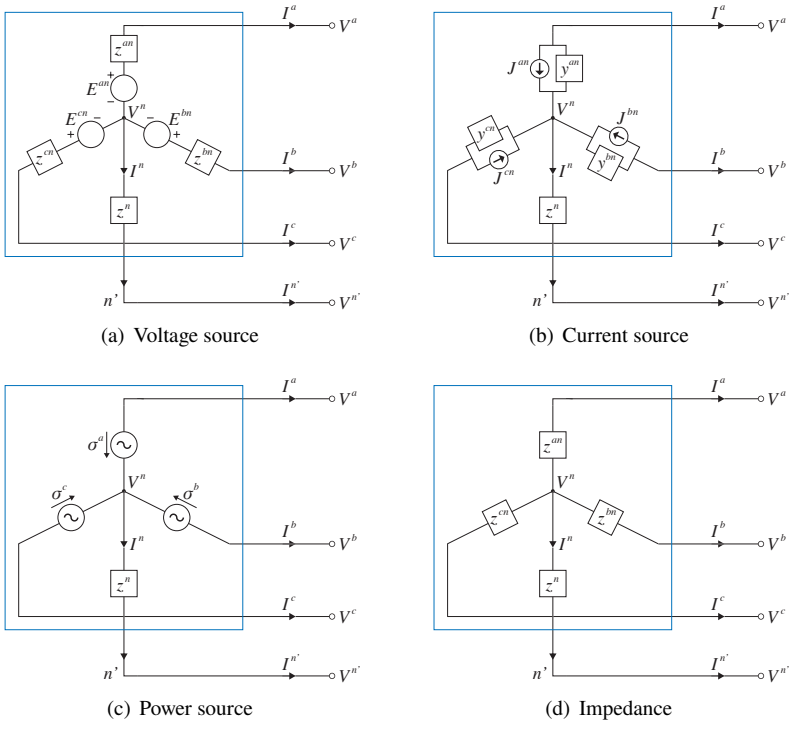


Figure 14.7 Three-phase devices in Y configuration. (a) A voltage source. (b) A current source. (c) A power source. (d) An impedance. Note that the direction of J^Y and σ^Y is terminal-to-neutral.

the internal voltage E^Y is typically balanced. It can also model the primary or secondary side of a transformer, or a grid-forming inverter.

2 *Current source* $\left(J^Y, y^Y, z^n\right)$. A current source is a single-terminal three or four-wire device. When the configuration is Y, as shown in Figure 14.7(b), it is specified by three parameters. Its internal current is fixed at $J^Y := (J^{an}, J^{bn}, J^{cn})$ and its shunt admittance matrix is $y^Y := \text{diag}\left(y^{an}, y^{bn}, y^{cn}\right)$. If there is a neutral wire then its impedance is a scalar z^n which may or may not be zero whether or not the neutral is grounded. An ideal current source is one with $y^Y = 0$ and $z^n = 0$. A current source can serve as a Norton equivalent circuit of a synchronous generator. It can also model a load such as an electric vehicle charger, or a grid-following inverter.

3 *Power source* $\left(\sigma^Y, z^n\right)$. A single-terminal three or four-wire power source in Y configuration is shown in Figure 14.7(c) and specified by two parameters. It consumes a constant power $\sigma^Y := \left(\sigma^{an}, \sigma^{bn}, \sigma^{cn}\right)$ or injects a constant power $-\sigma^Y$. If there is a neutral wire then its impedance is a scalar z^n which may or may not be zero whether or not the neutral is grounded. An ideal power source is one with $z^n = 0$. A power source can model a load, a generator, or the primary or secondary side of a transformer.

4 *Impedance* $\left(z^Y, z^n\right)$. A single-terminal three or four-wire impedance in Y configuration as shown in Figure 14.7(d) is specified by an impedance matrix $z^Y := \text{diag}\left(z^{an}, z^{bn}, z^{cn}\right)$. If there is a neutral wire then its impedance is a scalar z^n which may or may not be zero whether or not the neutral is grounded. An impedance can model a load.

Note that the direction of J^Y and σ^Y is defined to be terminal-to-neutral, opposite to that of the terminal current I.

The list above only specifies the internal parameters of a Y-configured device. When it is connected to a network, its neutral voltage V^n will need to be either specified or computed in order to translate between its internal voltage V^Y and external voltage $V = V^Y + V^n\mathbf{1}$ (from (14.8)) and determine voltages, currents, and powers at other parts of the network. We will discuss in Chapter 16.2, for each device in a typical three-phase analysis problem, what quantities are parameters that should be specified and what are variables to be computed through network equations. An assumption that is often made, sometimes implicitly, is:

C14.1: All neutrals are grounded either through an impedance z^n or directly ($z^n = 0$) and all voltages are defined with respect to the ground.

This assumption is often satisfied in practice. Under this assumption, $V^{n'} = 0$ (see Figure 14.7). Moreover the internal neutral voltage V^n is not independently specified but is determined by the current through the neutral impedance z^n:

$$V^n \;=\; z^n\left(\mathbf{1}^{\mathsf{T}} I^Y\right) \;=\; -z^n\left(\mathbf{1}^{\mathsf{T}} I\right) \tag{14.11}$$

If the neutral is directly grounded, i.e., $z^n = 0$, then $V^n = 0$. Without C14.1 or for an ungrounded voltage source, knowing the internal voltage and current $\left(V^Y, I^Y\right)$ alone may not be sufficient to determine the external voltage V. We will be explicit when we assume C14.1.

Voltage source (E^Y, z^Y, z^n).

Internal model. Referring to Figure 14.7(a) the internal model of a voltage source is

$$V^Y \;=\; E^Y + z^Y I^Y, \qquad V^n - V^{n'} \;=\; z^n\left(\mathbf{1}^{\mathsf{T}} I^Y\right), \qquad I^n \;=\; \mathbf{1}^{\mathsf{T}} I^Y \tag{14.12a}$$

This yields an internal power $s^Y := \text{diag}\left(V^Y I^{Y\mathsf{H}}\right)$ across the non-ideal voltage source and an internal power $s^n := \left(V^n - V^{n'}\right) I^{n\mathsf{H}}$ across the impedance z^n on the neutral line,

given by:

$$s^Y = \operatorname{diag}\left(E^Y I^{Y\mathsf{H}}\right) + \operatorname{diag}\left(z^Y I^Y I^{Y\mathsf{H}}\right) = \underbrace{\begin{bmatrix} E^{an} I^{an\mathsf{H}} \\ E^{bn} I^{bn\mathsf{H}} \\ E^{cn} I^{cn\mathsf{H}} \end{bmatrix}}_{s^Y_{\text{ideal}}} + \underbrace{\begin{bmatrix} z^{an} |I^{an}|^2 \\ z^{bn} |I^{bn}|^2 \\ z^{cn} |I^{cn}|^2 \end{bmatrix}}_{s_{\text{imp}}} \tag{14.12b}$$

$$s^n = z^n \left|\mathbf{1}^\mathsf{T} I^Y\right|^2 \tag{14.12c}$$

External model. To derive an external model, apply the conversion rule (14.8), reproduced here:

$$V = V^Y + V^n \mathbf{1}, \qquad I = -I^Y, \qquad -\mathbf{1}^\mathsf{T} I = I^n, \qquad s = -\left(s^Y + V^n \overline{I}^Y\right)$$

to the internal model (14.12) to eliminate the internal variables (here, $\overline{I}^Y$ is the complex conjugate of vector I^Y componentwise). This yields a relation between its terminal variables (V, I, s):

$$V = E^Y + V^n \mathbf{1} - z^Y I, \quad \mathbf{1}^\mathsf{T} I = -I^n, \quad s = \operatorname{diag}\left(E^Y I^\mathsf{H}\right) + V^n \overline{I} - \operatorname{diag}\left(z^Y I I^\mathsf{H}\right) \tag{14.13a}$$

The model (14.13a) holds whether there is a neutral line or whether the neutral line is grounded or ungrounded but connected to another device over a four-wire line. As discussed before, $I^n = 0$ if the neutral is ungrounded.

Suppose assumption C14.1 holds so that $V^{n'} = 0$ and $V^n = -z^n \left(\mathbf{1}^\mathsf{T} I\right)$. Then (14.13a) yields the external model:

$$V = E^Y - Z^Y I \tag{14.13b}$$

where

$$Z^Y := z^Y + z^n \mathbf{1}\mathbf{1}^\mathsf{T} = \begin{bmatrix} z^{an} + z^n & z^n & z^n \\ z^n & z^{bn} + z^n & z^n \\ z^n & z^n & z^{cn} + z^n \end{bmatrix}$$

This has the same form as that of a single-phase voltage source discussed in Chapter **??**. The neutral impedance z^n couples the phases. Substituting (14.13b) into $s = \operatorname{diag}\left(V I^\mathsf{H}\right)$ expresses the terminal power s as a quadratic function of V:

$$s = \operatorname{diag}\left(V \left(E^Y - V\right)^\mathsf{H} \left((Z^Y)^{-1}\right)^\mathsf{H}\right) \tag{14.13c}$$

assuming Z^Y is invertible. The inverse of Z^Y is calculated in Exercise 14.7.

The linear I-V relation and the nonlinear V-s or I-s relation in (14.2) takes the form of (14.13) for a voltage source.

If $z^n = 0$ then $Z^Y = z^Y$. From (14.13b) the phases are decoupled, i.e., $V^a = E^{an} -$

$z^{an} I^a$, whether or not the current I and the voltage V are balanced. For an ideal voltage source where both $z^n = 0$ and $z^Y = 0$, the internal and external models (14.12) (14.13) here reduce to, under assumption C14.1,

$$V = V = E^Y, \qquad s = s^Y = \operatorname{diag}(E^Y I^{\mathsf{H}})$$

Example 14.2. Unlike for an ideal voltage source, s^Y in (14.12b) includes both the power $s^Y_{\text{ideal}} := \operatorname{diag}\left(E^Y I^{Y\mathsf{H}}\right)$ across the ideal voltage source and the power $s_{\text{imp}} := \operatorname{diag}\left(z^Y I^Y I^{Y\mathsf{H}}\right)$ delivered to the series impedance z^Y. Hence the net power injection is

$$s = -\left(s^Y_{\text{ideal}} + s_{\text{imp}} + V^n \bar{I}^Y\right)$$

Summing across phases a, b, c shows that the total power generated is equal to the total power injection and total power consumed by the internal impedances of the voltage source:

$$-\mathbf{1}^{\mathsf{T}} s^Y_{\text{ideal}} = \mathbf{1}^{\mathsf{T}} s + \mathbf{1}^{\mathsf{T}} s_{\text{imp}} + s^n$$

where s^n given by (14.12c) is the power delivered to the impedance z^n on the neutral wire.

Current source (J^Y, y^Y, z^n).

Internal model. Referring to Figure 14.7(b) the internal model of a current source is given by

$$I^Y = J^Y + y^Y V^Y, \qquad V^n - V^{n'} = z^n \left(\mathbf{1}^{\mathsf{T}} I^Y\right), \qquad I^n = \mathbf{1}^{\mathsf{T}} I^Y \tag{14.14a}$$

This yields an internal power $s^Y := \operatorname{diag}\left(V^Y I^{Y\mathsf{H}}\right)$ across the non-ideal current source and an internal power $s^n := V^n I^{n\mathsf{H}}$ across the impedance z^n on the neutral line, given by (Exercise 14.8):

$$s^Y = \operatorname{diag}\left(V^Y J^{Y\mathsf{H}}\right) + \operatorname{diag}\left(V^Y V^{Y\mathsf{H}} y^{Y\mathsf{H}}\right) = \underbrace{\begin{bmatrix} V^{an} J^{an\mathsf{H}} \\ V^{bn} J^{bn\mathsf{H}} \\ V^{cn} J^{cn\mathsf{H}} \end{bmatrix}}_{s^Y_{\text{ideal}}} + \underbrace{\begin{bmatrix} y^{an\mathsf{H}} |V^{an}|^2 \\ y^{bn\mathsf{H}} |V^{bn}|^2 \\ y^{cn\mathsf{H}} |V^{cn}|^2 \end{bmatrix}}_{s_{\text{adm}}} \tag{14.14b}$$

$$s^n := V^n I^{n\mathsf{H}} = z^n \left| \mathbf{1}^{\mathsf{T}} J^Y + \operatorname{diag}\left(y^Y\right)^{\mathsf{T}} V^Y \right|^2 \tag{14.14c}$$

External model. The derivation here is analogous to that for a voltage source above. Applying the conversion rule (14.8) to the internal model (14.14a) yields an external model of a current source that relates its terminal variables:

$$I = -J^Y - y^Y \left(V - V^n \mathbf{1}\right), \quad \mathbf{1}^{\mathsf{T}} I = -I^n, \quad s = -\operatorname{diag}\left(V J^{Y\mathsf{H}}\right) - \operatorname{diag}\left(V \left(V - V^n \mathbf{1}\right)^{\mathsf{H}} y^{Y\mathsf{H}}\right) \tag{14.15a}$$

As discussed earlier, $I^n = 0$ if the neutral is ungrounded.

Suppose assumption C14.1 holds so that $V^n = -z^n\left(\mathbf{1}^\mathsf{T} I\right)$. Then (14.15a) yields (Exercise 14.9):

$$V \;=\; -\left(z^Y J^Y \,+\, Z^Y I\right), \qquad I \;=\; -A\left(J^Y + y^Y V\right) \tag{14.15b}$$

where, assuming Z^Y is invertible,

$$z^Y \;:=\; \left(y^Y\right)^{-1}, \quad Z^Y \;:=\; z^Y + z^n\mathbf{1}\mathbf{1}^\mathsf{T}, \quad A \;:=\; \mathbb{I} - \frac{z^n}{1+z^n\left(\mathbf{1}^\mathsf{T} y^Y \mathbf{1}\right)}\, y^Y\mathbf{1}\mathbf{1}^\mathsf{T}$$

and $\mathbb{I}$ denotes the identity matrix of size 3. The effective impedance matrix Z^Y is the same matrix in (14.13b) for a voltage source. Substituting (14.15b) into $s = \mathrm{diag}\left(VI^\mathsf{H}\right)$ expresses the terminal power s as a quadratic function of V:

$$s = -\mathrm{diag}\left(V\left(J^{Y\mathsf{H}} + V^\mathsf{H} y^{Y\mathsf{H}}\right)A^\mathsf{H}\right) \tag{14.15c}$$

The linear I-V relation and the nonlinear V-s or I-s relation in (14.2) takes the form of (14.15) for a current source.

Analogous to a voltage source, the phases are decoupled if $z^n = 0$. An ideal current source with $y^Y = 0$ and $z^n = 0$ has $I = -I^Y = -J^Y$ and $s = -\mathrm{diag}\left(V J^{Y\mathsf{H}}\right)$.

Power source $\left(\sigma^Y, z^n\right)$.

Internal model: By definition the power delivered to a constant-power source and the power delivered to the impedance z^n on the neutral line are respectively (Figure 14.7(c))

$$s^Y \;:=\; \mathrm{diag}\left(V^Y I^{Y\mathsf{H}}\right) \;=\; \sigma^Y, \quad s^n \;:=\; \left(V^n - V^{n'}\right) I^{n\mathsf{H}} \;=\; z^n\left|\mathbf{1}^\mathsf{T} I^Y\right|^2 \tag{14.16}$$

External model: Apply the conversion rule to the internal model (14.16) yields an external model that relates the terminal variables:

$$\sigma^Y \;=\; \mathrm{diag}\left(I^{Y\mathsf{H}}\right)V^Y \;=\; -\mathrm{diag}\left(I^\mathsf{H}\right)\left(V - V^n\mathbf{1}\right), \quad s \;=\; -\sigma^Y + V^n\overline{I}, \quad \mathbf{1}^\mathsf{T} I \;=\; -I^n \tag{14.17a}$$

Suppose assumption C14.1 holds so that $V^{n'} = 0$ and $V^n = -z^n\left(\mathbf{1}^\mathsf{T} I\right)$. We can then rewrite the vector $V^n\overline{I}^Y$ as

$$V^n\overline{I} = -z^n\left(\mathbf{1}^\mathsf{T} I\right)\overline{I} \;=\; -z^n\left(\overline{I}I^\mathsf{T}\right)\mathbf{1}$$

This yields a quadratic relation between V and I (Exercise 14.10):

$$V = -\left(\mathrm{diag}\,\overline{I}\right)^{-1}\sigma^Y \;-\; z^n\left(\mathbf{1}\mathbf{1}^\mathsf{T}\right)I \tag{14.17b}$$

and between s and I:

$$s = -\left(\sigma^Y + z^n \left(\bar{I} I^{\mathsf{T}}\right) \mathbf{1}\right) \tag{14.17c}$$

It is generally not possible to solve (14.17b) for I in closed form and hence there is generally not an explicit V-s model for a power source. From (14.17c) the total power $-\mathbf{1}^{\mathsf{T}}\sigma^Y$ generated by the constant-power source is equal to the total power injection and the power delivered to the impedance on the neutral line:

$$-\mathbf{1}^{\mathsf{T}}\sigma^Y = \mathbf{1}^{\mathsf{T}} s + \underbrace{z^n \left(\mathbf{1}^{\mathsf{T}} I^Y\right)}_{-V^n} \underbrace{\left(\mathbf{1}^{\mathsf{T}} \bar{I}^Y\right)}_{-I^{n\mathsf{H}}} = \mathbf{1}^{\mathsf{T}} s + s^n$$

Clearly $s = -\sigma^Y$ if $z^n = 0$.

Impedance $\left(z^Y, z^n\right)$.

Internal model: Referring to Figure 14.7(d) the internal model of an impedance is

$$V^Y = z^Y I^Y, \quad s^Y := V^Y I^{Y\mathsf{H}}, \quad s^n := \left(V^n - V^{n'}\right) I^{n\mathsf{H}} = z^n \left|\mathbf{1}^{\mathsf{T}} I^Y\right|^2 \tag{14.18}$$

External model: Application of the conversion rule (14.8) to the internal model (14.18) yields an external model that relates the terminal variables:

$$V = -z^Y I + V^n \mathbf{1}, \qquad -\mathbf{1}^{\mathsf{T}} I = I^n \tag{14.19a}$$

If assumption C14.1 holds so that $V^{n'} = 0$ and $V^n = -z^n \left(\mathbf{1}^{\mathsf{T}} I\right)$, then the external model reduces to:

$$V = -Z^Y I \tag{14.19b}$$

where $Z^Y := z^Y + z^n \mathbf{1}\mathbf{1}^{\mathsf{T}}$ is the same effective impedance Z^Y in (14.13b) for a voltage source. Substituting (14.19b) into $s = \mathrm{diag}\left(V I^{\mathsf{H}}\right)$ expresses s as a quadratic function of V:

$$s = -\mathrm{diag}\left(V V^{\mathsf{H}} \left((Z^Y)^{-1}\right)^{\mathsf{H}}\right) \tag{14.19c}$$

assuming Z^Y is invertible. If $z^n = 0$ then $Z^Y = z^Y$ is diagonal.

Balanced impedance. When $z^n \neq 0$ but z^Y is balanced, i.e., $z^{an} = z^{bn} = z^{cn}$, then $Z^Y = z^{an}\mathbb{I} + z^n \mathbf{1}\mathbf{1}^{\mathsf{T}}$ and its off-diagonal entries will couple voltages and currents in different phases. One can perform a similarity transformation using the unitary matrix F to what is called the *sequence coordinate* as explained in Chapter 14.2.2. In the sequence coordinate, the transformed impedance $\tilde{Z}^Y$, called the sequence impedance, is diagonal:

$$\tilde{Z}^Y = \begin{bmatrix} z^{an} + 3z^n & 0 & 0 \\ 0 & z^{an} & 0 \\ 0 & 0 & z^{an} \end{bmatrix}$$

This leads to decoupled voltages and currents in the sequence coordinate called symmetrical components. The decoupled relation between the sequence voltages, currents and impedances can be interpreted as defining separate sequence networks that can be analyzed independently. This is explained in Chapter 16.4.1.

Remark 14.5 (Phase decoupling)**.** The matrix $Z^Y := z^Y + z^n \mathbf{1}\mathbf{1}^\mathsf{T}$ in (14.13) (14.15) (14.19) is called the *phase impedance matrix* or the *impedance matrix*.

1 If $z^n = 0$ in these four devices, i.e., the neutrals are directly grounded, then the phases are decoupled. This is because, for a power source, $s = -\sigma^Y$, and for the other devices, the impedance matrix $Z^Y = z^Y$ becomes diagonal and hence $V = z^Y I$.
2 If $z^n \neq 0$ but the currents are balanced, i.e., $I^a + I^b + I^c = 0$ then $I^n = 0$ and $V^{ng} = 0$. In this case the phases are also decoupled. If the voltage V is balanced and $z^{an} = z^{an} = z^{cn}$ then I^n will indeed be zero and the phases will be decoupled (Exercise 14.11).
3 In unbalanced operation, however, the neutral current I^n may be nonzero and Z^Y generally has nonzero off-diagonal entries that couple voltages and currents in different phases. As mentioned above, if $z^{an} = z^{an} = z^{cn}$ then the sequence impedance $\tilde{Z}^Y$ is diagonal and hence decoupled in the sequence domain (Chapter 16.4).

□

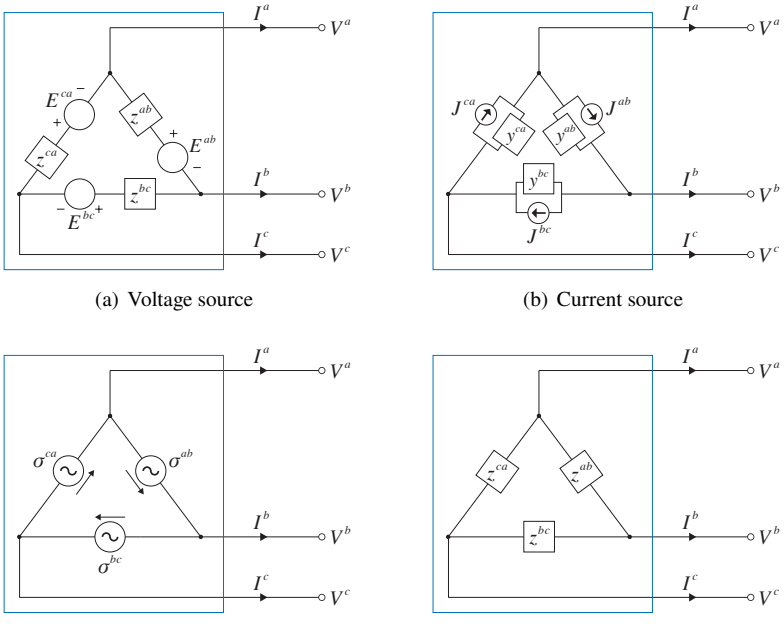
14.3.4 Devices in Δ configuration

In this subsection we first present parameters of the same single-phase devices studied in Chapter 14.3.3, but arranged in Δ rather than Y configuration. For each device we then specify its internal model. Finally we apply the conversion rule (14.9) (14.10) to the internal model of each device to derive its external models.

Internal specification.

The three-phase devices we study are shown in Figure 14.8.

1 *Voltage source* $\left(E^\Delta, z^\Delta\right)$. A three-wire voltage source in Δ configuration as shown in Figure 14.8(a) is specified by its internal line-to-line voltage $E^\Delta := (E^{ab}, E^{bc}, E^{ca})$ and series impedance matrix $z^\Delta := \operatorname{diag}\left(z^{ab}, z^{bc}, z^{ca}\right)$. We assume that $z^{ab} + z^{bc} + z^{ca} \neq 0$. An ideal voltage source is one with $z^\Delta = 0$.
2 *Current source* $\left(J^\Delta, y^\Delta\right)$. A three-wire current source in Δ configuration as shown in Figure 14.8(b) is specified by its internal line-to-line current $J^\Delta := (J^{ab}, J^{bc}, J^{ca})$ and shunt admittance matrix $y^\Delta := \operatorname{diag}\left(y^{ab}, y^{bc}, y^{ca}\right)$. An ideal current source is one with $y^\Delta = 0$.



(a) Voltage source (b) Current source

(c) Power source (d) Impedance

Figure 14.8 Three-phase devices in Δ configuration. (a) A voltage source. (b) A current source. (c) A power load. (d) An impedance. Note the direction of J^Δ and σ^Δ.

3 *Power source* σ^Δ. A three-wire power source in Δ configuration as shown in Figure 14.8(c) consumes a constant power $\sigma^\Delta := \left(\sigma^{ab}, \sigma^{bc}, \sigma^{ca}\right)$ or injects a constant power $-\sigma^\Delta$.

4 *Impedance* z^Δ. A three-wire impedance in Δ configuration as shown in Figure 14.8(d) is specified by an impedance matrix $z^\Delta := \text{diag}\left(z^{ab}, z^{bc}, z^{ca}\right)$. We assume that $z^{ab} + z^{bc} + z^{ca} \neq 0$.

Voltage source $\left(E^\Delta, z^\Delta\right)$.

Internal model. Referring to Figure 14.8(a) the internal model of a voltage source in Δ configuration is

$$V^\Delta \;=\; E^\Delta + z^\Delta I^\Delta, \quad s^\Delta \;:=\; \text{diag}\left(V^\Delta I^{\Delta H}\right) \;=\; \text{diag}\left(E^\Delta I^{\Delta \mathsf{H}}\right) + \text{diag}\left(z^\Delta I^\Delta I^{\Delta \mathsf{H}}\right) \tag{14.20}$$

External model. The terminal voltage and current (V, I) are related to the internal voltage and current $\left(V^\Delta, I^\Delta\right)$ according to the conversion rule (14.9a) for Δ-configured devices, reproduced here

$$V^\Delta \;=\; \Gamma V, \qquad I \;=\; -\Gamma^{\mathsf{T}} I^\Delta$$

We will derive two equivalent relations between the terminal (V, I). Given V, the first relation uniquely determines I in terms of V. Given I, the second relation however determines V in terms of I only up to an arbitrary zero-sequence voltage γ. The asymmetry between these two cases is because V contains more information ($\gamma := \frac{1}{3}\mathbf{1}^{\mathsf{T}}V$) than I and uniquely determines the internal voltage V^{Δ} and hence I^{Δ} (from (14.20)) and I. In contrast I contains no information about the zero-sequence current $\beta := \frac{1}{3}\mathbf{1}^{\mathsf{T}}I^{\Delta}$ and hence does not uniquely determine the internal current I^{Δ}.

For the first relation that maps V to I, define $y^{\Delta} := \left(z^{\Delta}\right)^{-1}$ and write from (14.20)

$$I^{\Delta} \;=\; y^{\Delta}\left(V^{\Delta} - E^{\Delta}\right)$$

Multiplying both sides by $-\Gamma^{\mathsf{T}}$ and substituting the conversion rule we have

$$I \;=\; \left(\Gamma^{\mathsf{T}} y^{\Delta}\right) E^{\Delta} \;-\; Y^{\Delta} V \tag{14.21a}$$

where Y_{Δ} is a complex symmetric Laplacian matrix of the graph in Figure 1.9:[1]

$$Y^{\Delta} \;:=\; \Gamma^{\mathsf{T}} y^{\Delta} \Gamma \;=\; \begin{bmatrix} y^{ab}+y^{ca} & -y^{ab} & -y^{ca} \\ -y^{ab} & y^{bc}+y^{ab} & -y^{bc} \\ -y^{ca} & -y^{bc} & y^{ca}+y^{bc} \end{bmatrix}$$

Note that the terminal current I given by (14.21a) satisfies $\mathbf{1}^{\mathsf{T}} I = 0$.

For the second relation that maps I to V, substitute the conversion rule into the internal model (14.20) to eliminate the internal variable $\left(V^{\Delta}, I^{\Delta}\right)$:

$$\Gamma V \;=\; E^{\Delta} \;+\; z^{\Delta}\left(-\Gamma^{\mathsf{T}\dagger} I + \beta \mathbf{1}\right)$$

where we have used $I^{\Delta} = -\Gamma^{\mathsf{T}\dagger} I + \beta \mathbf{1}$ from (14.9c) and this is valid if and only if we require

$$\mathbf{1}^{\mathsf{T}} I = 0$$

Here $\beta \in \mathbb{C}$ is not arbitrary but depends on E^{Δ} and I.[2] Multiplying both sides by $\mathbf{1}^{\mathsf{T}}$ gives

$$0 \;=\; \mathbf{1}^{\mathsf{T}}\Gamma V \;=\; \mathbf{1}^{\mathsf{T}} E^{\Delta} \;-\; \underbrace{\mathbf{1}^{\mathsf{T}} z^{\Delta}}_{\tilde{z}^{\Delta\mathsf{T}}} \Gamma^{\mathsf{T}\dagger} I \;+\; \beta \underbrace{\left(\mathbf{1}^{\mathsf{T}} z^{\Delta} \mathbf{1}\right)}_{\zeta}$$

Define the column vector $\tilde{z}^{\Delta} := z^{\Delta}\mathbf{1} = \left(z^{ab}, z^{bc}, z^{ca}\right)$ and the scalar $\zeta := \mathbf{1}^{\mathsf{T}} z^{\Delta} \mathbf{1} = z^{ab} +$

[1] Note however that y^{Δ} is a complex matrix and therefore Y^{Δ} is complex symmetric, not Hermitian. Therefore $\mathrm{span}(\mathbf{1})$ is a subset of the null space of Y^{Δ}. For a sufficient condition for the null space of Y^{Δ} to be $\mathrm{span}(\mathbf{1})$, see Exercise 4.2.

[2] To gain intuition, imagine the voltage source is connected to a constant-voltage device that fixes the terminal voltage V of the voltage source, and hence its internal voltage $V^{\Delta} = \Gamma V$. Therefore, on each phase line, say, line ab, we have $V^{ab} - E^{ab} = z^{ab} I^{ab}$. Hence I^{Δ} is uniquely determined which fixes both I and $\beta := \frac{1}{3}\mathbf{1}^{\mathsf{T}} I^{\Delta}$.

$z^{bc} + z^{ca}$. Then

$$\beta \;=\; \frac{1}{\zeta}\left(\tilde{z}^{\Delta\mathsf{T}}\Gamma^{\mathsf{T}\dagger} I \;-\; \mathbf{1}^{\mathsf{T}} E^{\Delta}\right)$$

Note that $\mathbf{1}^{\mathsf{T}} E^{\Delta}$ is the zero-sequence internal voltage and $\tilde{z}^{\Delta}$ is the vector of internal impedances. Both are zero, and hence $\beta = 0$, if the internal voltage E^{Δ} and impedances $\tilde{z}^{\Delta}$ are balanced. Therefore

$$\begin{aligned}\Gamma V \;&=\; E^{\Delta} \;-\; z^{\Delta}\Gamma^{\mathsf{T}\dagger} I \;+\; \frac{1}{\zeta} z^{\Delta}\mathbf{1}\left(\tilde{z}^{\Delta\mathsf{T}}\Gamma^{\mathsf{T}\dagger} I \;-\; \mathbf{1}^{\mathsf{T}} E^{\Delta}\right)\\ &=\; \left(\mathbb{I} - \frac{1}{\zeta}\tilde{z}^{\Delta}\mathbf{1}^{\mathsf{T}}\right) E^{\Delta} \;-\; z^{\Delta}\left(\mathbb{I} - \frac{1}{\zeta}\mathbf{1}\tilde{z}^{\Delta\mathsf{T}}\right)\Gamma^{\mathsf{T}\dagger} I\end{aligned}$$

or

$$V \;=\; \hat{\Gamma} E^{\Delta} \;-\; Z^{\Delta} I \;+\; \gamma\mathbf{1}, \qquad\qquad \mathbf{1}^{\mathsf{T}} I \;=\; 0 \tag{14.21b}$$

where (using Theorem 14.2)

$$\hat{\Gamma} \;:=\; \frac{1}{3}\Gamma^{\mathsf{T}}\left(\mathbb{I} - \frac{1}{\zeta}\tilde{z}^{\Delta}\mathbf{1}^{\mathsf{T}}\right), \qquad\qquad Z^{\Delta} \;:=\; \frac{1}{9}\Gamma^{\mathsf{T}} z^{\Delta}\left(\mathbb{I} - \frac{1}{\zeta}\mathbf{1}\tilde{z}^{\Delta\mathsf{T}}\right)\Gamma$$

and γ is fixed by a given reference voltage. This is similar to (14.13b) for the Y-configured voltage source.

The two external models (14.21a) and (14.21b) are equivalent in the following sense.

Theorem 14.3. Given the conversion rules $V^{\Delta} = \Gamma V$ and $I = -\Gamma^{\mathsf{T}} I^{\Delta}$ between the terminal and internal voltages and currents, the following are equivalent:

1 Internal model: $V^{\Delta} \;=\; E^{\Delta} + z^{\Delta} I^{\Delta}$ and $\mathbf{1}^{\mathsf{T}}\left(E^{\Delta} + z^{\Delta} I^{\Delta}\right) = 0$.
2 External model: $I \;=\; \left(\Gamma^{\mathsf{T}} y^{\Delta}\right) E^{\Delta} - Y^{\Delta} V$ where $Y^{\Delta} := \Gamma^{\mathsf{T}} y^{\Delta}\Gamma$.
3 External model: $V = \hat{\Gamma} E^{\Delta} - Z^{\Delta} I + \gamma\mathbf{1}$, $\mathbf{1}^{\mathsf{T}} I = 0$ for some $\gamma \in \mathbb{C}$ where

$$\hat{\Gamma} \;:=\; \frac{1}{3}\Gamma^{\mathsf{T}}\left(\mathbb{I} - \frac{1}{\zeta}\tilde{z}^{\Delta}\mathbf{1}^{\mathsf{T}}\right), \qquad\qquad Z^{\Delta} \;:=\; \frac{1}{9}\Gamma^{\mathsf{T}} z^{\Delta}\left(\mathbb{I} - \frac{1}{\zeta}\mathbf{1}\tilde{z}^{\Delta\mathsf{T}}\right)\Gamma$$

□

The proof of the theorem is similar to that of Theorem 14.4 and left as Exercise 14.14.

Hence given V, I is uniquely determined by (14.21a) and given I, V is determined by (14.21b) up to a reference voltage specified by γ. These equations allow us to relate terminal power injection s to V or to I as:

$$s \;=\; \operatorname{diag}\left(VI^{\mathsf{H}}\right) \;=\; \operatorname{diag}\left(V\left(\Gamma^{\mathsf{T}} y^{\Delta} E^{\Delta} \;-\; Y^{\Delta} V\right)^{\mathsf{H}}\right) \tag{14.21c}$$

$$s \;=\; \operatorname{diag}\left(VI^{\mathsf{H}}\right) \;=\; \operatorname{diag}\left(\left(\hat{\Gamma} E^{\Delta} \;-\; Z^{\Delta} I\right) I^{\mathsf{H}}\right) \;+\; \gamma\bar{I} \tag{14.21d}$$

For an ideal voltage source where $z^\Delta = 0$ we have $\hat{\Gamma} := \frac{1}{3}\Gamma^\mathsf{T}$ and $Z^\Delta = 0$. The external model is, provided $\mathbf{1}^\mathsf{T} E^\Delta = 0$,

$$V = \frac{1}{3}\Gamma^\mathsf{T} E^\Delta + \gamma \mathbf{1}, \qquad \mathbf{1}^\mathsf{T} I = 0, \qquad s = \frac{1}{3}\mathrm{diag}\left(\Gamma^\mathsf{T} E^\Delta I^\mathsf{H}\right) + \gamma \bar{I}$$

where γ is fixed by a reference voltage.

Current source $\left(J^\Delta, y^\Delta\right)$.

Internal model. Referring to Figure 14.8(b) the internal model of a current source in Δ configuration is

$$I^\Delta = J^\Delta + y^\Delta V^\Delta, \quad s^\Delta := \mathrm{diag}\left(V^\Delta I^{\Delta H}\right) = \mathrm{diag}\left(V^\Delta J^{\Delta\mathsf{H}}\right) + \mathrm{diag}\left(V^\Delta V^{\Delta\mathsf{H}} y^{\Delta\mathsf{H}}\right) \tag{14.22}$$

External model. Multiplying both sides of $I^\Delta = J^\Delta + y^\Delta V^\Delta$ by $-\Gamma^\mathsf{T}$ and substituting the general conversion rule

$$V^\Delta = \Gamma V, \qquad I = -\Gamma^\mathsf{T} I^\Delta$$

for Δ-configured devices, we have

$$I = -\left(\Gamma^\mathsf{T} J^\Delta + Y^\Delta V\right) \tag{14.23a}$$

where $Y^\Delta := \Gamma^\mathsf{T} y^\Delta \Gamma$ is the matrix in (14.21a). The power injection is

$$s = \mathrm{diag}\left(V I^\mathsf{H}\right) = -\mathrm{diag}\left(V J^{\Delta\mathsf{H}} \Gamma + V V^\mathsf{H} Y^{\Delta\mathsf{H}}\right) \tag{14.23b}$$

For an ideal current source where $y^\Delta = 0$ we have $I = -\Gamma^\mathsf{T} J^\Delta$ and $s = -\mathrm{diag}(V J^{\Delta\mathsf{H}}\Gamma)$.

Remark 14.6 (Voltage and current sources)**.** A Δ-configured current source specifies its internal current J^Δ which then uniquely determines its terminal current I through the conversion rule (14.9a), as well as its zero-sequence current $\beta := \frac{1}{3}\mathbf{1}^\mathsf{T} J^\Delta$, whereas a voltage source specifies its internal voltage E^Δ which does not uniquely determine its terminal voltage V. This is why the external voltage source model (14.21b) determines V only up to an arbitrary zero-sequence voltage γ and requires $\mathbf{1}^\mathsf{T} I = 0$ while both (14.21a) and (14.23a) are valid without any extra condition as their derivation does not involve pseudo-inverse of conversion matrices.

Power source σ^Δ.

Internal model. Referring to Figure 14.8(c) the internal model of a constant-power source is

$$s^\Delta := \mathrm{diag}\left(V^\Delta I^{\Delta\mathsf{H}}\right) = \sigma^\Delta \tag{14.24}$$

This specifies the powers $(\sigma^{ab}, \sigma^{bc}, \sigma^{ca})$ delivered to these single-phase devices.

External model. Applying the power conversion rule (14.10b) to the internal model $s^\Delta = \sigma^\Delta$ yields an external model of a constant-power source that relates its terminal voltage and current (V, I):

$$\sigma^\Delta \;=\; -\frac{1}{3}\operatorname{diag}\left(\Gamma\left(VI^{\mathsf{H}}\right)\Gamma^{\mathsf{T}}\right) + \overline{\beta}\Gamma V, \qquad \mathbf{1}^{\mathsf{T}} I \;=\; 0 \tag{14.25a}$$

where the first equality follows because $\left(\Gamma^{\mathsf{T}\dagger}\right)^{\mathsf{H}} = \frac{1}{3}\Gamma^{\mathsf{H}} = \frac{1}{3}\Gamma^{\mathsf{T}}$ from Theorem 14.2. Here β represents the amount of loop flow in the internal current I^Δ. All three quantities (V, I, β) are variables to be determined by the interaction with other devices through the network; see Chapter 16.1. Here (V, I) are terminal variables but, unlike the external models of other devices, β is a quantity internal to the Δ configuration.

An alternative model of a constant-power source is (14.10c) that relates its terminal voltage V with its *internal* current I^Δ:

$$\sigma^\Delta \;:=\; \operatorname{diag}\left(V^\Delta I^{\Delta\mathsf{H}}\right) \;=\; \operatorname{diag}\left(\Gamma V I^{\Delta\mathsf{H}}\right) \tag{14.25b}$$

An advantage of this model is that it contains implicitly both the zero-sequence terminal voltage $\gamma := \frac{1}{3}\mathbf{1}^{\mathsf{T}} V$ and zero-sequence internal current $\beta := \frac{1}{3}\mathbf{1}^{\mathsf{T}} I^\Delta$.

We now study the connection between the two equivalent models (14.25a) and (14.25b) of a constant-power source that relate (V, I) and $\left(V, I^\Delta\right)$ respectively. Expand the first equation in (14.25a) to get

$$\sigma^\Delta \;=\; -\frac{1}{3}\begin{bmatrix}\left(I^a - I^b\right)^{\mathsf{H}}\left(V^a - V^b\right)\\ \left(I^b - I^c\right)^{\mathsf{H}}\left(V^b - V^c\right)\\ \left(I^c - I^a\right)^{\mathsf{H}}\left(V^c - V^a\right)\end{bmatrix} + \overline{\beta}\begin{bmatrix}V^a - V^b\\ V^b - V^c\\ V^c - V^a\end{bmatrix} \;=\; \underbrace{\left(\operatorname{diag}\left(\left(-\Gamma^{\mathsf{T}\dagger} I\right)^{H}\right) + \overline{\beta}\mathbb{I}\right)}_{\operatorname{diag}\left(I^{\Delta\mathsf{H}}\right)}(\Gamma V)$$

which is equivalent to (14.25b). Given a terminal voltage V, the currents I and I^Δ can be uniquely determined in these models (14.25a) and (14.25b) respectively. Given a current I or I^Δ in (14.25a) and (14.25b) respectively, however, V cannot be uniquely determined.

Specifically, given a terminal voltage V, the model (14.25b) provides three linear equations in three unknowns I^Δ, which determines I^Δ uniquely. Both the terminal current I and β are then determined uniquely. Conversely, given I^Δ (and hence β), (14.25b) provides three linear equations in three unknowns V but only $\left(V^a - V^b, V^b - V^c, V^c - V^a\right)$, i.e., $V^\Delta = \Gamma V$, can be uniquely determined. The terminal voltage V (or equivalently, its zero-sequence voltage γ) needs to be determined through network equations or from a reference voltage.

Similarly for the model (14.25b), given a terminal voltage V, (14.25a) provides four linear equations in four unknowns $I := (I^a, I^b, I^c)$ and β which determine (I, β) uniquely (Exercise 14.15). Intuitively, the given terminal voltage V fixes the internal voltage V^Δ which then fixes the internal current I^Δ since $\operatorname{diag}\left(V^\Delta I^{\Delta\mathsf{H}}\right) = \sigma^\Delta$. This then produces a unique terminal current I and the zero-sequence current $\beta := \frac{1}{3}\mathbf{1}^{\mathsf{T}} I^\Delta$.

On the other hand, consider the situation where the terminal current I with $\mathbf{1}^{\mathsf{T}} I = 0$ is given, instead of I^{Δ} as for the model (14.25b) above. In this case (14.25a) also does not uniquely determine the terminal voltage V because (14.25a) provides three quadratic equations in four unknowns (V, β), quadratic due to the term $\beta\Gamma V$. Moreover since I contains less information than I^{Δ}, there is ambiguity in β in addition to γ; see Exercise 14.16. As for the model (14.25b) the terminal voltage V (hence γ) and β will be determined through network equations or from a reference voltage.

For a balanced system however the loop flow β and the internal voltages V^{Δ} are uniquely determined by σ^{Δ} and a terminal current I, as the next example illustrates.

Example 14.3 (Balanced systems)**.** Consider a constant-power source with a given σ^{Δ} whose external behavior is described by (14.25a). Given a terminal current $I = i\alpha_+$ which is a positive-sequence balanced vector with $\mathbf{1}^{\mathsf{T}} I = 0$:

1 Show that the given σ^{Δ} and I must satisfy
$$\sigma^{\Delta} \in \operatorname{span}\left(-\frac{1-\alpha}{3} i\mathbf{1} + \overline{\beta}\alpha_+\right)$$
for some $\beta \in \mathbb{C}$. Note that the internal power σ^{Δ} is different in each phase (with different phase angles separated by 120°) if and only if the loop flow $\beta \neq 0$.
2 Show that the loop flow β and the internal voltage V^{Δ} are uniquely determined by σ^{Δ} and I, and that the terminal voltage V is unique only up to an arbitrary reference voltage.

Assume that the internal voltage V^{Δ} is also a positive-sequence balanced vector.

Solution. By Corollary 1.3 we have for any balanced vector $x \in \mathbb{C}^3$ in positive sequence
$$\Gamma x = (1-\alpha)x, \qquad \Gamma^{\mathsf{T}} x = (1-\alpha^2)x$$
Hence the internal current is
$$I^{\Delta} = -\Gamma^{\mathsf{T}\dagger} I + \beta\mathbf{1} = -\frac{1}{3}\Gamma I + \beta\mathbf{1} = -\frac{1-\alpha}{3} i\alpha_+ + \beta\mathbf{1}$$
where the second equality follows from Theorem 14.2. By assumption V^{Δ} is a positive-sequence balanced vector, i.e., $V^{\Delta} = v\alpha_+$ where $v \in \mathbb{C}$ is a scalar to be determined. Then
$$\begin{aligned}\sigma^{\Delta} = \operatorname{diag}\left(V^{\Delta} I^{\Delta \mathsf{H}}\right) &= v \operatorname{diag}\left(\alpha_+ \left(-\frac{(1-\alpha)i}{3}\alpha_+ + \beta\mathbf{1}\right)^{\mathsf{H}}\right) \\ &= v\left(-\frac{(1-\alpha)i}{3}\operatorname{diag}\left(\alpha_+\alpha_+^{\mathsf{H}}\right) + \overline{\beta}\operatorname{diag}\left(\alpha_+\mathbf{1}^{\mathsf{T}}\right)\right) \\ &= v\left(-\frac{(1-\alpha)i}{3}\mathbf{1} + \overline{\beta}\alpha_+\right)\end{aligned}$$
i.e., σ^{Δ} lies in $\operatorname{span}\left(-\frac{(1-\alpha)i}{3}\mathbf{1} + \overline{\beta}\alpha_+\right)$ for some β. To determine v, multiplying both

sides by $\mathbf{1}^\mathsf{T}$ to get

$$v \;=\; \frac{-\mathbf{1}^\mathsf{T}\sigma^\Delta}{(1-\alpha)i}$$

Then $V^\Delta = v\alpha_+$. The terminal voltage V is given by

$$V = \Gamma^\dagger V^\Delta + \gamma\mathbf{1} \;=\; \frac{v}{3}\Gamma^\mathsf{T}\alpha_+ + \gamma\mathbf{1} \;=\; \frac{-\mathbf{1}^\mathsf{T}\sigma^\Delta(1+\alpha)}{3i}\alpha_+ + \gamma\mathbf{1}, \qquad \gamma \in \mathbb{C}$$

which is unique up to an arbitrary reference voltage specified by $\gamma \in \mathbb{C}$.

Note that neither V^Δ nor V depends on β, even though from the expression above for σ^Δ in part 1, the internal powers $\sigma^\Delta := (\sigma^{ab}, \sigma^{bc}, \sigma^{ca})$ depend on the loop flow specified by β. Moreover the expression uniquely determines β:

$$\sigma^{ab} \;=\; v\left(-\frac{(1-\alpha)i}{3} + \overline{\beta}\right), \quad \sigma^{bc} \;=\; v\left(-\frac{(1-\alpha)i}{3} + \alpha\overline{\beta}\right) \;\Longrightarrow\; \overline{\beta} \;=\; \frac{\sigma^{bc}-\sigma^{ab}}{\sigma^{ab}+\sigma^{bc}+\sigma^{ca}}\,i$$

□

Whereas (14.25a) relates the internal power σ^Δ to the external voltage and current (V, I), we can also use the conversion rule (14.10a) to relate the external power s to the internal voltage and current (V^Δ, I^Δ). Specifically, the internal voltage and current (V^Δ, I^Δ) and the terminal power s of a constant-power source must satisfy:

$$s \;=\; -\frac{1}{3}\operatorname{diag}\left(\Gamma^\mathsf{T}\left(V^\Delta I^{\Delta\mathsf{H}}\right)\Gamma\right) - \gamma\Gamma^\mathsf{T}\overline{I}^\Delta, \quad \sigma^\Delta \;=\; \operatorname{diag}\left(V^\Delta I^{\Delta\mathsf{H}}\right), \quad \mathbf{1}^\mathsf{T}V^\Delta \;=\; 0 \tag{14.25c}$$

where γ is fixed by a reference voltage. An equivalent model in terms of (V, I^Δ) is (using (14.10c))

$$s \;=\; -\operatorname{diag}\left(VI^{\Delta\mathsf{H}}\Gamma\right), \qquad \sigma^\Delta \;=\; \operatorname{diag}\left(\Gamma VI^{\Delta\mathsf{H}}\right) \tag{14.25d}$$

The choice of different models in (14.25) for three-phase analysis depends on the specification of the problem. See Example 16.11 in Chapter 16.2.1.

Remark 14.7 (Total power)**.** Since σ^Δ is the power delivered to the single-phase devices while s is the power injected from the three-phase power source to the network it is connected to, (14.25) implies that (the negative of) its total internal power is equal to its total terminal power, i.e., $\mathbf{1}^\mathsf{T}s = -\mathbf{1}^\mathsf{T}\sigma^\Delta$ (Exercise 14.17). In particular the total terminal power $\mathbf{1}^\mathsf{T}s$ is independent of the loop-flow β and zero-sequence voltage γ even when s does.

Impedance z^Δ.

Internal model. Referring to Figure 14.8(d) the internal model of an impedance z^Δ in Δ configuration is

$$V^\Delta \;=\; z^\Delta I^\Delta, \qquad s^\Delta \;=\; \operatorname{diag}\left(V^\Delta I^{\Delta\mathsf{H}}\right) \;:=\; \operatorname{diag}\left(z^\Delta I^\Delta I^{\Delta\mathsf{H}}\right) \tag{14.26}$$

External model. The external model can be derived in a similar way to that for a voltage source, by applying the conversion rule $V^\Delta = \Gamma V$, $I = -\Gamma^\mathsf{T} I^\Delta$ to the internal model (14.26). We will derive first a relation that maps a terminal voltage V (which also determines its zero-sequence component γ) uniquely to a terminal current I and then a converse relation that maps I to V up to an arbitrary γ.

Define the admittance matrix $y^\Delta := \left(z^\Delta\right)^{-1}$. Substituting into (14.26), multiplying both sides by $-\Gamma^\mathsf{T}$ and applying the conversion rule, we get

$$I \;=\; -Y^\Delta V \tag{14.27a}$$

where $Y^\Delta := \Gamma^\mathsf{T} y^\Delta \Gamma$ is the same complex symmetric Laplacian matrix in (14.21a) for a voltage source. Note that the terminal current I given by (14.27a) satisfies $\mathbf{1}^\mathsf{T} I = 0$.

For the converse relation, given any terminal current I that satisfies $\mathbf{1}^\mathsf{T} I = 0$, substitute the conversion rule into the internal model (14.26) to eliminate $\left(V^\Delta, I^\Delta\right)$:

$$\Gamma V \;=\; z^\Delta \left(-\Gamma^{\mathsf{T}\dagger} I + \beta \mathbf{1}\right)$$

where $\beta \in \mathbb{C}$ is not arbitrary but depends on I. Multiplying both sides by $\mathbf{1}^\mathsf{T}$ gives

$$0 \;=\; \mathbf{1}^\mathsf{T}\Gamma V \;=\; -\underbrace{\mathbf{1}^\mathsf{T} z^\Delta}_{\tilde{z}^{\Delta\mathsf{T}}} \Gamma^{\mathsf{T}\dagger} I \;+\; \beta \underbrace{\left(\mathbf{1}^\mathsf{T} z^\Delta \mathbf{1}\right)}_{\zeta}$$

where $\tilde{z}^\Delta := z^\Delta \mathbf{1}$ and $\zeta := z^{ab} + z^{bc} + z^{ca}$. Hence

$$\beta \;=\; \frac{1}{\zeta}\left(\tilde{z}^{\Delta\mathsf{T}} \Gamma^{\mathsf{T}\dagger}\right) I$$

Therefore

$$\Gamma V \;=\; -z^\Delta \left(\mathbb{I} - \frac{1}{\zeta}\mathbf{1}\tilde{z}^{\Delta\mathsf{T}}\right)\Gamma^{\mathsf{T}\dagger} I$$

or

$$V \;=\; -Z^\Delta I + \gamma \mathbf{1}, \qquad\qquad \mathbf{1}^\mathsf{T} I \;=\; 0 \tag{14.27b}$$

where γ is a variable to be determined together with V and (using Theorem 14.2)

$$Z^\Delta \;:=\; \frac{1}{9}\Gamma^\mathsf{T} z^\Delta \left(\mathbb{I} - \frac{1}{\zeta}\mathbf{1}\tilde{z}^{\Delta\mathsf{T}}\right)\Gamma$$

is the same matrix in (14.21b).

Remark 14.8. Note that (14.27b) is a system of at most 4 linearly independent equations in 7 variables (V, I, γ). We can also eliminate the variable $\gamma := \frac{1}{3}\mathbf{1}^\mathsf{T} V$ and write (14.27b) equivalently in terms of only (V, I):

$$\left(\mathbb{I} - \frac{1}{3}\mathbf{1}\mathbf{1}^\mathsf{T}\right) V \;=\; -Z^\Delta I, \qquad\qquad \mathbf{1}^\mathsf{T} I \;=\; 0$$

Since the matrices on both sides of the first equation are singular, this is a system of at most 3 linearly independent equations in 6 variables. It is often more convenient to

use (14.27b) in analysis as it expresses V explicitly in terms of I despite the additional variable γ; see Example 16.8. □

As for a voltage source, the two external models (14.27a) and (14.27b) of an impedance are equivalent in the following sense. The theorem also implies that Z^Δ and Y^Δ are pseudo-inverses of each other.

Theorem 14.4. Given the conversion rules $V^\Delta = \Gamma V$ and $I = -\Gamma^\mathsf{T} I^\Delta$ between the terminal and internal voltages and currents, the following are equivalent:

1 Internal model: $V^\Delta = z^\Delta I^\Delta$ and hence $\mathbf{1}^\mathsf{T} z^\Delta I^\Delta = 0$.
2 External model: $I = -Y^\Delta V$ where $Y^\Delta := \Gamma^\mathsf{T} y^\Delta \Gamma$.
3 External model: $V = -Z^\Delta I + \gamma \mathbf{1}$, $\mathbf{1}^\mathsf{T} I = 0$ for some $\gamma \in \mathbb{C}$ where

$$Z^\Delta := \frac{1}{9}\Gamma^\mathsf{T} z^\Delta \left(\mathbb{I} - \frac{1}{\zeta}\mathbf{1}\tilde{z}^{\Delta\mathsf{T}}\right)\Gamma$$

Proof The derivation above of the two external models (14.27a) and (14.27b) shows that 1 ⇒ 2 and 3. For the converse we will show that 2 ⇒ 1 and 3 ⇒ 1.

Suppose $I = -Y^\Delta V = -\left(\Gamma^\mathsf{T} y^\Delta \Gamma\right) V$. Substitute the conversion rules to get

$$\Gamma^\mathsf{T}\left(y^\Delta V^\Delta - I^\Delta\right) = 0$$

i.e., $y^\Delta V^\Delta - I^\Delta$ is in the null space of Γ^T, or $y^\Delta V^\Delta - I^\Delta = \beta\mathbf{1}$ for some $\beta \in \mathbb{C}$. Therefore

$$V^\Delta = z^\Delta I^\Delta + \beta z^\Delta \mathbf{1}$$

It is important to note that this expression is not of the form $V^\Delta = z'^\Delta I^\Delta + \beta'\mathbf{1}$ for some *diagonal* matrix $z'^\Delta \in \mathbb{C}^3$ and *scalar* $\beta' \in \mathbb{C}$. Since $\mathbf{1}^\mathsf{T} V^\Delta = 0$ because of the conversion rule, multiplying both sides by $\mathbf{1}^\mathsf{T}$ yields

$$\beta = -\frac{1}{\zeta}\tilde{z}^{\Delta\mathsf{T}} I^\Delta$$

where $\tilde{z}^\Delta := z^\Delta\mathbf{1}$ and $\zeta := z^{ab} + z^{bc} + z^{ca}$. Hence

$$V^\Delta = z^\Delta I^\Delta - \frac{1}{\zeta}\tilde{z}^{\Delta\mathsf{T}} I^\Delta z^\Delta \mathbf{1} = \underbrace{z^\Delta\left(\mathbb{I} - \frac{1}{\zeta}\mathbf{1}\mathbf{1}^\mathsf{T} z^\Delta\right)}_{z'^\Delta} I^\Delta$$

For z'^Δ to be a valid three-phase impedance, it must be a diagonal matrix. This is the case if and only if $z^\Delta\mathbf{1}\left(\mathbf{1}^\mathsf{T} z^\Delta I^\Delta\right) = 0$ in which case $V^\Delta = z^\Delta I^\Delta$, as desired.

Suppose $V = -Z^\Delta I + \gamma\mathbf{1}$, $\mathbf{1}^\mathsf{T} I = 0$ for some $\gamma \in \mathbb{C}$. Then †

$$V^\Delta = \Gamma V = -\frac{1}{3}\Gamma\Gamma^\dagger z^\Delta\left(\mathbb{I} - \frac{1}{\zeta}\mathbf{1}\tilde{z}^{\Delta\mathsf{T}}\right)\Gamma I$$

Since $\mathbf{1}^\mathsf{T} I = 0$, there exists I^Δ such that $I = -\Gamma^\mathsf{T} I^\Delta$. Hence

$$V^\Delta = \Gamma\Gamma^\dagger z^\Delta \left(\mathbb{I} - \frac{1}{\zeta}\mathbf{1}\tilde{z}^{\Delta\mathsf{T}}\right)\Gamma\Gamma^\mathsf{T} I^\Delta = \underbrace{z^\Delta\left(\mathbb{I} - \frac{1}{\zeta}\mathbf{1}\mathbf{1}^\mathsf{T} z^\Delta\right)}_{z'^\Delta} I^\Delta$$

where we have used $\Gamma\Gamma^\dagger = \mathbb{I} - \frac{1}{3}\mathbf{1}\mathbf{1}^\mathsf{T}$ from Theorem 14.2. As before, z'^Δ must be a diagonal matrix to be a valid three-phase impedance. This is the case if and only if $z^\Delta \mathbf{1}\left(\mathbf{1}^\mathsf{T} z^\Delta I^\Delta\right) = 0$ in which case $V^\Delta = z^\Delta I^\Delta$, as desired. □

Hence given a V, I is uniquely determined by (14.27a) and given an I with $\mathbf{1}^\mathsf{T} I = 0$, V is determined by (14.27b) up to a reference voltage specified by γ. These equations allow us to relate terminal power injection s to V or to I as:

$$s = \operatorname{diag}\left(VI^\mathsf{H}\right) = -\operatorname{diag}\left(VV^\mathsf{H} Y^{\Delta\mathsf{H}}\right) \tag{14.27c}$$

$$s = \operatorname{diag}\left(VI^\mathsf{H}\right) = -\operatorname{diag}\left(Z^\Delta II^\mathsf{H}\right) + \gamma\bar{I} \tag{14.27d}$$

Balanced impedance. When the impedance is balanced, i.e., $z^{ab} = z^{bc} = z^{ca}$ then (Exercise 14.18)

$$Z^\Delta = \frac{z^{ab}}{3}\left(\mathbb{I} - \frac{1}{3}\mathbf{1}\mathbf{1}^\mathsf{T}\right)$$

i.e., Z^Δ is not diagonal and the off-diagonal entries will couple voltages and currents in different phases. As we will see in Chapter 16.4.1, in this case, one can perform a similarity transformation using the unitary matrix F to what is called the *sequence coordinate* as explained in Chapter 14.2.2. In the sequence coordinate, the transformed impedance $\tilde{Z}^\Delta$, called the sequence impedance, is diagonal:

$$\tilde{Z}^\Delta = \frac{z^{ab}}{3}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

This leads to decoupled voltages and currents in the sequence coordinate called symmetrical components. The zero-sequence component (first row and column of $\tilde{Z}^\Delta$) is zero, reflecting the fact that $I^a + I^b + I^c = 0$ in a Δ configuration since there is no neutral line. The decoupled relation between the sequence voltages, currents and impedances can be interpreted as defining separate sequence networks that can be analyzed independently.

Remark 14.9 (Phase decoupling)**.** Determine conditions under which phases become decoupled (Exercise 14.19). □

14.3.5 Δ-Y transformation

Ideal voltage source $\left(E^{\Delta},\gamma\right)$.

The terminal voltage of an ideal Δ-configured voltage source $\left(E^{\Delta},\gamma\right)$ with zero internal impedance $z^{\Delta}=0$ is, from (14.21b):

$$V \;=\; \frac{1}{3}\Gamma^{\mathsf{T}}E^{\Delta} + \gamma\mathbf{1}, \qquad\qquad \mathbf{1}^{\mathsf{T}}I \;=\; 0$$

where γ is fixed by a given reference voltage. The terminal voltage of an ideal Y-configured voltage source $\left(E^{Y},V^{n}\right)$ with zero internal impedance $z^{Y}=0$ is, from (14.13a):

$$V \;=\; E^{Y} + V^{n}\,\mathbf{1}, \qquad\qquad \mathbf{1}^{\mathsf{T}}I \;=\; -I^{n}$$

Hence the Y equivalent of an ideal voltage source $\left(E^{\Delta},\gamma\right)$, not necessarily balanced, is given by

$$E^{Y} \;:=\; \frac{1}{3}\Gamma^{\mathsf{T}}E^{\Delta}, \qquad V^{n} \;:=\; \gamma, \qquad \text{no neutral line so that } \; I^{n} \;:=\; 0$$

Note that this does not satisfy assumption C14.1 since the neutral is not grounded unless $\gamma=0$. If E^{Δ} is balanced then $\Gamma^{\mathsf{T}}E^{\Delta}=(1-\alpha^{2})E^{\Delta}=\sqrt{3}\,e^{-\mathbf{i}\pi/6}E^{\Delta}$ (by Corollary 1.3) and E^{Y} reduces to the expression (1.32a) derived in Chapter 1.2.4 for balanced systems:

$$E^{Y} \;=\; \frac{1}{\sqrt{3}\,e^{\mathbf{i}\pi/6}}E^{\Delta}, \qquad V^{n} \;:=\; \gamma, \qquad \text{no neutral line so that } \; I^{n} \;:=\; 0$$

For a non-ideal Δ-configured voltage source $\left(E^{\Delta},z^{\Delta},\gamma\right)$, its terminal voltage is, from (14.21b):

$$V \;=\; \hat{\Gamma}^{\mathsf{T}}E^{\Delta} \;-\; Z^{\Delta}I \;+\; \gamma\mathbf{1}$$

where

$$\hat{\Gamma} \;:=\; \frac{1}{3}\Gamma^{\mathsf{T}}\left(\mathbb{I}-\frac{1}{\zeta}\tilde{z}^{\Delta}\mathbf{1}^{\mathsf{T}}\right), \qquad\qquad Z^{\Delta} \;:=\; \frac{1}{9}\Gamma^{\mathsf{T}}z^{\Delta}\left(\mathbb{I}-\frac{1}{\zeta}\mathbf{1}\,\tilde{z}^{\Delta\mathsf{T}}\right)\Gamma$$

It generally does not have a Y equivalent. Indeed, since the Y equivalent needs to be ungrounded so that $\mathbf{1}^{\mathsf{T}}I=0$, its external model is $V \;=\; E^{Y}-z^{Y}I+V^{n}\mathbf{1}$ from (14.13a). In general the effective impedance Z^{Δ} is not diagonal and hence may not be interpreted as an internal series impedance matrix z^{Y} of an Y-configured source, even if the impedance is balanced $z^{\Delta}:=z^{ab}\mathbb{I}$ (in which case $Z^{\Delta}=\frac{z^{ab}}{3}\left(\mathbb{I}-\frac{1}{3}\mathbf{1}\mathbf{1}^{\mathsf{T}}\right)$).

Remark 14.10 (Y-equivalent with equal line-to-line voltage)**.** Given a general Δ-configured device with internal voltage V^{Δ}, its equivalent line-to-neutral voltage is defined in [43, p.204] to be

$$V^{Y} \;:=\; \frac{1}{3}\begin{bmatrix}2 & 1 & 0\\ 0 & 2 & 1\\ 1 & 0 & 2\end{bmatrix}V^{\Delta} \tag{14.28}$$

This definition is the same as the Y-equivalent of an ideal voltage source V^Δ derived above with a particular choice of the neutral voltage:

$$V^Y := \frac{1}{3}\Gamma^{\mathsf{T}} V^\Delta, \qquad\qquad V^n := \gamma = 0$$

in the sense that they have the same line-to-line voltages.

To see this, recall that the line-to-line voltage $\tilde{V}^Y$ (not the terminal voltage) of a Y-configured device with internal voltage V^Y is $\tilde{V}^Y = \Gamma V^Y$. If it is equivalent to the given V^Δ then $V^\Delta = \tilde{V}^Y = \Gamma V^Y$. Theorem 14.2 then implies

$$V^Y = \frac{1}{3}\Gamma^{\mathsf{T}} V^\Delta + \gamma \mathbf{1} \qquad\qquad \text{for any } \gamma \in \mathbb{C}$$

Here γ being arbitrary means that the Δ-configured device has an arbitrary zero-sequence terminal voltage and its Y-equivalent has an arbitrary neutral voltage. Take $\gamma := 0$. Since $\mathbf{1}^{\mathsf{T}} V^\Delta = \mathbf{1}^{\mathsf{T}} \left(\Gamma V^Y\right) = 0$ we can add $\frac{1}{3}\mathbf{1}^{\mathsf{T}} V^\Delta$ to Y^Y to get

$$V^Y = \frac{1}{3}\left(\Gamma^{\mathsf{T}} + \mathbf{1}\mathbf{1}^{\mathsf{T}}\right) V^\Delta = \frac{1}{3}\left(\begin{bmatrix} 1 & 0 & -1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}\right) V^\Delta = \frac{1}{3}\begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 1 & 0 & 2 \end{bmatrix} V^\Delta$$

The model (14.28) is applicable only if the zero-sequence voltage $\gamma := \frac{1}{3}\mathbf{1}^{\mathsf{T}} V$ of the given Δ-configured device is zero. Otherwise its Y-equivalent must have a nonzero neutral voltage $V^j = \gamma$. □

Ideal current source J^Δ.

An ideal Δ-configured current source J^Δ has an external model of $I = -\Gamma^{\mathsf{T}} J^\Delta$. Note that $\mathbf{1}^{\mathsf{T}} I = 0$. The external model of a Y-configured current source is $I = -J^Y$, $\mathbf{1}^{\mathsf{T}} I = -I^n$. Hence the Y equivalent is

$$J^Y = \Gamma^{\mathsf{T}} J^\Delta, \qquad\qquad \text{no neutral line so that } I^n := 0$$

If J^Δ is balanced then Corollary 1.3 implies

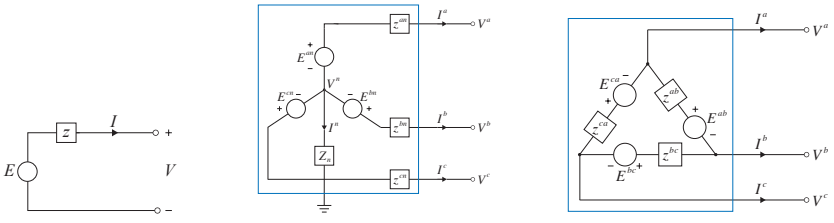
$$J^Y = (1-\alpha^2) J^\Delta = \frac{\sqrt{3}}{e^{\mathbf{i}\pi/6}} J^\Delta$$

the same expression (1.32a) for balanced systems.

14.3.6 Comparison with single-phase devices

Assume C14.1 holds, i.e., neutrals are grounded and voltages are defined with respect to the ground. We compare the external models of three-phase devices to those of their single-phase counterparts. As we will see they are structurally the same, except for the Δ-configured power source.

Voltage source.



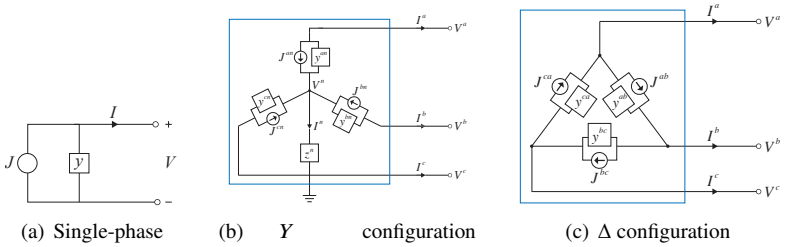
(a) Single-phase (b) Y configuration (grounded) (c) Δ configuration

Figure 14.9 Comparison of single-phase and three-phase voltage sources.

Figure 14.9 shows a single-phase voltage source specified by an internal voltage E and a series impedance z and the three-phase voltage sources in Y and Δ configurations studied in this section. Their external models are, from (14.13b) and (14.21b):

$$\begin{aligned}
&\text{single-phase:} && V = E - zI \\
&Y\text{-configuration:} && V = E^Y - Z^Y I, && Z^Y := z^Y + z^n \mathbf{1}\mathbf{1}^\mathsf{T} \\
&\Delta\text{-configuration:} && V = \hat{\Gamma} E^\Delta - Z^\Delta I + \gamma \mathbf{1}, && \mathbf{1}^\mathsf{T} I = 0
\end{aligned}$$

Current source.



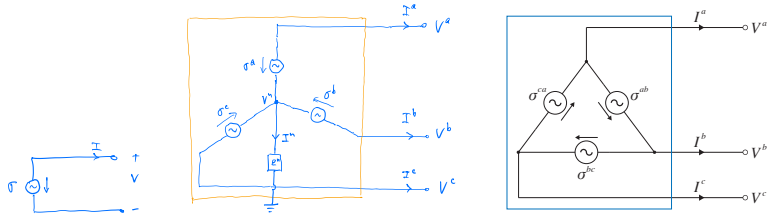
(a) Single-phase (b) Y configuration (grounded) (c) Δ configuration

Figure 14.10 Comparison of single-phase and three-phase current sources.

Figure 14.10 shows a single-phase current source specified by an internal current J and a shunt admittance y and the three-phase current sources in Y and Δ configurations studied in this section. Their external models are, from (14.15b) and (14.23a):

$$\begin{aligned}
&\text{single-phase:} && I = -(J + yV) \\
&Y\text{-configuration:} && I = -A\left(J^Y + y^Y V\right), && A := \mathbb{I} - \frac{z^n}{1 + z^n \left(\mathbf{1}^\mathsf{T} y^Y \mathbf{1}\right)} y^Y \mathbf{1}\mathbf{1}^\mathsf{T} \\
&\Delta\text{-configuration:} && I = -\left(\Gamma^\mathsf{T} J^\Delta + Y^\Delta V\right), && Y^\Delta := \Gamma^\mathsf{T} y^\Delta \Gamma
\end{aligned}$$

Power source.



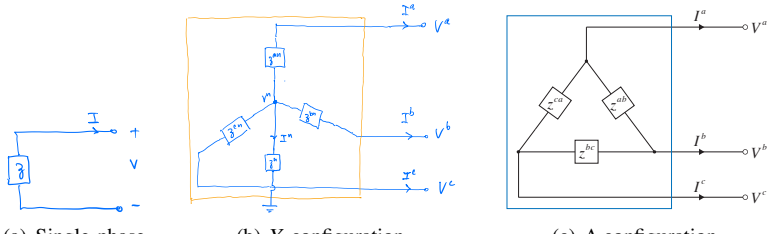
(a) Single-phase (b) Y configuration (c) Δ configuration

Figure 14.11 Comparison of single-phase and three-phase power sources.

Figure 14.11 shows a single-phase power source specified by an internal power σ and the three-phase power sources in Y and Δ configurations studied in this section. Their external models are, from (14.17c) and (14.25d):

$$\begin{aligned}
&\text{single-phase:} && s = -\sigma \\
&Y\text{-configuration:} && s = -\left(\sigma^Y + z^n \left(\overline{I} I^{\mathsf{T}}\right)\mathbf{1}\right) \\
&\Delta\text{-configuration:} && s = -\operatorname{diag}\left(V I^{\Delta \mathsf{H}} \Gamma\right), \qquad \sigma^\Delta = \operatorname{diag}\left(\Gamma V I^{\Delta \mathsf{H}}\right)
\end{aligned}$$

Impedance.



(a) Single-phase (b) Y configuration (c) Δ configuration

Figure 14.12 Comparison of single-phase and three-phase impedances.

Figure 14.12 shows a single-phase impedance specified by z and the three-phase power sources in Y and Δ configurations studied in this section. Their external models are, from (14.19b) and (14.27a):

$$\begin{aligned}
&\text{single-phase:} && V = -zI \\
&Y\text{-configuration:} && V = -Z^Y I, && Z^Y := z^Y + z^n \mathbf{1}\mathbf{1}^{\mathsf{T}} \\
&\Delta\text{-configuration:} && I = -Y^\Delta V, && Y^\Delta := \Gamma^{\mathsf{T}} y^\Delta \Gamma
\end{aligned}$$

14.3.7 Summary

The external models of three-phase devices are summarized in Table 14.2 and will be used to compose network models in Chapters 16 and 17.

Device	Y configuration			Δ configuration		
	Specification	Internal	External	Specification	Internal	External
Voltage source	$\left(E^Y, z^Y, z^n\right)$	(14.12)	(14.13)	$\left(E^\Delta, z^\Delta\right)$	(14.20)	(14.21)
Current source	$\left(J^Y, y^Y, z^n\right)$	(14.14)	(14.15)	$\left(J^\Delta, y^\Delta\right)$	(14.22)	(14.23)
Power source	$\left(\sigma^Y, z^n\right)$	(14.16)	(14.17)	σ^Δ	(14.24)	(14.25)
Impedance	$\left(z^Y, z^n\right)$	(14.18)	(14.19)	z^Δ	(14.26)	(14.27)
Line (3-wire model)		(15.8)				

Table 14.2 Specification, internal and external models of three-phase devices.

When the devices are ideal these models reduce to a simpler form summarized in Tables 14.3 and 14.4. The internal models of ideal devices are:

1 Ideal voltage source $E^{Y/\Delta}$:

$$V^{Y/\Delta} = E^{Y/\Delta}, \qquad s^{Y/\Delta} = \operatorname{diag}\left(E^{Y/\Delta}\left(I^{Y/\Delta}\right)^{\mathsf{H}}\right) \tag{14.29a}$$

2 Ideal current source $J^{Y/\Delta}$:

$$I^{Y/\Delta} = J^{Y/\Delta}, \qquad s^{Y/\Delta} = \operatorname{diag}\left(V^{Y/\Delta}\left(J^{Y/\Delta}\right)^{\mathsf{H}}\right) \tag{14.29b}$$

3 Ideal power source $\sigma^{Y/\Delta}$:

$$s^{Y/\Delta} = \sigma^{Y/\Delta}, \qquad \sigma^{Y/\Delta} = \operatorname{diag}\left(V^{Y/\Delta}\left(I^{Y/\Delta}\right)^{\mathsf{H}}\right) \tag{14.29c}$$

4 Impedance $z^{Y/\Delta}$:

$$V^{Y/\Delta} = z^{Y/\Delta} I^{Y/\Delta}, \qquad s^{Y/\Delta} = \operatorname{diag}\left(V^{Y/\Delta}\left(I^{Y/\Delta}\right)^{\mathsf{H}}\right) \tag{14.29d}$$

In each case the internal specification of the three-phase device fixes one of the terminal variables (V, I, s) and the relation between the remaining variables characterizes its external behavior. In the rest of this book we often assume sources are ideal and characterized by Tables 14.3 and 14.4 (see Chapter 15.1.4 for a justification).

Consider a network of three-phase voltage sources, current sources, power sources, and impedances connected by three-phase lines and transformers. A power flow problem typically specifies a set of these devices and the objective is to determine other

Device	Assumption	Y configuration	
Voltage source	$z^n = 0,\ z^Y = 0$	$V = E^Y + \gamma \mathbf{1}$	$s = \text{diag}\left(E^Y I^{\mathsf{H}}\right) + \gamma \bar{I}$
Current source	$z^n = 0,\ y^Y = 0$	$I = -J^Y$	$s = -\text{diag}\left(V J^{Y\mathsf{H}}\right)$
Power source	$z^n = 0$	$\text{diag}\left(I^{\mathsf{H}}\right)(V - \gamma \mathbf{1}) = -\sigma$	$s = -\sigma^Y + \gamma \bar{I}$
Impedance	$z^n = 0$	$V = -z^Y I + \gamma \mathbf{1}$	$s = -\text{diag}\left(V\,(V - \gamma \mathbf{1})^{\mathsf{H}}\, y^{Y\mathsf{H}}\right)$

Table 14.3 External models of ideal single-terminal devices in Y configuration. The quantity $\gamma := V^n$ is the neutral voltage. If all neutrals are directly grounded and voltages are defined with respect to the ground, then $\gamma := V^n = 0$ for all Y-configured devices.

Device	Assumption	Δ configuration	
Voltage source	$z^\Delta = 0,\ \mathbf{1}^{\mathsf{T}} E^\Delta = 0$	$V = \Gamma^\dagger E^\Delta + \gamma \mathbf{1},\ \mathbf{1}^{\mathsf{T}} I = 0$	$s = \text{diag}\left(\Gamma^\dagger E^\Delta I^{\mathsf{H}}\right) + \gamma \bar{I}$
Current source	$y^\Delta = 0$	$I = -\Gamma^{\mathsf{T}} J^\Delta$	$s = -\text{diag}\left(V J^{\Delta\mathsf{H}} \Gamma\right)$
Power source			$\sigma^\Delta = \text{diag}\left(\Gamma V I^{\Delta\mathsf{H}}\right)$
		$\mathbf{1}^{\mathsf{T}} I = 0$	$\sigma^\Delta = -\text{diag}\left(\Gamma^{\mathsf{T}\dagger}\left(V I^{\mathsf{H}}\right)\Gamma^{\mathsf{T}}\right) + \bar{\beta}\Gamma V$
			$s = \text{diag}\left(V I^{\Delta\mathsf{H}} \Gamma\right)$
		$\mathbf{1}^{\mathsf{T}} V^\Delta = 0$	$s = -\text{diag}\left(\Gamma^\dagger\left(V^\Delta I^{\Delta\mathsf{H}}\right)\Gamma\right) - \gamma \Gamma^{\mathsf{T}} \bar{I}^\Delta$
Impedance		$I = -Y^\Delta V$	$s = -\text{diag}\left(V V^{\mathsf{H}} Y^{\Delta\mathsf{H}}\right)$
		$V = -Z^\Delta I + \gamma \mathbf{1},\ \mathbf{1}^{\mathsf{T}} I = 0$	$s = -\text{diag}\left(Z^\Delta I I^{\mathsf{H}}\right) + \gamma \bar{I}$

Table 14.4 External models of ideal single-terminal devices in Δ configuration. The quantity $\gamma := \frac{1}{3}\mathbf{1}^{\mathsf{T}} V$ is the zero-sequence voltage of V and $\beta := \frac{1}{3}\mathbf{1}^{\mathsf{T}} I^\Delta$ is the zero-sequence current of I^Δ.

voltages, currents, and powers on the network. The specification of these devices include not only internal voltages, currents, or powers, but also some of the zero-sequence quantities (γ, β). We will clarify in Chapter 16.2 the parameters that should be specified versus variables to be computed of the external models in Tables 14.3 and 14.4.

14.4 Voltage regulators

14.5 Bibliographical notes

The concept of symmetrical component is described in another seminal paper [152] by C. L. Fortescue to simplify the analysis of unbalanced operation of a multiphase system. The use of symmetrical components for fault current analysis is explained in

e.g. [155] which also proposes a different transformation called $(\alpha,\beta,0)$ components. The paper [153] explains that Fortescue's transformation matrix as a particular choice of orthogonal basis for three-dimensional vectors over the complex field (the similarity transformation matrix F in Chapter 14.2.2 is the normalized version of Fortescue's original matrix so that the basis are orthonormal). It shows that other well-known transformations such as those of Clarke, Concordia, Kimbark, and Park can be obtained from Forescue's matrix through elementary row and column transformations and have different advantages and disadvantages mostly for fault analysis. Park transformation [154] is applicable not only to steady state voltage and current phasors, but also to instantaneous voltages, currents, and flux linkages in modeling synchronous machines.

As we will see in Chapter 16 a three-phase network has a single-phase equivalent circuit where the network equations have the form as a single-phase network. The main difference with a single-phase network is the models of three-phase devices in the equivalent circuit, such as models for constant-power devices [5, Chapter 11], loads and voltage regulars [43], as we have studied in Chapter 14.3, as well as three-phase lines and transformers, to be studied in Chapter 15. See also [156, Chapter 3] for comprehensive models of three-phase components including distribution lines, transformers and switches.

14.6 Problems

Chapter 14.2.

Exercise 14.1 (Proof of Theorem 14.2)**.** Let

$$\Gamma := \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{bmatrix}, \qquad \Gamma^{\mathsf{T}} := \begin{bmatrix} 1 & 0 & -1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}$$

Prove Theorem 14.2:

1 The null spaces of Γ and Γ^{T} are both $\mathrm{span}(1,1,1)$.
2 Their pseudo-inverses are

$$\Gamma^{\dagger} = \frac{1}{3}\Gamma^{\mathsf{T}}, \qquad \Gamma^{\mathsf{T}\dagger} = \frac{1}{3}\Gamma$$

3 Consider $\Gamma x = b$. If $\mathbf{1}^{\mathsf{T}} b = 0$ then the solutions x are given by $x = \Gamma^{\dagger} b + \beta\mathbf{1}$ for all $\beta \in \mathbb{C}^3$.
4 Consider $\Gamma^{\mathsf{T}} x = b$. If $\mathbf{1}^{\mathsf{T}} b = 0$ then the solutions x are given by $x = \Gamma^{\mathsf{T}\dagger} b + \beta\mathbf{1}$ for all $\beta \in \mathbb{C}^3$.
5 $\Gamma\Gamma^{\dagger} = \Gamma^{\dagger}\Gamma = \frac{1}{3}\Gamma\Gamma^{\mathsf{T}} = \frac{1}{3}\Gamma^{\mathsf{T}}\Gamma = \mathbb{I} - \frac{1}{3}\mathbf{1}\mathbf{1}^{\mathsf{T}}$ where $\mathbb{I}$ is the identity matrix of appropriate size.

Exercise 14.2. Use $\Gamma^\dagger = \frac{1}{3}\Gamma^\mathsf{T}$ (Theorem 14.2) to verify the four defining properties of pseudo-inverse of Γ:

1 $(\Gamma\Gamma^\dagger)\Gamma = \Gamma$.
2 $\Gamma^\dagger(\Gamma\Gamma^\dagger) = \Gamma^\dagger$.
3 $\Gamma\Gamma^\dagger$ is Hermitian.
4 $\Gamma^\dagger\Gamma$ is Hermitian.

Exercise 14.3. Suppose $I = -\Gamma^\mathsf{T} I^\Delta$. Show that $VI^\mathsf{H}\left(\Gamma^\dagger\Gamma\right) = VI^\mathsf{H}$.

Chapter 14.3.1.

Exercise 14.4 (Terminal power s)**.** Consider the three-phase voltage source serving a three-phase impedance load shown in Figure 14.13. Both the source and the load are

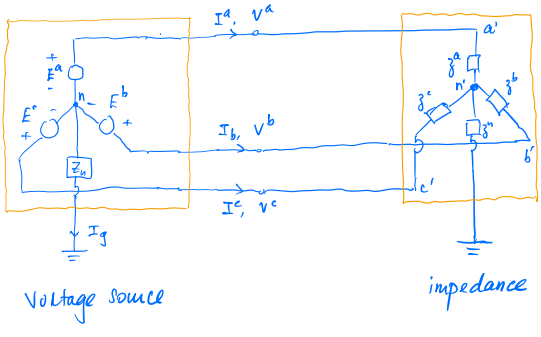


Figure 14.13 Terminal power s and internal power s^Y.

grounded. Suppose the terminal voltage V is defined with respect to the ground. The terminal current I^a flows from terminal a of the source to the load and returns from the ground, and $s^a := V^a I^{a\mathsf{H}}$ is the power delivered across terminal a and the ground. Relate the terminal power $\mathbf{1}^\mathsf{T} s := V^a I^{a\mathsf{H}} + V^b I^{b\mathsf{H}} + V^c I^{c\mathsf{H}}$ and the internal power $\mathbf{1}^\mathsf{T} s^Y$ for both the voltage source and the impedance.

Exercise 14.5 (Terminal power s)**.** Repeat Exercise 14.4 but for the case where the neutrals are not grounded, as shown in Figure 14.14. All voltages are defined with respect to an arbitrary but common reference point, e.g., the ground.

Exercise 14.6 (Total powers)**.** Show that $\mathbf{1}^\mathsf{T}\mathrm{diag}\left(\Gamma\left(VI^\mathsf{H}\right)\Gamma^\dagger\right) = \mathbf{1}^\mathsf{T}\mathrm{diag}\left(VI^\mathsf{H}\right)$ and hence the total internal and terminal powers are equal, i.e., $\mathbf{1}^\mathsf{T} s^\Delta = \mathbf{1}^\mathsf{T} s$.

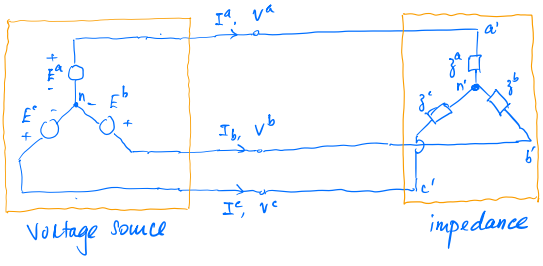


Figure 14.14 Terminal power s and internal power s^Y.

Chapter 14.3.3.

Exercise 14.7 (Y-configured voltage source)**.** Compute the inverse of $Z^Y := z^Y + z_n \mathbf{1}\mathbf{1}^\mathsf{T}$ in (14.13c) using the matrix inversion formula.

Exercise 14.8 (Y-configured current source)**.** Consider the current source in Figure 14.7(b). Derive (14.14) for internal power s^Y and s^n.

Exercise 14.9 (Y-configured current source)**.** Consider the current source in Figure 14.7(b). Suppose assumption C14.1 holds. Derive (14.15b):

$$V = -\left(z^Y J^Y + Z^Y I\right), \qquad I = -A\left(J^Y + y^Y V\right)$$

where

$$z^Y := \left(y^Y\right)^{-1}, \quad Z^Y := z^Y + z^n \mathbf{1}\mathbf{1}^\mathsf{T}, \quad A := \mathbb{I} - \frac{z^n}{1 + z^n\left(\mathbf{1}^\mathsf{T} y^Y \mathbf{1}\right)} y^Y \mathbf{1}\mathbf{1}^\mathsf{T}$$

assuming Z^Y is invertible.

Exercise 14.10 (Y-configured power device)**.** Suppose all voltages are defined with respect to the ground, so that $V^n = -z^n\left(\mathbf{1}^\mathsf{T} I\right)$. Derive (14.17b).

Exercise 14.11 (Y-configured impedance)**.** Consider a three-phase load in Y configuration specified by a series impedance matrix Z^Y:

$$V := \begin{bmatrix} V_{ag} \\ V_{bg} \\ V_{cg} \end{bmatrix} = \begin{bmatrix} z_a + z_n & z_n & z_n \\ z_n & z_b + z_n & z_n \\ z_n & z_n & z_c + z_n \end{bmatrix} \begin{bmatrix} I_a \\ I_b \\ I_c \end{bmatrix}$$

Show that if V is balanced and $z_a = z_b = z_c$ then the neutral current $I_n = 0$ and the phases are decoupled.

Chapter 14.3.4.

Exercise 14.12 (Voltage source in Δ configuration)**.** Consider the voltage source in Figure 14.8(a). Let $V^\Delta = \Gamma V$.

1 Show that $\mathbf{1}^\mathsf{T} I = 0$ implies $\mathbf{1}^\mathsf{T}\left(E^\Delta - z^\Delta \Gamma^{\mathsf{T}\dagger} I\right) = 0$.
2 Show that the converse is not true.

Exercise 14.13 (Voltage source in Δ configuration)**.** Suppose A is a complex symmetric matrix A with zero row sums. Show that its pseudo-inverse $A^\dagger$ is also complex symmetric with zero row sums. (Hint: Use Takagi factorization for complex symmetric matrices in Theorem A.17 of Appendix A.6.)

Exercise 14.14 (Voltage source in Δ configuration)**.** Prove Theorem 14.3: Given the conversion rules $V^\Delta = \Gamma V$ and $I = -\Gamma^\mathsf{T} I^\Delta$ between the terminal and internal voltages and currents, the following are equivalent:

1 Internal model: $V^\Delta = E^\Delta + z^\Delta I^\Delta$ and hence $\mathbf{1}^\mathsf{T}\left(E^\Delta + z^\Delta I^\Delta\right) = 0$.
2 External model: $I = \left(\Gamma^\mathsf{T} y^\Delta\right) E^\Delta - Y^\Delta V$ where $Y^\Delta := \Gamma^\mathsf{T} y^\Delta \Gamma$.
3 External model: $V = \hat{\Gamma} E^\Delta - Z^\Delta I + \gamma \mathbf{1}$, $\mathbf{1}^\mathsf{T} I = 0$ for some $\gamma \in \mathbb{C}$ where

$$\hat{\Gamma} := \frac{1}{3}\Gamma^\mathsf{T}\left(\mathbb{I} - \frac{1}{\zeta}\tilde{z}^\Delta \mathbf{1}^\mathsf{T}\right), \qquad Z^\Delta := \frac{1}{9}\Gamma^\mathsf{T} z^\Delta \left(\mathbb{I} - \frac{1}{\zeta}\mathbf{1}\tilde{z}^{\Delta\mathsf{T}}\right)\Gamma$$

(Hint: See the proof of Theorem 14.4.)

Exercise 14.15 (Voltage source in Δ configuration)**.** Consider (14.25a), reproduced here:

$$\sigma^\Delta = -\frac{1}{3}\operatorname{diag}\left(\Gamma\left(VI^\mathsf{H}\right)\Gamma^\mathsf{T}\right) + \overline{\beta}\Gamma V, \qquad \mathbf{1}^\mathsf{T} I = 0$$

Given any terminal voltage V, show that I and β are uniquely determined in terms of V and σ^Δ.

Exercise 14.16 (Voltage source in Δ configuration)**.** Consider the model of a constant-power source (14.25a), reproduced here:

$$\sigma^\Delta = -\frac{1}{3}\operatorname{diag}\left(\Gamma\left(VI^\mathsf{H}\right)\Gamma^\mathsf{T}\right) + \overline{\beta}\Gamma V, \qquad \mathbf{1}^\mathsf{T} I = 0, \qquad \beta \in \mathbb{C}$$

Given a terminal current I with $\mathbf{1}^\mathsf{T} I = 0$, show that the zero-sequence current $\beta := \frac{1}{3}\mathbf{1}^\mathsf{T} I^\Delta$ can take two values.

Exercise 14.17 (Total power in Δ)**.** Consider a power source with internal power $\sigma^{\Delta} := (\sigma^{ab}, \sigma^{bc}, \sigma^{ca})$ in Δ configuration. Show that (the negative of) its total internal power is equal to its total terminal power, i.e., $\mathbf{1}^{\mathsf{T}} s = -\mathbf{1}^{\mathsf{T}} \sigma^{\Delta}$.

Exercise 14.18 (Balanced impedance z^{Δ})**.** Consider a Δ-configured impedance z^{Δ} whose external equivalent is (from (14.27b)):

$$Z^{\Delta} := \frac{1}{9}\Gamma^{\mathsf{T}} \underbrace{z^{\Delta}\left(\mathbb{I} - \frac{1}{\zeta}\mathbf{1}\hat{z}^{\Delta\mathsf{T}}\right)}_{\hat{z}^{\Delta}} \Gamma$$

If the impedance is balanced, i.e., $z^{ab} = z^{bc} = z^{ca}$, show that

$$Z^{\Delta} = \frac{z^{ab}}{3}\left(\mathbb{I} - \frac{1}{3}\mathbf{1}\mathbf{1}^{\mathsf{T}}\right)$$

Exercise 14.19 (Devices in Δ configuration)**.** Show that the phases are decoupled, i.e., phase a variables (s^a, V^a, I^a) do not depend on variables in phases b and c, if the terminal currents are balanced $I^a + I^b + I^c = 0$ and the terminal voltages $V^a + V^b + V^c = 0$ for the four types of devices in Δ configuration discussed in Chapter 14.3.4.

Chapter 14.3.5.

Exercise 14.20 (Δ-Y transformation)**.** Show that the external behavior of a symmetric non-ideal voltage source $\left(E^{\Delta}, z^{ab}\mathbb{I}\right)$ with identical series impedance $z^{\Delta} := z^{ab}\mathbb{I}$ and zero-sequence voltage $\gamma = 0$ is equivalent to a non-ideal Y-configured voltage source $\left(E^{Y}, z^{Y}, z^{n}\right)$ whose neutral is grounded through an impedance z^{n} with:

$$E^{Y} := \frac{1}{3}\Gamma^{\mathsf{T}} E^{\Delta}, \qquad z^{Y} := \frac{z^{ab}}{3}\mathbb{I}, \qquad z^{n} := -\frac{z^{ab}}{9}$$

under assumption C14.1.

Exercise 14.21 (Δ-Y transformation)**.** Consider a symmetric non-ideal current source $\left(J^{\Delta}, y^{ab}\mathbb{I}\right)$ with identical shunt admittance $y^{\Delta} := y^{ab}\mathbb{I}$. Show that it cannot be equivalent to a non-ideal Y-configured current source $\left(J^{Y}, y^{Y}, z^{n}\right)$ under assumption C14.1.

15 Component models, II: line and transformers

In this chapter we continue the modeling of three-phase components. In Chapter 15.1 we model a three-phase transmission or distribution line. In Chapter 15.2 we extend the simplified model of transformers of Chapter 3.1.4 from single-phase to three-phase setting. In Chapter 15.3 we extend the transformer model based on unitary voltage network of Chapter 3.1.5 from single-phase to three-phase setting. In Chapter 15.4 we explain how to identify model parameters from measurements. We will use these component models in Chapters 16 and 17 to construct network models and study unbalanced three-phase analysis.

15.1 Three-phase transmission or distribution line models

As explained Chapter 2.1 the electromagnetic interactions among the electric charges in wires of different phases couple the voltages on and currents in these wires. The relation between the voltages and currents in these phases can be modeled by a linear mapping that depends on the line characteristics (resistances, inductances, capacitances).

15.1.1 Review: single-phase model

The linear mapping becomes decoupled when the phases are balanced, leading to a per-phase model of a transmission or distribution line as a two-terminal device specified by a Π-equivalent circuit $(y^s_{jk}, y^m_{jk}, y^m_{kj})$, as explained in Chapter 2.2.2. The terminal (or bus) voltages (V_j, V_k) and sending-end line currents (I_{jk}, I_{kj}) on this two-terminal device describes the end-to-end behavior of the line. They are linearly related according to Kirchhoff's and Ohm's laws:

$$I_{jk} = y^s_{jk}\left(V_j - V_k\right) + y^m_{jk} V_j, \qquad I_{kj} = y^s_{kj}\left(V_k - V_j\right) + y^m_{kj} V_k \quad (15.1a)$$

For a transmission or distribution line, $y^s_{jk} = y^s_{kj}$. The terms $y^m_{jk} V_j$ and $y^m_{kj} V_k$ assume that the shunt admittances connect the buses j and k both to the common reference point for terminal voltages, e.g., the ground. The sending-end line power $\left(S_{jk}, S_{kj}\right)$ is

related to (V_j, V_k) by

$$S_{jk} = \left(y^s_{jk}\right)^{\mathsf{H}} V_j \left(V_j - V_k\right)^{\mathsf{H}} + \left(y^m_{jk}\right)^{\mathsf{H}} V_j V_j^{\mathsf{H}} \tag{15.1b}$$

$$S_{kj} = \left(y^s_{kj}\right)^{\mathsf{H}} V_k \left(V_k - V_j\right)^{\mathsf{H}} + \left(y^m_{kj}\right)^{\mathsf{H}} V_k V_k^{\mathsf{H}} \tag{15.1c}$$

When ($y^s_{jk} = y^s_{kj}$ and) the shunt admittances are zero, i.e., $y^m_{jk} = y^m_{kj} = 0$, then $I_{jk} = -I_{kj}$ and this relation reduces to

$$V_j \; - \; V_k = z^s_{jk} \, I_{jk} \tag{15.1d}$$

where $z^s_{jk} := \left(y^s_{jk}\right)^{-1}$ is the series impedance of the line. We now extend these relations to an unbalanced three-phase transmission or distribution line.

15.1.2 Four-wire three-phase model

A three-phase line has three wires one for each phase a, b, c. It may also have a neutral wire which may be grounded at one or both ends if the device connected to that end of the line is in Y configuration. Consider then a four-wire three-phase line where the total current $i^a(t) + i^b(t) + i^c(t)$ and the total charge $q^a(t) + q^b(t) + q^c(t)$ may be nonzero and they flow through the neutral wire (if present) and the earth return. The effect of neutral or earth return on the impedance of a transmission line depends on details such as how many neutral wires are present, whether they are grounded along the lines at regular spacing, etc.

To build intuition we first omit line charging. In this case the three-phase voltages and currents are related by a series impedance matrix, similar to (15.1d) for a single-phase system. We then incorporate the effect of line charging by including shunt admittances to obtain a model that generalizes (15.1a) to a three-phase system.

Without shunt admittances.

Consider a four-wire three-phase line with a neutral wire. The voltage between one end of a wire to the other end depends linearly on the current in each of the four wires. Let $\hat{V}_j := \left(V^a_j, V^b_j, V^c_j, V^n_j\right)$ and $\hat{V}_k := \left(V^a_k, V^b_k, V^c_k, V^n_k\right)$ be the *terminal (or nodal or bus)* voltages at terminals j and k respectively of the phase and neutral wire (j, k), with respect to an arbitrary but common reference point, e.g., the ground. Let $\hat{I}_{jk} := \left(I^a_{jk}, I^b_{jk}, I^c_{jk}, I^n_{jk}\right)$ denote the currents in these lines. Then the four-wire three-phase line can be modeled by a *series impedance matrix*[1] $\hat{z}^s_{jk}$ that linearly relates these

[1] It is sometimes called a series *phase* impedance matrix to differentiate it from a series *sequence* impedance matrix for sequence variables; see Chapter 16.4.

voltages and currents:

$$\begin{bmatrix} V_j^a \\ V_j^b \\ V_j^c \\ V_j^n \end{bmatrix} - \begin{bmatrix} V_k^a \\ V_k^b \\ V_k^c \\ V_k^n \end{bmatrix} = \underbrace{\begin{bmatrix} \hat{z}_{jk}^{aa} & \hat{z}_{jk}^{ab} & \hat{z}_{jk}^{ac} & \hat{z}_{jk}^{an} \\ \hat{z}_{jk}^{ba} & \hat{z}_{jk}^{bb} & \hat{z}_{jk}^{bc} & \hat{z}_{jk}^{bn} \\ \hat{z}_{jk}^{ca} & \hat{z}_{jk}^{cb} & \hat{z}_{jk}^{cc} & \hat{z}_{jk}^{cn} \\ \hat{z}_{jk}^{na} & \hat{z}_{jk}^{nb} & \hat{z}_{jk}^{nc} & \hat{z}_{jk}^{nn} \end{bmatrix}}_{\hat{z}_{jk}^s} \begin{bmatrix} I_{jk}^a \\ I_{jk}^b \\ I_{jk}^c \\ I_{jk}^n \end{bmatrix} \tag{15.2a}$$

or in vector form

$$\hat{V}_j - \hat{V}_k = \hat{z}_{jk}^s \hat{I}_{jk} \tag{15.2b}$$

For example, the series impedance matrix $\hat{z}_{jk}^s$ can model an overhead three-phase line with an overhead neutral wire and earth return. Here $\hat{z}_{jk}^{\phi\phi}$ are called the *self-impedances* of phase ϕ wires, including the effect of earth return, and $\hat{z}_{jk}^{\phi\phi'}$ the *mutual impedances* between phase ϕ and phase ϕ' wires, including the effect of earth return. Their values depend on the wire materials, their lengths, distances between them, the operating frequency, and the resistivity of the earth. To relate these impedances to the physical system, suppose a voltage is applied between the phase a terminals and therefore completing the phase a circuit, while circuits of phases b, c, n are open. Then the current I_{jk}^a in the phase a wire is nonzero while all other currents $I_{jk}^{\phi} = 0$, $\phi \neq a$, so that

$$\begin{bmatrix} V_j^a \\ V_j^b \\ V_j^c \\ V_j^n \end{bmatrix} - \begin{bmatrix} V_k^a \\ V_k^b \\ V_k^c \\ V_k^n \end{bmatrix} = \begin{bmatrix} \hat{z}_{jk}^{aa} & \hat{z}_{jk}^{ab} & \hat{z}_{jk}^{ac} & \hat{z}_{jk}^{an} \\ \hat{z}_{jk}^{ba} & \hat{z}_{jk}^{bb} & \hat{z}_{jk}^{bc} & \hat{z}_{jk}^{bn} \\ \hat{z}_{jk}^{ca} & \hat{z}_{jk}^{cb} & \hat{z}_{jk}^{cc} & \hat{z}_{jk}^{cn} \\ \hat{z}_{jk}^{na} & \hat{z}_{jk}^{nb} & \hat{z}_{jk}^{nc} & \hat{z}_{jk}^{nn} \end{bmatrix} \begin{bmatrix} I_{jk}^a \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Hence the self-impedance

$$\hat{z}_{jk}^{aa} = \frac{V_j^a - V_k^a}{I_{jk}^a}$$

is the ratio of the voltage applied between the phase a terminals to the current in the phase a wire when all other circuits are open. The current I_{jk}^a induces voltages in other phases and the mutual impedance

$$\hat{z}_{jk}^{ba} = \frac{V_j^b - V_k^b}{I_{jk}^a}$$

is the ratio of the voltage induced across the phase b terminals to the phase a current when only the phase a circuit is complete.

With shunt admittances.

To incorporate the effect of line charging, let the *series admittance matrix* be $\hat{y}_{jk}^s := \left(\hat{z}_{jk}^s\right)^{-1}$, assuming $\hat{z}_{jk}^s$ is invertible. Let $\left(\hat{y}_{jk}^m, \hat{y}_{kj}^m\right)$ denote the *shunt admit-*

tance matrices. The terminal voltages $(V_j, V_k) \in \mathbb{C}^8$ and the sending-end currents $(I_{jk}, I_{kj}) \in \mathbb{C}^8$ respectively are related according to

$$I_{jk} = \hat{y}^s_{jk}(V_j - V_k) + \hat{y}^m_{jk} V_j, \qquad I_{kj} = \hat{y}^s_{jk}(V_k - V_j) + \hat{y}^m_{kj} V_k \quad (15.3)$$

This model is illustrated in Figure 15.1. It has exactly the same form as (15.1a), except

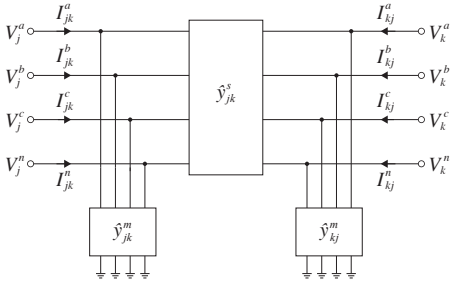


Figure 15.1 A four-wire line characterized by 4×4 series and shunt admittance matrices $\left(\hat{y}^s_{jk}, \hat{y}^m_{jk}, \hat{y}^m_{kj}\right)$.

that the variables and admittances are vectors and matrices respectively. It generalizes (15.1a) from a single-phase model to a three-phase model. The terms $y^m_{jk} V_j$ and $y^m_{kj} V_k$ in (15.3) assume that the shunt admittances connect the buses j and k both to the common reference point for terminal voltages, e.g., the ground.

15.1.3 Three-wire three-phase model

An equivalent three-wire model can be derived from the four-wire models (15.2) and (15.3). To this end denote the phase voltages by $V_j := (V^a_j, V^b_j, V^b_j)$ and $V_k := (V^a_k, V^b_k, V^b_k)$ and phase currents by $I_{jk} := (I^a_{jk}, I^b_{jk}, I^c_{jk})$.

Without shunt admittances.

Ignore first shunt admittances. Decompose the impedance matrix $\hat{z}^s_{jk}$ in (15.2a) into

$$\hat{z}^s_{jk} = \left[\begin{array}{c|c} \hat{z}^{\phi\phi}_{jk} & \hat{z}^{\phi n}_{jk} \\ \hline \hat{z}^{n\phi}_{jk} & \hat{z}^{nn}_{jk} \end{array}\right] := \left[\begin{array}{ccc|c} \hat{z}^{aa}_{jk} & \hat{z}^{ab}_{jk} & \hat{z}^{ac}_{jk} & \hat{z}^{an}_{jk} \\ \hat{z}^{ba}_{jk} & \hat{z}^{bb}_{jk} & \hat{z}^{bc}_{jk} & \hat{z}^{bn}_{jk} \\ \hat{z}^{ca}_{jk} & \hat{z}^{cb}_{jk} & \hat{z}^{cc}_{jk} & \hat{z}^{cn}_{jk} \\ \hline \hat{z}^{na}_{jk} & \hat{z}^{nb}_{jk} & \hat{z}^{nc}_{jk} & \hat{z}^{nn}_{jk} \end{array}\right] \quad (15.4a)$$

where $\hat{z}^{\phi\phi}_{jk} \in \mathbb{C}^{3\times 3}$, $\hat{z}^{nn}_{jk} \in \mathbb{C}$, and $\hat{z}^{\phi n}_{jk}, \hat{z}^{n\phi}_{jk}$ are of matching dimensions. Then (15.2a) can be rewritten as

$$\begin{bmatrix} V_j \\ V^n_j \end{bmatrix} - \begin{bmatrix} V_k \\ V^n_k \end{bmatrix} = \begin{bmatrix} \hat{z}^{\phi\phi}_{jk} & \hat{z}^{\phi n}_{jk} \\ \hat{z}^{n\phi}_{jk} & \hat{z}^{nn}_{jk} \end{bmatrix} \begin{bmatrix} I_{jk} \\ I^n_{jk} \end{bmatrix} \quad (15.4b)$$

The Schur complement of $\hat{z}^{nn}_{jk}$ of $\hat{z}^{s}_{jk}$ is

$$z^{\text{schur}}_{jk} := \hat{z}^{\phi\phi}_{jk} - \frac{1}{\hat{z}^{nn}_{jk}} \hat{z}^{\phi n}_{jk} \hat{z}^{n\phi}_{jk} = \begin{bmatrix} \hat{z}^{aa}_{jk} & \hat{z}^{ab}_{jk} & \hat{z}^{ac}_{jk} \\ \hat{z}^{ba}_{jk} & \hat{z}^{bb}_{jk} & \hat{z}^{bc}_{jk} \\ \hat{z}^{ca}_{jk} & \hat{z}^{cb}_{jk} & \hat{z}^{cc}_{jk} \end{bmatrix} - \frac{1}{\hat{z}^{nn}_{jk}} \begin{bmatrix} \hat{z}^{an}_{jk} \\ \hat{z}^{bn}_{jk} \\ \hat{z}^{cn}_{jk} \end{bmatrix} \begin{bmatrix} \hat{z}^{na}_{jk} & \hat{z}^{nb}_{jk} & \hat{z}^{nc}_{jk} \end{bmatrix} \tag{15.5a}$$

Then we can perform Kron reduction on (15.4) to obtain an equivalent three-wire model that relates $V_j - V_k$ and I^n_{jk} to I_{jk} and $V^n_j - V^n_k$:

$$V_j - V_k = z^{\text{schur}}_{jk} I_{jk} + \frac{\hat{z}^{\phi n}_{jk}}{\hat{z}^{nn}_{jk}} \left(V^n_j - V^n_k \right) \tag{15.5b}$$

$$I^n_{jk} = -\frac{\hat{z}^{n\phi}_{jk}}{\hat{z}^{nn}_{jk}} I_{jk} + \frac{1}{\hat{z}^{nn}_{jk}} \left(V^n_j - V^n_k \right) \tag{15.5c}$$

i.e., a complete three-wire model expresses the phase voltages $V_j - V_k$ and the neutral current I^n_{jk} in terms of the phase currents I_{jk} and neutral voltage difference $V^n_j - V^n_k$. It is equivalent to the four-wire model (15.2) for the case where shunt admittances are assumed zero. Therefore in using three-wire models we generally have to keep track of neutral voltages for Y-configured devices because $V^n_j - V^n_k$ affects the phase voltages and currents $(V_j - V_k, I_{jk})$ through (15.5b).

We refer to the complete model (15.5) as a three-wire model because when the neutral wire is absent or open circuited, e.g., when connecting devices in Δ configuration, or when the neutral is grounded at both the sending and the receiving ends of the line, the phase voltages and currents (V_{jk}, I_{jk}) are related simply by a 3×3 impedance matrix:

1 *Neutral wire absent:* $I^n_{jk} = 0$. Then (15.5) reduces to

$$V_j - V_k = \hat{z}^{\phi\phi}_{jk} I_{jk}, \qquad V^n_j - V^n_k = \hat{z}^{n\phi}_{jk} I_{jk} \tag{15.6a}$$

where $\hat{z}^{\phi\phi}_{jk} \in \mathbb{C}^{3\times3}$ is defined in (15.4a). The neutral voltages V^n_j, V^n_k are generally nonzero since they are not grounded (assuming voltages are defined with respect to the ground) and their difference depends on the phase currents according to (15.6a).

2 *Neutral wire grounded:* $V^n_j = V^n_k$.[2] Then (15.5) reduces to

$$V_j - V_k = z^{\text{schur}}_{jk} I_{jk}, \qquad I^n_{jk} = -\frac{\hat{z}^{n\phi}_{jk}}{\hat{z}^{nn}_{jk}} I_{jk} \tag{15.6b}$$

Even though $V^n_j = V^n_k$ across the neutral wire, the current I^n_{jk} in the neutral wire is generally nonzero and given by (15.6b).

[2] The neutral n' of a Y-configured four-wire device may be through a neutral impedance z^n_j to the external terminal n of the device which is then connected to the neutral of the line. The neutral impedance z^n_j of the device may or may not be zero but $V^n_j = V^n_k$.

Hence when $I^n_{jk} = 0$ or $V^n_j = V^n_k$, we can use a simplified three-wire model and characterize a three-phase line by a 3×3 series impedance matrix z^s_{jk} that relates the phase voltages and currents:

$$V_j \;-\; V_k \;=\; z^s_{jk}\, I_{jk} \tag{15.7}$$

where $z^s_{jk} := \hat{z}^{\phi\phi}_{jk}$ if $I^n_{jk} = 0$ and $z^s_{jk} := z^{\text{schur}}_{jk}$ if $V^n_j = V^n_k$. This is a direct generalization of (15.1d) from a single-phase model to a three-phase model. Even though the three-wire model (15.7) involves no neutral voltage or current, the 3×3 impedance matrix z^s_{jk} includes the effect of neutral lines and earth return (see (15.6)).

Example 15.1. For the case where the neutrals of the sending and receiving ends are grounded through nonzero impedances, derive the three-wire model from the four-wire model (15.2). □

With shunt admittances.

To incorporate the effect of line charging, let the *series admittance matrix* be $y^s_{jk} := \left(z^s_{jk}\right)^{-1}$, assuming z^s_{jk} is invertible. Let $\left(y^m_{jk}, y^m_{kj}\right)$ denote the *shunt admittance matrices*. The terminal voltages $\left(V_j, V_k\right) \in \mathbb{C}^6$ and the sending-end currents $\left(I_{jk}, I_{kj}\right) \in \mathbb{C}^6$ respectively are related according to

$$I_{jk} \;=\; y^s_{jk}\left(V_j - V_k\right) + y^m_{jk} V_j, \qquad I_{kj} \;=\; y^s_{jk}\left(V_k - V_j\right) + y^m_{kj} V_k \tag{15.8a}$$

This model is the three-wire version of (15.3). It is illustrated in Figure 15.2 which is a three-wire version of Figure 15.1. The terms $y^m_{jk} V_j$ and $y^m_{kj} V_k$ in (15.8a) assume that

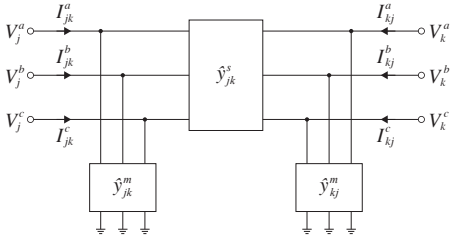


Figure 15.2 A three-wire line characterized by 3×3 series and shunt admittance matrices $\left(y^s_{jk}, y^m_{jk}, y^m_{kj}\right)$.

the shunt admittances connect the buses j and k both to the common reference point for terminal voltages, e.g., the ground.

Example 15.2. Derive the three-wire model (15.8a) directly from the four-wire model (15.3) with nonzero shunt admittances. □

To describe the relation between the sending-end line power and the voltages

(V_j, V_k), define the matrices $S_{jk}, S_{kj} \in \mathbb{C}^{3\times 3}$ by

$$S_{jk} := V_j \left(I_{jk}\right)^{\mathsf{H}} = V_j \left(V_j - V_k\right)^{\mathsf{H}} \left(y^s_{jk}\right)^{\mathsf{H}} + V_j V_j^{\mathsf{H}} \left(y^m_{jk}\right)^{\mathsf{H}} \tag{15.8b}$$

$$S_{kj} := V_k \left(I_{kj}\right)^{\mathsf{H}} = V_k \left(V_k - V_j\right)^{\mathsf{H}} \left(y^s_{jk}\right)^{\mathsf{H}} + V_k V_k^{\mathsf{H}} \left(y^m_{kj}\right)^{\mathsf{H}} \tag{15.8c}$$

The three-phase sending-end line power from terminals j to k along the line is the vector $\text{diag}\left(S_{jk}\right)$ of diagonal entries and that in the opposite direction is the vector $\text{diag}\left(S_{kj}\right)$. The off-diagonal entries of these matrices represent electromagnetic coupling between phases. This generalizes (15.1b)(15.1c) from a single-phase model to a three-phase model.

Example 15.3 (External vs internal variables)**.** Figure 15.3 shows a three-phase voltage source connected to a three-phase impedance load through the line in Figure 15.2. As the figure highlights, the voltages (V_j, V_k) and currents (I_{jk}, I_{kj}) in (15.8a) are

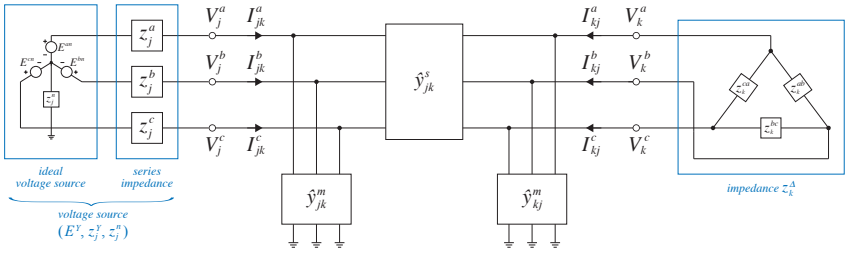


Figure 15.3 A voltage source connected to an impedance load through the line in Figure 15.2.

terminal voltages and currents regardless of whether the three-phase devices connected to terminals j and k are in Y or Δ configuration. The relation between the terminal variables and internal variables are derived in Chapters 14.3.3 and 14.3.4.

The terminal variable (V_j, I_j, s_j) at each bus j satisfies both the external device model and the line model (15.8):

$$0 = f^{\text{ext}}_j\left(V_j, I_j\right), \qquad s_j = \text{diag}\left(V_j I_j^{\mathsf{H}}\right)$$

$$I_j = I_{jk}\left(V_j, V_k\right), \qquad s_j = \text{diag}\left(S_{jk}\left(V_j, V_k\right)\right)$$

In particular the nodal balance equation (15.8) relate (V_j, I_j, s_j) to the terminal voltage V_k at bus k. □

Remark 15.1 (Three-wire model)**.** We will mostly use three-wire line models (15.8) for simplicity, but all analysis extends to four-wire models (including a neutral line) or five-wire models (including a neutral line and the ground return) almost without change with proper definitions that include neutral and ground variables; see Example 16.5 in Chapter 16.2 and Exercise 16.7. □

In most practical situations the series impedance matrix z^s_{jk} is symmetric, i.e.,

$\left(z^s_{jk}\right)^{\phi\phi'} = \left(z^s_{jk}\right)^{\phi'\phi}$, $\phi,\phi' = a,b,c$, meaning that the coupling between phases ϕ and ϕ' does not depend on direction. It is also common in practice that the shunt admittance matrices y^m_{jk} and y^m_{kj} are symmetric. Formally, we assume throughout this chapter:

C15.1: z^s_{jk} is symmetric and invertible. Moreover $z^s_{jk} = z^s_{kj}$.
C15.2: y^m_{jk} and y^m_{kj} are symmetric matrices.

These matrices are generally complex symmetric, but not Hermitian. By Theorem 4.2, z^s_{jk} is invertible and $\mathrm{Re}(y^s_{jk}) \succ 0$ if $\mathrm{Re}(z^s_{jk}) \succ 0$. Assumption C15.1 implies that y^s_{jk} is symmetric and $y^s_{jk} = y^s_{kj}$ (Exercise 15.1).

Symmetric line.

When the line geometry is symmetric (e.g. through transposition) then the series impedance matrix z^s_{jk} has the following important property:

$$z^{aa}_{jk} = z^{bb}_{jk} = z^{cc}_{jk} =: z_{jk} \text{ and } z^{ab}_{jk} = z^{ba}_{jk} = z^{bc}_{jk} = z^{cb}_{jk} = z^{ca}_{jk} = z^{ac}_{jk} =: \epsilon_{jk}$$

so that

$$z^s_{jk} = \begin{bmatrix} z_{jk} & \epsilon_{jk} & \epsilon_{jk} \\ \epsilon_{jk} & z_{jk} & \epsilon_{jk} \\ \epsilon_{jk} & \epsilon_{jk} & z_{jk} \end{bmatrix} = (z_{jk} - \epsilon_{jk})\mathbb{I} + \epsilon_{jk}\mathbf{1}\mathbf{1}^\mathsf{T} \tag{15.9a}$$

Typically $|z_{jk}| > |\epsilon_{jk}|$. Then the line admittance $y^s_{jk} := \left(z^s_{jk}\right)^{-1}$ has the same structure

$$y^s_{jk} = \begin{bmatrix} y^1_{jk} & y^2_{jk} & y^2_{jk} \\ y^2_{jk} & y^1_{jk} & y^2_{jk} \\ y^2_{jk} & y^2_{jk} & y^1_{jk} \end{bmatrix} = (y_{jk} - \delta_{jk})\mathbb{I} + \delta_{jk}\mathbf{1}\mathbf{1}^\mathsf{T} \tag{15.9b}$$

where

$$y_{jk} := \frac{z_{jk} + \epsilon_{jk}}{(z_{jk} - \epsilon_{jk})(z_{jk} + 2\epsilon_{jk})}, \quad \delta_{jk} := -\frac{\epsilon_{jk}}{(z_{jk} - \epsilon_{jk})(z_{jk} + 2\epsilon_{jk})} \tag{15.9c}$$

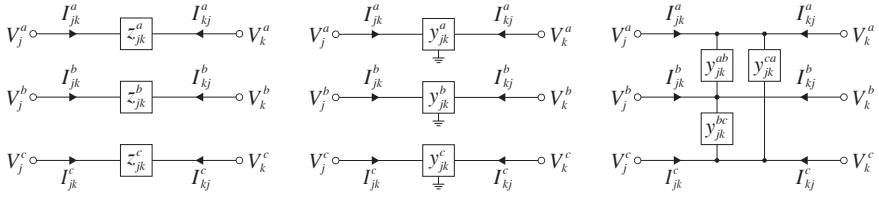
and (15.9c) follows from:

$$\begin{aligned} \mathbb{I} = y^s_{jk} z^s_{jk} &= \left((y_{jk} - \delta_{jk})\mathbb{I} + \delta\mathbf{1}\mathbf{1}^\mathsf{T}\right)\left((z_{jk} - \epsilon_{jk})\mathbb{I} + \epsilon_{jk}\mathbf{1}\mathbf{1}^\mathsf{T}\right) \\ &= (y_{jk} - \delta_{jk})(z_{jk} - \epsilon_{jk})\mathbb{I} + (\epsilon_{jk} y_{jk} + z_{jk}\delta_{jk} + \epsilon_{jk}\delta_{jk})\mathbf{1}\mathbf{1}^\mathsf{T} \end{aligned}$$

Typically $|y_{jk}| > |\delta_{jk}|$. If the sources and loads are balanced so that currents sum to zero $i^a(t) + i^b(t) + i^c(t) = 0$ and charges sum to zero $q^a(t) + q^b(t) + q^c(t) = 0$ across phases then $\epsilon_{jk} = 0$ (see Chapter 2.1.4), i.e., z^s_{jk} is diagonal and the voltages and currents of different phases are decoupled. Otherwise z^s_{jk} is not diagonal and therefore the voltages and currents of different phases are coupled even if the line is symmetric, i.e., even if the series impedance z^s_{jk} satisfies (15.9). As we will see in Chapter 16.4.4, in this case, when shunt admittances are assumed zero, a similarity transformation using the

unitary matrix F yields a diagonal impedance matrix $\tilde{z}^s_{jk}$ in the sequence coordinate. This leads to decoupled relation between the sequence voltages and currents across the three-phase line that can be interpreted as defining separate sequence networks.

Example 15.4 (Special lines)**.** The line in (15.8a) is an abstraction that can model a transmission or distribution line, a transformer, or parts of series impedances or shunt admittances of generators or loads. We discuss some degenerate forms of (15.8a) that will be used for this purpose, e.g., for modeling non-ideal voltage and current sources in Chapter 15.1.4. The series impedance z^Y_{jk} in Figure 15.4(a) can be treated as a line



(a) Series impedance (b) Shunt admittance in Y (c) Shunt admittance in Δ

Figure 15.4 Special three-wire lines characterized by (15.10).

$\left(y^s_{jk}, y^m_{jk}, y^m_{kj}\right)$ with a diagonal series impedance, i.e., $y^m_{jk} = y^m_{kj} = 0$, and

$$y^s_{jk} := \operatorname{diag}^{-1}\left(z^a_{jk}, z^b_{jk}, z^c_{jk}\right), \quad I_{jk} := y^s_{jk}\left(V_j - V_k\right), \quad I_{kj} := -I_{jk} \tag{15.10a}$$

The Y-configured shunt admittance y^Y_{jk} in Figure 15.4(b) can be treated as a line $\left(y^s_{jk}, y^m_{jk}, y^m_{kj}\right)$ with a shunt admittance in Y configuration, i.e., $z^s_{jk} = 0$, $y^m_{kj} = 0$, and

$$y^m_{jk} := \operatorname{diag}\left(y^a_{jk}, y^b_{jk}, y^c_{jk}\right), \quad V_j = V_k, \quad I_{jk} + I_{kj} = y^m_{jk} V_j \tag{15.10b}$$

The Δ-configured shunt admittance y^Δ_{jk} in Figure 15.4(c) can be treated as a line $\left(y^s_{jk}, y^m_{jk}, y^m_{kj}\right)$ with a shunt admittance in Δ configuration, i.e., $z^s_{jk} = 0$, $y^m_{kj} = 0$, and

$$y^m_{jk} := \operatorname{diag}\left(y^{ab}_{jk}, y^{bc}_{jk}, y^{ca}_{jk}\right), \qquad V_j = V_k, \qquad I^\Delta = y^m_{jk} \Gamma V_j$$

where $I^\Delta := \left(I^{ab}, I^{bc}, I^{ca}\right)$ are the line-to-line current internal to the Δ configuration. Therefore for any currents I_{jk} and I_{kj} with $\mathbf{1}^\mathsf{T} I_{jk} = \mathbf{1}^\mathsf{T} I_{kj} = 0$, the degenerate line in Figure 15.4(c) is characterized by

$$y^m_{jk} := \operatorname{diag}\left(y^{ab}_{jk}, y^{bc}_{jk}, y^{ca}_{jk}\right), \quad V_j = V_k, \quad \Gamma^{\mathsf{T}\dagger}\left(I_{jk} + I_{kj}\right) + \beta \mathbf{1} = y^m_{jk} \Gamma V_j \tag{15.10c}$$

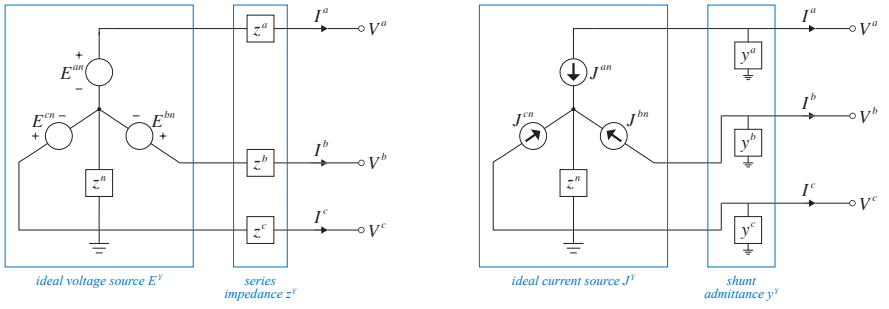
where $\beta \in \mathbb{C}$ depends on the amount of loop flow in the internal current I^Δ.

□

We next use these special lines to simplify models for non-ideal voltage and current sources in Y and Δ configurations.

15.1.4 Ideal voltage and current sources

A voltage or current source in Y configuration may or may not have a neutral line which may or may not be grounded. Figure 15.5 shows the case where the neutral is grounded through an impedance z^n. In this case the voltage source $\left(E^Y, z^Y, z^n\right)$ can be treated as an ideal voltage source $\left(E^Y, z^n\right)$ connected to a (degenerate) three-phase line with a series impedance z^Y characterized by (15.10a). Similarly a grounded current



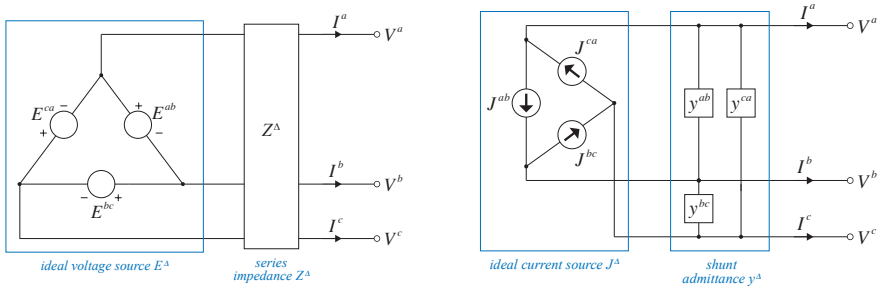
(a) Voltage source (b) Current source

Figure 15.5 Three-wire sources in Y configuration. (a) A voltage source. (b) A current source.

source $\left(J^Y, y^y, z^n\right)$ in Y configuration, as shown in Figure 15.5(b), can be treated as an ideal current source $\left(J^Y, z^n\right)$ connected to a three-phase line with a shunt admittance y^Y characterized by (15.10b). In both cases the ideal source has no series impedance or shunt admittance. In general the neutral voltage V^n is nonzero whether or not there is a neutral line and whether or not the neutral is grounded.

A voltage source $\left(E^\Delta, z^\Delta\right)$ in Δ configuration, as shown in Figure 15.6(a), can be treated as an ideal voltage source E^Δ in Δ configuration connected to a three-phase line with a series impedance $Z^\Delta := \frac{1}{9}\Gamma^{\mathsf{T}} z^\Delta \left(\mathbb{I} - \frac{1}{\zeta}\mathbf{1}\tilde{z}^{\Delta\mathsf{T}}\right)\Gamma$ in (14.21b). A current source $\left(J^\Delta, y^\Delta\right)$ in Δ configuration, as shown in Figure 15.6(b), can be treated as an ideal current source J^Δ in Δ configuration connected to a three-phase line with a shunt admittance y^Δ in Δ configuration characterized by (15.10c).

Example 15.5 (Ideal sources)**.** Figure 15.7 shows a three-phase voltage source in Y configuration connected to a three-phase current source in Δ configuration through the line in Figure 15.2. The shunt admittance $y_k^\Delta := \operatorname{diag}\left(y_k^{ab}, y_k^{bc}, y_k^{ca}\right)$ of the current source can be absorbed into the shunt admittance matrix y_{kj}^m of the line so that the system is equivalent to an ideal current source J_k^Δ connected to terminal k of a line



(a) Voltage source (b) Current source

Figure 15.6 Three-wire sources in Δ configuration. (a) A voltage source. (b) A current source.

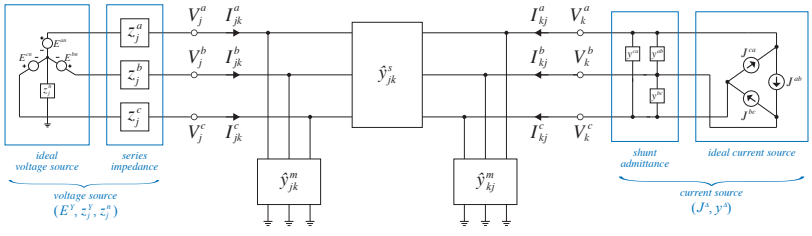


Figure 15.7 A voltage source connected to a current source through the line in Figure 15.2.

with an equivalent shunt admittance matrix $\tilde{y}^m_{kj}$ given by:

$$\tilde{y}^m_{kj} := \underbrace{\begin{bmatrix} y^{aa}_{kj} & y^{ab}_{kj} & y^{ac}_{kj} \\ y^{ba}_{kj} & y^{bb}_{kj} & y^{bc}_{kj} \\ y^{ca}_{kj} & y^{cb}_{kj} & y^{cc}_{kj} \end{bmatrix}}_{y_{kj}} + \underbrace{\begin{bmatrix} 0 & y^{ab}_{k} & y^{ca}_{k} \\ y^{ab}_{k} & 0 & y^{bc}_{k} \\ y^{ca}_{k} & y^{bc}_{k} & 0 \end{bmatrix}}_{\text{from } y^{\Delta}_{k}}$$

Note that in this equivalent model the two shunt admittance matrices y^m_{jk} and $\tilde{y}^m_{kj}$ are generally unequal even if $y^m_{jk} = y^m_{kj}$ originally. Note also that the series impedance matrix z^Y_j of the voltage source cannot be directly absorbed into the line parameters. □

15.2 Three-phase transformer models: simplified circuit

In this section we show that, as for a three-phase line, the external model of a three-phase transformer takes the form of an admittance matrix Y. The general method is similar to that for other three-phase devices: (i) define internal and terminal variables; (ii) derive conversion rules that relate internal and terminal variables; (ii) define internal models that relate these internal variables; and finally (iv) eliminate the internal variables to arrive at the external model. We start by reviewing the single-phase transformer. The

notation and the derivation generalize naturally when these transformers are configured into a three-phase transformer.

15.2.1 Review: single-phase transformer

Consider the simplified mode of a single-phase transformer in Figure 3.5 of Chapter 3.1.4, reproduced in Figure 15.8, consisting of an ideal transformer with a voltage gain n, a leakage admittance y^s and a shunt admittance y^m on the primary side. Let the turns ratio be $a := n^{-1}$ (even though a is used to denote both a phase and a turns ratio its meaning should be clear from the context). The currents entering/leaving and

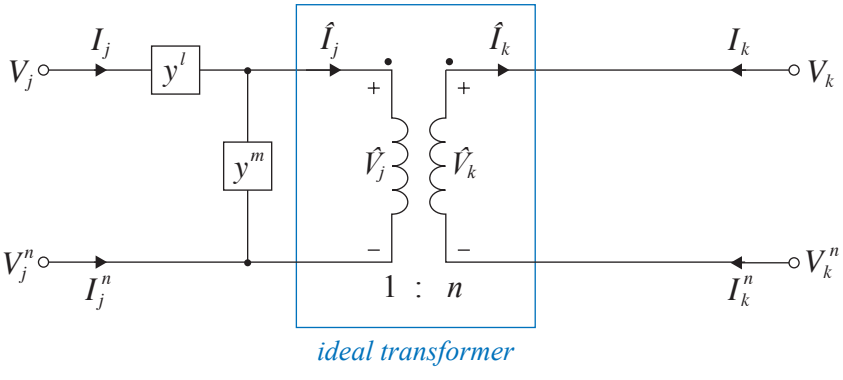


Figure 15.8 Single-phase transformer: simplified model. The internal variables $(\hat{V}_j, \hat{I}_j)$, $(\hat{V}_k, \hat{I}_k)$ and terminal variables (V_j, V_j^n, I_j), (V_k, V_k^n, I_k).

the voltages across the ideal transformer are denoted by variables with a hat: $(\hat{V}_j, \hat{I}_j)$, $(\hat{V}_k, \hat{I}_k)$. They are called *internal* variables. The dot notation on the ideal transformer indicates that the internal currents are defined to be positive when $\hat{I}_j$ flows into and $\hat{I}_k$ flows out of the dotted terminals, as indicated in Figure 15.8.

The *terminal voltages* $\left(V_j, V_j^n, V_k, V_k^n\right)$ are defined with respect to an arbitrary but common reference point, e.g., the ground. We emphasize that, while the internal voltages $(\hat{V}_j, \hat{V}_k)$ are defined to be the voltage drops across the ideal transformer windings, the terminal voltages $\left(V_j, V_j^n, V_k, V_k^n\right)$ are defined with respect to a common reference point; in particular the primary and secondary windings are not assumed to be grounded. The *terminal currents* (I_j, I_k) are defined to be the sending-end currents from buses j and k respectively to the other side, as shown in Figure 15.8. The terminal and internal variables are related by the *conversion rule*:

$$I_j = y^l\left(V_j - V_j^n - \hat{V}_j\right), \qquad I_j = y^m \hat{V}_j + \hat{I}_j, \qquad I_j^n = -I_j \tag{15.11a}$$

$$\hat{V}_k = V_k - V_k^n, \qquad \hat{I}_k = -I_k, \qquad I_k^n = -I_k \tag{15.11b}$$

where the neutral currents (I_j^n, I_k^n) are injections from the neutral terminals into the ideal transformer and follow from $I_j^n = -(y^m \hat{V}_j + \hat{I}_j) = -I_j$ and $I_k^n = \hat{I}_k = -I_k$ respectively. The *internal model* of the single-phase (ideal) transformer is defined by its transformer gains (n, a):

$$\hat{V}_k \;=\; n\hat{V}_j, \qquad\qquad \hat{I}_k \;=\; \frac{1}{n}\hat{I}_j \;=:\; a\hat{I}_j \tag{15.11c}$$

Eliminating the internal variables from (15.11) yields an *external model* that relates the terminal variables:

$$I_j \;=\; y^l\left(\left(V_j - V_j^n\right) - a\left(V_k - V_k^n\right)\right), \quad I_k \;=\; -a\hat{I}_j \;=\; ay^m\left(V_j - V_j^n\right) - a\left(1 + \frac{y^m}{y^l}\right)I_j$$

or in terms of an admittance matrix Y:

$$\begin{bmatrix} I_j \\ I_k \end{bmatrix} \;=\; \underbrace{\begin{bmatrix} y^l & -ay^l \\ -ay^l & a^2(y^l + y^m) \end{bmatrix}}_{Y} \left(\begin{bmatrix} V_j \\ V_k \end{bmatrix} - \begin{bmatrix} V_j^n \\ V_k^n \end{bmatrix}\right) \tag{15.12a}$$

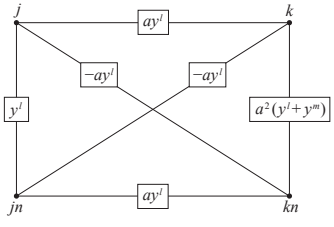
We can add neutral currents from (15.11) to (15.12a):

$$\begin{bmatrix} I_j^n \\ I_k^n \end{bmatrix} \;=\; -\begin{bmatrix} I_j \\ I_k \end{bmatrix} \;=\; -Y\left(\begin{bmatrix} V_j \\ V_k \end{bmatrix} - \begin{bmatrix} V_j^n \\ V_k^n \end{bmatrix}\right)$$

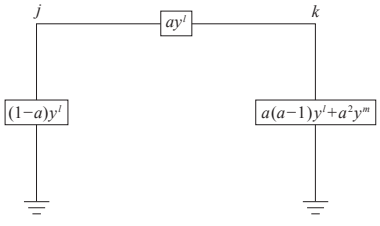
to obtain a two-wire model of a single-phase transformer:

$$\begin{bmatrix} I_j \\ I_k \\ I_j^n \\ I_k^n \end{bmatrix} \;=\; \underbrace{\begin{bmatrix} Y & -Y \\ -Y & Y \end{bmatrix}}_{Y^{2\text{wire}}} \begin{bmatrix} V_j \\ V_k \\ V_j^n \\ V_k^n \end{bmatrix} \tag{15.12b}$$

Both Y and the 4×4 admittance matrix $Y^{2\text{wire}}$ are complex symmetric. While Y generally has nonzero row and column sums, $Y^{2\text{wire}}$ has zero row and column sums. The admittance matrix $Y^{2\text{wire}}$ is represented by a four-node network in Figure 15.9(a). Since $Y^{2\text{wire}}$ has zero row and column sums, there are no shunt admittances in the



(a) General circuit model.



(b) Π circuit model.

Figure 15.9 (a) Circuit model of admittance matrix $Y^{2\text{wire}}$ and (b) when neutrals are grounded with zero grounding impedances, $V_j^n = V_k^n = 0$.

four-node network in Figure 15.9(a).

It is often assume implicitly (e.g., in Chapter 3 and Chapter 4.1.3) that neutrals are grounded with zero grounding impedance and voltages are defined with respect to the ground (assumption C14.1). In this case, $V_j^n = V_k^n = 0$ and the model (15.12a) reduces to a Π circuit model:

$$\begin{bmatrix} I_j \\ I_k \end{bmatrix} = Y \begin{bmatrix} V_j \\ V_k \end{bmatrix}$$

The four-node network in Figure 15.9(b) then reduces to a Π circuit in which parallel branches to the ground are combined into shunt admittances, i.e., it can be characterized by series and shunt admittances given by

$$\tilde{y}^s_{jk} := ay^l, \qquad \tilde{y}^m_{jk} := (1-a)y^l, \qquad \tilde{y}^m_{kj} := a(a-1)y^l + a^2 y^m \tag{15.12c}$$

like a transmission or distribution line.

We now explain how these relations (15.11)(15.12) extend naturally to three-phase transformers in an unbalanced setting.

15.2.2 General derivation method

The external model of a three-phase transformer depends on the models of its constituent single-phase transformers and their configuration on each side of the three-phase transformer. In particular each of the primary and secondary sides can be in Y or Δ configuration, giving four configurations for a standard three-phase transformer. The external model can be derived in four simple steps, similar to the derivation for a single-phase transformer or other three-phase devices:

1. *Conversion rule*: For the primary side, define the internal variables $(\hat{V}_j, \hat{I}_j)$ and external variables (V_j, V_j^n, I_j) (defined precisely below) and relate them.
2. *Conversion rule*: For the secondary side, define the internal variables $(\hat{V}_k, \hat{I}_k)$ and external variables (V_k, V_k^n, I_k) and relate them.
3. *Internal model*: Couple these relations through the transformer gains (15.11c) on $(\hat{V}_j, \hat{I}_j)$, $(\hat{V}_k, \hat{I}_k)$ for each of the single-phase transformers.
4. *External model*: Derive the external model, a relation between external variables (V_j, I_j) and (V_k, I_k), by eliminating the internal variables.

This method is modular and applicable in a general setting where the single-phase transformers may have different admittances or turns ratios, the neutrals of Y configurations may or may not be connected to the other side, may or may not be grounded, with zero or nonzero grounding impedances. The method can also be generalized to non-standard transformers such as open transformers.

We now describe these steps in more detail.

1. Primary side.

Consider the primary circuit of a three-phase transformer in Y or Δ configuration in Figure 15.10. The internal voltages and currents associated with the ideal transformer

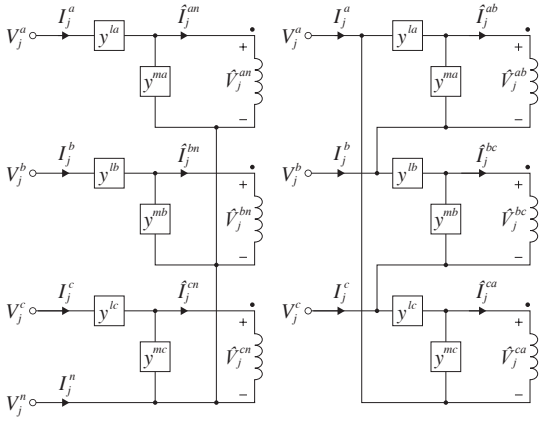


Figure 15.10 Primary side of a three-phase transformer in Y (left) or Δ (right) configuration.

are denoted by

$$\hat{V}_j^Y := \begin{bmatrix} \hat{V}_j^{an} \\ \hat{V}_j^{bn} \\ \hat{V}_j^{cn} \end{bmatrix}, \quad \hat{I}_j^Y := \begin{bmatrix} \hat{I}_j^{an} \\ \hat{I}_j^{bn} \\ \hat{I}_j^{cn} \end{bmatrix}, \quad \hat{V}_j^\Delta := \begin{bmatrix} \hat{V}_j^{ab} \\ \hat{V}_j^{bc} \\ \hat{V}_j^{ca} \end{bmatrix}, \quad \hat{I}_j^\Delta := \begin{bmatrix} \hat{I}_j^{ab} \\ \hat{I}_j^{bc} \\ \hat{I}_j^{ca} \end{bmatrix}$$

The terminal voltages and currents are denoted by

$$V_j := \begin{bmatrix} V_j^a \\ V_j^b \\ V_j^c \end{bmatrix}, \qquad I_j := \begin{bmatrix} I_j^a \\ I_j^b \\ I_j^c \end{bmatrix}$$

regardless of the configuration. For Y configuration the (terminal) neutral voltage and current are denoted by $\left(V_j^n, I_j^n\right)$ in the direction shown in Figure 15.10. As for the single-phase model, these voltages are defined with respect to a common reference point (e.g., the ground); in particular the neutrals are not assumed to be grounded. Note that the internal voltages and currents $\left(\hat{V}_j^{Y/\Delta}, \hat{I}_j^{Y/\Delta}\right)$ are defined across the ideal transformers. In general, $V_j \neq \hat{V}_j^Y + V_j^n \mathbf{1}$ and $\hat{V}_j^\Delta \neq \Gamma V_j$. Moreover, $I_j \neq \hat{I}_j^Y$ and $I_j \neq \Gamma^\top \hat{I}_j^\Delta$, unless $y^m = 0$.

The leakage admittances of the transformer are denoted by the diagonal matrix $y^l := \text{diag}\left(y^{la}, y^{lb}, y^{lc}\right)$ and the shunt admittances are denoted by $y^m := \text{diag}\left(y^{ma}, y^{mb}, y^{mc}\right)$. From (15.11a) for each single-phase transformer the terminal

variables are related to the internal variables according to the *conversion rule*:

$$Y \text{ configuration:} \quad I_j = y^l \left(V_j - V_j^n \mathbf{1} - \hat{V}_j^Y \right), \quad I_j = y^m \hat{V}_j^Y + \hat{I}_j^Y, \qquad I_j^n = -\mathbf{1}^\mathsf{T} I_j \tag{15.13a}$$

$$\Delta \text{ configuration:} \quad \hat{I}_j^\Delta = y^l \Gamma V_j - (y^l + y^m) \hat{V}_j^\Delta, \quad I_j = \Gamma^\mathsf{T} \left(\hat{I}_j^\Delta + y^m \hat{V}_j^\Delta \right) \tag{15.13b}$$

For Y configuration the neutral current I_j^n in (15.13a) follows from $I_j^n = -\mathbf{1}^\mathsf{T} \left(y^m \hat{V}_j^Y + \hat{I}_j^Y \right) = -\mathbf{1}^\mathsf{T} I_j$. For Δ configuration $\hat{I}_j^\Delta$ in (15.13b) follows from $\hat{I}_j^{ab} + y^{ma} \hat{V}_j^{ab} = y^{la} \left(V_j^a - V_j^b - \hat{V}_j^{ab} \right)$. Clearly $\mathbf{1}^\mathsf{T} I_j = 0$ for Δ configuration. Moreover (15.13) implies that the internal and terminal voltages are related according to

$$Y \text{ configuration:} \quad V_j = \hat{V}_j^Y + V_j^n \mathbf{1} + z^l I_j \tag{15.13c}$$

$$\Delta \text{ configuration:} \quad \hat{V}_j^\Delta = \Gamma V_j + y^l z^m \Gamma V_j - (z^l + z^m) \hat{I}_j^\Delta \tag{15.13d}$$

where $z^l := (y^l)^{-1}$ and $z^m := (y^m)^{-1}$.

2. Secondary side.

Consider the secondary side of a three-phase transformer in Y or Δ configuration in Figure 15.11. The internal voltages and currents associated with the transformer are

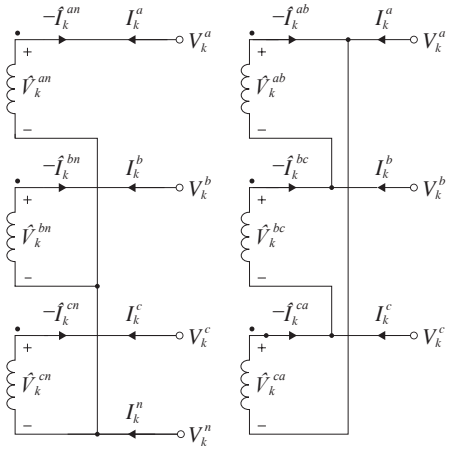


Figure 15.11 Secondary side of a three-phase transformer in Y (left) or Δ (right) configuration.

denoted by

$$\hat{V}_k^Y := \begin{bmatrix} \hat{V}_k^{an} \\ \hat{V}_k^{bn} \\ \hat{V}_k^{cn} \end{bmatrix}, \quad \hat{I}_k^Y := \begin{bmatrix} \hat{I}_k^{an} \\ \hat{I}_k^{bn} \\ \hat{I}_k^{cn} \end{bmatrix}, \quad \hat{V}_k^\Delta := \begin{bmatrix} \hat{V}_k^{ab} \\ \hat{V}_k^{bc} \\ \hat{V}_k^{ca} \end{bmatrix}, \quad \hat{I}_k^\Delta := \begin{bmatrix} \hat{I}_k^{ab} \\ \hat{I}_k^{bc} \\ \hat{I}_k^{ca} \end{bmatrix}$$

The terminal voltages and currents are denoted by

$$V_k := \begin{bmatrix} V_k^a \\ V_k^b \\ V_k^c \end{bmatrix}, \qquad I_k := \begin{bmatrix} I_k^a \\ I_k^b \\ I_k^c \end{bmatrix}$$

regardless of the configuration. For Y configuration the neutral voltage and current are denoted by (V_k^n, I_k^n) in the direction shown in Figure 15.11.

From (15.11b) for each single-phase transformer the terminal variables are related to the internal variables according to the *conversion rule*:

$$Y \text{ configuration:} \quad V_k = \hat{V}_k^Y + V_k^n \mathbf{1}, \quad I_k = \hat{I}_k^Y, \quad I_k^n = -\mathbf{1}^\mathsf{T} \hat{I}_k^Y = -\mathbf{1}^\mathsf{T} I_k \tag{15.14a}$$

$$\Delta \text{ configuration:} \quad \hat{V}_k^\Delta = \Gamma V_k, \quad I_k = \Gamma^\mathsf{T} \hat{I}_k^\Delta \tag{15.14b}$$

For Δ configuration, $\mathbf{1}^\mathsf{T} I_k = 0$.

3. Internal model.

The voltage and current gains across the ideal transformer define an *internal model* which couples the internal variables in the primary and secondary circuits and connects the relations (15.13) and (15.14). These gains are determined by the turns ratios of the constituent single-phase ideal transformers according to (15.11c), but tailored for different configurations. Denote the voltage gain of the ideal three-phase transformer by a real diagonal matrix $n := \text{diag}\,(n^a, n^b, n^c) \in \mathbb{R}^{3\times 3}$ and its turns ratio by $a := n^{-1} \in \mathbb{R}^{3\times 3}$. Then

$$YY \text{ configuration:} \quad \hat{V}_k^Y = n \hat{V}_j^Y, \quad -\hat{I}_k^Y = a \hat{I}_j^Y \tag{15.15a}$$

$$\Delta\Delta \text{ configuration:} \quad \hat{V}_k^\Delta = n \hat{V}_j^\Delta, \quad -\hat{I}_k^\Delta = a \hat{I}_j^\Delta \tag{15.15b}$$

$$\Delta Y \text{ configuration:} \quad \hat{V}_k^Y = n \hat{V}_j^\Delta, \quad -\hat{I}_k^Y = a \hat{I}_j^\Delta \tag{15.15c}$$

$$Y\Delta \text{ configuration:} \quad \hat{V}_k^\Delta = n \hat{V}_j^Y, \quad -\hat{I}_k^\Delta = a \hat{I}_j^Y \tag{15.15d}$$

These are internal models of a three-phase (ideal) transformer. The negative signs on $\hat{I}_k^Y$ and $\hat{I}_k^\Delta$ are due to the convention that the transformer current gain is defined for secondary current leaving the dotted terminal of the secondary winding (see Figure 15.11).

4. External model.

The *external model* of a three-phase transformer relates the terminal variables (V_j, V_j^n, I_j) and (V_k, V_k^n, I_k) on both sides of the transformer in terms of the leakage admittance y^s, the shunt admittance y^m, and the turns ratio a. It can be derived by eliminating the internal variables $\left(\hat{V}_j^{Y/\Delta}, \hat{I}_j^{Y/\Delta}\right)$ and $\left(\hat{V}_k^{Y/\Delta}, \hat{I}_k^{Y/\Delta}\right)$ from the conversion rules (15.13) (15.14) and the internal model (15.15).

The external models, derived in detail below, turn out to have a striking modular structure. To describe the general form let $V := (V_j, V_k) \in \mathbb{C}^6$ and $I := (I_j, I_k) \in \mathbb{C}^6$. Define a 6×6 admittance matrix Y_{YY} and a column vector $\gamma \in \mathbb{C}^6$:

$$Y_{YY} := \begin{bmatrix} y^l & -ay^l \\ -ay^l & a^2(y^l + y^m) \end{bmatrix}, \qquad \gamma := \begin{bmatrix} V_j^n \mathbf{1} \\ V_k^n \mathbf{1} \end{bmatrix} \tag{15.16a}$$

where $\mathbf{1} := (1,1,1)$. Let D denote a 6×6 block diagonal matrix whose value depends on configuration. As we will explain below Y_{YY} is the admittance matrix of a transformer in YY configuration. It is the same as that in (15.12a) for a single-phase transformer, except that a, y are now 3×3 diagonal matrices rather than scalars. The vector γ is the neutral voltages of a transformer in YY configuration. For $\Delta\Delta$ configuration, $D\gamma = 0 \in \mathbb{C}^6$ in (15.16b), reflecting that a Δ configuration contains no neutral voltage; similarly for ΔY and $Y\Delta$ configurations. The external models of three-phase transformers in YY, $\Delta\Delta$, ΔY and $Y\Delta$ configurations take the form

$$I \;=\; D^{\mathsf{T}} Y_{YY} D\,(V - \gamma) \tag{15.16b}$$

where D is a 6×6 block diagonal matrix that depends on configuration:

$$YY \text{ configuration:} \qquad D := \begin{bmatrix} \mathbb{I} & 0 \\ 0 & \mathbb{I} \end{bmatrix} \tag{15.16c}$$

$$\Delta\Delta \text{ configuration:} \qquad D := \begin{bmatrix} \Gamma & 0 \\ 0 & \Gamma \end{bmatrix} \tag{15.16d}$$

$$\Delta Y \text{ configuration:} \qquad D := \begin{bmatrix} \Gamma & 0 \\ 0 & \mathbb{I} \end{bmatrix} \tag{15.16e}$$

$$Y\Delta \text{ configuration:} \qquad D := \begin{bmatrix} \mathbb{I} & 0 \\ 0 & \Gamma \end{bmatrix} \tag{15.16f}$$

Hence the external models of $\Delta\Delta$, ΔY, $Y\Delta$ configurations can be obtained by pre-multiplying the admittance matrix Y_{YY} of the YY configuration by Γ^{T} and post-multiplying it by Γ for a (primary or secondary) circuit that is in Δ configuration and setting its neutral voltage to zero. This has a simple interpretation. Take $\Delta\Delta$ configuration as an example: $D(V-\gamma) = DV = (\Gamma V_j, \Gamma V_k)$ can be interpreted as the internal line-to-line voltages of a certain three-phase device in Δ configuration, $Y_{YY}DV$ can be interpreted as the corresponding internal currents, and hence $D^{\mathsf{T}}(Y_{YY}DV)$ converts this internal current to terminal currents that are externally observable.

Remark 15.2. 1 Neither the voltage gains $n := (n^a, n^b, n^c)$ nor the admittances $y^l := (y^{la}, y^{lb}, y^{lc})$, $y^m := (y^{ma}, y^{mb}, y^{mc})$ may be equal across phases a, b, c. Unless otherwise specified we assume n and a are real matrices. This is the case if they represent voltage gains and turns ratios of constituent single-phase transformers (they can be complex if phase-shifting transformers are involved or if the three-phase transformer is the YY equivalent model of a ΔY-configured transformer in a balanced setting; see Example 15.7).

2 The derivation method is modular. If a different single-phase transformer model

is used, e.g., with complex transformer gains, then the relations (15.13) or (15.14) need to be modified but the structure of the derivation remains unchanged.

3 The model (15.16) is a three-wire model that does not include neutral currents. See (15.19c) for a four-wire model that does.

4 The method is also applicable to non-standard transformers such as open transformers. Indeed the external model of an open $\Delta\Delta$ transformer is also given by (15.16b) (15.16d) but with the diagonal matrices y^l, y^m in Y_{YY} in (15.16a) replaced by $\text{diag}\left(y^{la}, y^{lb}, 0\right)$ and $\text{diag}\left(y^{ma}, y^{mb}, 0\right)$ with $y^{lc} = y^{mc} = 0$ on the third leg that has no transformer.

□

We will illustrate this general method by deriving the external models (15.16) of three-phase transformers in YY, $\Delta\Delta$, ΔY and $Y\Delta$ configurations and then show how to adapt the method to non-standard transformers such as open transformers. We start by explaining when a three-phase transformer can be represented by a three-phase Π circuit.

15.2.3 Three-phase Π circuit, block symmetry, symmetry

Refer to the Π circuit model in Figure 15.9(b) for a single-phase transformer where the neutral voltages $V_j^n = V_k^n = 0$. The series and shunt admittances $\left(\tilde{y}_{jk}^s, \tilde{y}_{jk}^m, \tilde{y}_{kj}^m\right)$ of the Π circuit are given by (15.12c). They define a 2×2 admittance matrix Y_{jk} that relates $\left(V_j, V_k\right)$ to $\left(I_{jk}, I_{kj}\right)$ that is complex symmetric. This is because the application of Kirchhoff's laws to this circuit yields

$$I_{jk} \;=\; \tilde{y}_{jk}^s\left(V_j - V_k\right) + \tilde{y}_{jk}^m V_j, \qquad I_{kj} \;=\; \tilde{y}_{jk}^s\left(V_k - V_j\right) + \tilde{y}_{jk}^m V_k \tag{15.17}$$

Therefore a single-phase transformer always has a Π circuit representation and, in this sense, behaves like a single-phase transmission line.

This is not the case for three-phase transformers. Consider a three-phase transformer and denote by Y_{jk} the 6×6 admittance matrix that maps its voltage vectors $\left(V_j, V_k\right) \in \mathbb{C}^6$ to its current vectors $\left(I_{jk}, I_{kj}\right) \in \mathbb{C}^6$, i.e.,

$$\begin{bmatrix} I_{jk} \\ I_{kj} \end{bmatrix} = \underbrace{\begin{bmatrix} Y_{jk,11} & Y_{jk,12} \\ Y_{jk,21} & Y_{jk,22} \end{bmatrix}}_{Y_{jk}} \begin{bmatrix} V_j \\ V_k \end{bmatrix}$$

If Y_{jk} can be represented by a *three-phase* Π circuit model, i.e., if it behaves like a three-phase transmission line as shown in Figure 15.2, then (15.17) must also hold but $\left(\tilde{y}_{jk}^s, \tilde{y}_{jk}^m, \tilde{y}_{kj}^m\right)$ are now 3×3 matrices, not scalars. This means that the two off-diagonal

submtrices of $Y_{jk} \in \mathbb{C}^6$ must be *equal* $Y_{jk,12} = Y_{jk,21}$ and Y_{jk} must be of the form

$$Y_{jk} = \begin{bmatrix} \tilde{y}^s_{jk} + \tilde{y}^m_{jk} & -\tilde{y}^s_{jk} \\ -\tilde{y}^s_{jk} & \tilde{y}^s_{jk} + \tilde{y}^m_{kj} \end{bmatrix}$$

We call such a matrix *block symmetric* (see Definition 16.1). In contrast, if Y_{jk} is symmetric then $Y^{\mathsf{T}}_{jk,12} = Y_{jk,21}$. As we will see a three-phase transformer may not be block symmetric and hence may not have a three-phase Π circuit representation. For balanced systems, this manifests itself as the per-phase model of a ΔY or $Y\Delta$-configured transformer having no single-phase Π circuit representation because of the its complex voltage gain $K(n)$, as discussed in Chapter 4.1.3. This phenomenon is generalized in the rest of this section for unbalanced systems.

Whether or not Y_{jk} is block symmetric we can always interpret Y_{jk} as the 6×6 admittance matrix of a single-phase network consisting of 6 buses, indexed by $i\phi$, $i = j, k$ and $\phi \in \{a, b, c\}$, as studied in Chapter 4.2. This is referred to as its single-phase equivalent circuit and studied in Chapter 16.1.2.

A matrix can be symmetric but not block symmetric, and vice versa. Symmetry of a matrix is determined only by its off-diagonal entries but its diagonal entries can be arbitrary. Block symmetry is determined only by its off-diagonal blocks but its diagonal blocks can be arbitrary. A symmetric Y_{jk} is block symmetric if $Y^{\mathsf{T}}_{jk,12} = Y_{jk,12}$. A block symmetric Y_{jk} is symmetric if all submatrices $Y_{jk,12}, Y_{jk,11}, Y_{jk,22}$ are symmetric. These are reasonable assumptions for modeling a three-phase transmission or distribution line, i.e., Y_{jk} for a transmission or distribution line can be assumed to be both block symmetric and symmetric and therefore has both a three-phase Π circuit representation and a single-phase equivalent circuit. This is not necessarily the case for three-phase transformers.

We will generalize the concepts of block symmetry and single-phase equivalent circuit in Chapter 16.1.2 to a network setting.

15.2.4 *YY* configuration

Referring to Figure 15.12 and combining the variables defined in Chapter 15.2.2 for each configuration, the internal voltages and currents associated with the ideal transformer are:

$$\hat{V}^Y_j := \begin{bmatrix} \hat{V}^{an}_j \\ \hat{V}^{bn}_j \\ \hat{V}^{cn}_j \end{bmatrix}, \quad \hat{I}^Y_j := \begin{bmatrix} \hat{I}^{an}_j \\ \hat{I}^{bn}_j \\ \hat{I}^{cn}_j \end{bmatrix}, \quad \hat{V}^Y_k := \begin{bmatrix} \hat{V}^{an}_k \\ \hat{V}^{bn}_k \\ \hat{V}^{cn}_k \end{bmatrix}, \quad \hat{I}^Y_k := \begin{bmatrix} \hat{I}^{an}_k \\ \hat{I}^{bn}_k \\ \hat{I}^{cn}_k \end{bmatrix}$$

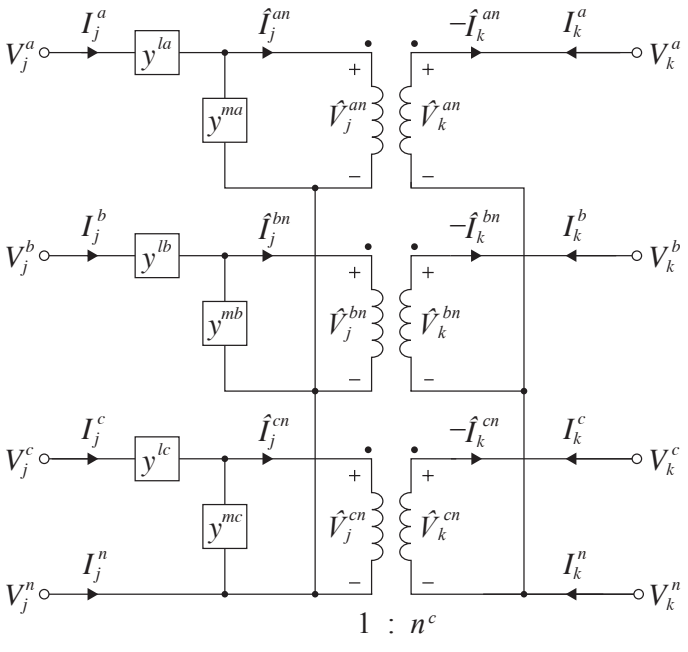


Figure 15.12 YY-configured transformer.

The terminal voltages and currents are:

$$V_j := \begin{bmatrix} V_j^a \\ V_j^b \\ \hat{V}_j^c \end{bmatrix}, \qquad I_j := \begin{bmatrix} I_j^a \\ I_j^b \\ \hat{I}_j^c \end{bmatrix}, \qquad V_k := \begin{bmatrix} V_k^a \\ V_k^b \\ \hat{V}_k^c \end{bmatrix}, \qquad I_k := \begin{bmatrix} I_k^a \\ I_k^b \\ \hat{I}_k^c \end{bmatrix}$$

as well as the the neutral voltages and currents $\left(V_j^n, I_j^n\right)$ and $\left(V_k^n, I_k^n\right)$ as shown in the figure. The relation between the internal and terminal variables is given by (15.13a) and (15.14a) for Y configurations on the primary and secondary sides respectively:

$$I_j = y^l\left(V_j - V_j^n\mathbf{1} - \hat{V}_j^Y\right), \qquad I_j = y^m\hat{V}_j^Y + \hat{I}_j^Y, \qquad I_j^n = -\mathbf{1}^\mathsf{T} I_j \tag{15.18a}$$

$$V_k = \hat{V}_k^Y + V_k^n\mathbf{1}, \qquad I_k = \hat{I}_k^Y, \qquad I_k^n = -\mathbf{1}^\mathsf{T} I_k \tag{15.18b}$$

The transformer gains that relate the internal variables are:

$$\hat{V}_k^Y = n\hat{V}_j^Y, \qquad \hat{I}_k^Y = -a\hat{I}_j^Y \tag{15.18c}$$

Here $y^l := \operatorname{diag}\left(y^{la}, y^{lb}, y^{lc}\right)$ is the leakage admittance matrix, $y^m := \operatorname{diag}\left(y^{ma}, y^{mb}, y^{mc}\right)$ is the shunt admittance matrix, $n := \operatorname{diag}\left(n^a, n^b, n^c\right)$ is the voltage gain matrix and $a := n^{-1}$ is the turns ratio matrix.

We can derive an external model that relates the terminal variables by eliminating

the internal variables from (15.18). Specifically we have from (15.18a)(15.18b)

$$\begin{aligned}\hat{V}_j^Y &= (V_j - V_j^n\mathbf{1}) - \left(y^l\right)^{-1} I_j, & \hat{V}_k^Y &= V_k - V_k^n\mathbf{1}\\ \hat{I}_j^Y &= I_j - y^m(V_j - V_j^n\mathbf{1}) + y^m\left(y^l\right)^{-1} I_j, & \hat{I}_k^Y &= I_k\end{aligned}$$

Substituting it into (15.18c) yields the external model of a three-phase transformer in YY configuration:

$$\begin{bmatrix} I_j \\ I_k \end{bmatrix} = \underbrace{\begin{bmatrix} y^l & -ay^l \\ -ay^l & a^2(y^l+y^m) \end{bmatrix}}_{Y_{YY}} \left(\begin{bmatrix} V_j \\ V_k \end{bmatrix} - \begin{bmatrix} V_j^n\mathbf{1} \\ V_k^n\mathbf{1} \end{bmatrix} \right) \tag{15.19a}$$

$$I_j^n = -\mathbf{1}^{\mathsf{T}} I_j, \qquad I_k^n = -\mathbf{1}^{\mathsf{T}} I_k \tag{15.19b}$$

where we have used $y^l a = a y^l$ and $a(y^l+y^m)a = a^2(y^l+y^m)$ since they are all diagonal matrices. The expression (15.19a) is the same as the external model (15.12a) for a single-phase transformer, except that, instead of scalars, the variables (V_j, I_j, V_k, I_k) are vectors in $\mathbb{C}^3$ and the parameters a, y^l, y^m are 3×3 matrices. It is the expression (15.16).

We can also express the neutral currents $\left(I_j^n, I_k^n\right)$ in terms of the terminal voltages instead of the terminal currents using (15.19a)(15.19b):

$$\begin{bmatrix} I_j^n \\ I_k^n \end{bmatrix} = -\underbrace{\begin{bmatrix} \mathbf{1}^{\mathsf{T}} & 0 \\ 0 & \mathbf{1}^{\mathsf{T}} \end{bmatrix} Y_{YY}}_{Y_{YY}^n} \left(\begin{bmatrix} V_j \\ V_k \end{bmatrix} - \begin{bmatrix} V_j^n\mathbf{1} \\ V_k^n\mathbf{1} \end{bmatrix} \right)$$

A four-wire model includes the neutral currents. To derive the four-wire model we rewrite this and (15.19a) as

$$\begin{bmatrix} I_j \\ I_k \end{bmatrix} = \underbrace{\begin{bmatrix} y^l & -ay^l \\ -ay^l & a^2(y^l+y^m) \end{bmatrix}}_{Y_{YY}} \begin{bmatrix} V_j \\ V_k \end{bmatrix} - \underbrace{\begin{bmatrix} y^l\mathbf{1} & -ay^l\mathbf{1} \\ -ay^l\mathbf{1} & a^2(y^l+y^m)\mathbf{1} \end{bmatrix}}_{Y_{YY}\,(\mathbb{I}_2\otimes\mathbf{1})} \begin{bmatrix} V_j^n \\ V_k^n \end{bmatrix}$$

$$\begin{bmatrix} I_j^n \\ I_k^n \end{bmatrix} = -\underbrace{\begin{bmatrix} \mathbf{1}^{\mathsf{T}}y^l & -\mathbf{1}^{\mathsf{T}}ay^l \\ -\mathbf{1}^{\mathsf{T}}ay^l & \mathbf{1}^{\mathsf{T}}a^2(y^l+y^m) \end{bmatrix}}_{(\mathbb{I}_2\otimes\mathbf{1}^{\mathsf{T}})Y_{YY}} \begin{bmatrix} V_j \\ V_k \end{bmatrix} + \underbrace{\begin{bmatrix} \mathbf{1}^{\mathsf{T}}y^l\mathbf{1} & -\mathbf{1}^{\mathsf{T}}ay^l\mathbf{1} \\ -\mathbf{1}^{\mathsf{T}}ay^l\mathbf{1} & \mathbf{1}^{\mathsf{T}}a^2(y^l+y^m)\mathbf{1} \end{bmatrix}}_{(\mathbb{I}_2\otimes\mathbf{1}^{\mathsf{T}})Y_{YY}\,(\mathbb{I}_2\otimes\mathbf{1})} \begin{bmatrix} V_j^n \\ V_k^n \end{bmatrix}$$

where $\mathbb{I}_2$ is the identity matrix of size 2, $\mathbf{1}^{\mathsf{T}}y^l\mathbf{1} = \sum_\phi y^{l\phi}$, $\mathbf{1}^{\mathsf{T}}ay^l\mathbf{1} = \sum_\phi a^\phi y^{l\phi}$, and $\mathbf{1}^{\mathsf{T}}a^2(y^l+y^m)\mathbf{1} = \sum_\phi (a^\phi)^2(y^{l\phi}+y^{m\phi})$. Hence the four-wire model of a three-phase transformer in YY configuration is:

$$\begin{bmatrix} I_j \\ I_k \\ I_j^n \\ I_k^n \end{bmatrix} = \underbrace{\begin{bmatrix} Y_{YY} & -Y_{YY}\,(\mathbb{I}_2\otimes\mathbf{1}) \\ -\left(\mathbb{I}_2\otimes\mathbf{1}^{\mathsf{T}}\right)Y_{YY} & \left(\mathbb{I}_2\otimes\mathbf{1}^{\mathsf{T}}\right)Y_{YY}\,(\mathbb{I}_2\otimes\mathbf{1}) \end{bmatrix}}_{Y_{YY}^{\text{4wire}}} \begin{bmatrix} V_j \\ V_k \\ V_j^n \\ V_k^n \end{bmatrix} \tag{15.19c}$$

This model extends (15.12b) with neutral currents to three-phase transformers. The matrix Y_{YY} in (15.19a) is both symmetric and block symmetric (see Chapter 15.2.3) because a, y^l and y^m are diagonal. This, together with $(A \otimes B)^\mathsf{T} = A^\mathsf{T} \otimes B^\mathsf{T}$, imply that the four-wire admittance matrix Y_{YY}^{4wire} is also symmetric. While the admittance matrix Y_{YY} generally has nonzero row and column sums, Y_{YY}^{4wire} has zero row and column sums.

If both neutrals are grounded with zero impedances and voltages are defined with respect to the ground, then $V_j^n = V_k^n = 0$ and (15.19a) reduces to

$$\begin{bmatrix} I_j \\ I_k \end{bmatrix} = Y_{YY} \begin{bmatrix} V_j \\ V_k \end{bmatrix} = \begin{bmatrix} y^l & -ay^l \\ -ay^l & a^2(y^l + y^m) \end{bmatrix} \begin{bmatrix} V_j \\ V_k \end{bmatrix}$$

which can be represented as a three-phase Π circuit. This means that the external behavior of a YY transformer, when its neutral voltages are zero, has the same structure as that of a three-phase transmission line and can be specified by 3×3 series and shunt admittance matrices $\left(\tilde{y}_{jk}^s, \tilde{y}_{jk}^m, \tilde{y}_{kj}^m\right)$ where

$$\tilde{y}_{jk}^s := ay^l, \qquad \tilde{y}_{jk}^m := (\mathbb{I} - a)\, y^l, \qquad \tilde{y}_{kj}^m := a(a - \mathbb{I})y^l + a^2 y^m \tag{15.19d}$$

This extends the single-phase Π circuit model (15.12c) to the three-phase setting.

15.2.5 $\Delta\Delta$ configuration

Referring to Figure 15.13, and combining the variables defined in Chapter 15.2.2 for each configuration, the internal voltages and currents associated with the ideal transformer are:

$$\hat{V}_j^\Delta := \begin{bmatrix} \hat{V}_j^{ab} \\ \hat{V}_j^{bc} \\ \hat{V}_j^{ca} \end{bmatrix}, \qquad \hat{I}_j^\Delta := \begin{bmatrix} \hat{I}_j^{ab} \\ \hat{I}_j^{bc} \\ \hat{I}_j^{ca} \end{bmatrix}, \qquad \hat{V}_k^\Delta := \begin{bmatrix} \hat{V}_k^{ab} \\ \hat{V}_k^{bc} \\ \hat{V}_k^{ca} \end{bmatrix}, \qquad \hat{I}_k^\Delta := \begin{bmatrix} \hat{I}_k^{ab} \\ \hat{I}_k^{bc} \\ \hat{I}_k^{ca} \end{bmatrix}$$

The terminal voltages and currents are denoted by (V_j, I_j), (V_k, I_k), as for a YY-configured transformer. The relation between the internal and terminal variables is given by (15.13b) and (15.14b) for Δ configurations:

$$\hat{I}_j^\Delta = y^l \Gamma V_j - (y^l + y^m)\hat{V}_j^\Delta, \qquad I_j = \Gamma^\mathsf{T}\left(\hat{I}_j^\Delta + y^m \hat{V}_j^\Delta\right) \tag{15.20a}$$

$$\hat{V}_k^\Delta = \Gamma V_k, \qquad I_k = \Gamma^\mathsf{T} \hat{I}_k^\Delta \tag{15.20b}$$

The transformer gains that relate the internal variables are:

$$\hat{V}_k^\Delta = n \hat{V}_j^\Delta, \qquad \hat{I}_k^\Delta = -a \hat{I}_j^\Delta \tag{15.20c}$$

To derive an external model, eliminate the internal variables from (15.20). We obtain from (15.20b)(15.20c):

$$\hat{V}_j^\Delta = n^{-1} \hat{V}_k^\Delta = a \Gamma V_k, \qquad \Gamma^\mathsf{T} a \hat{I}_j^\Delta = -I_k$$

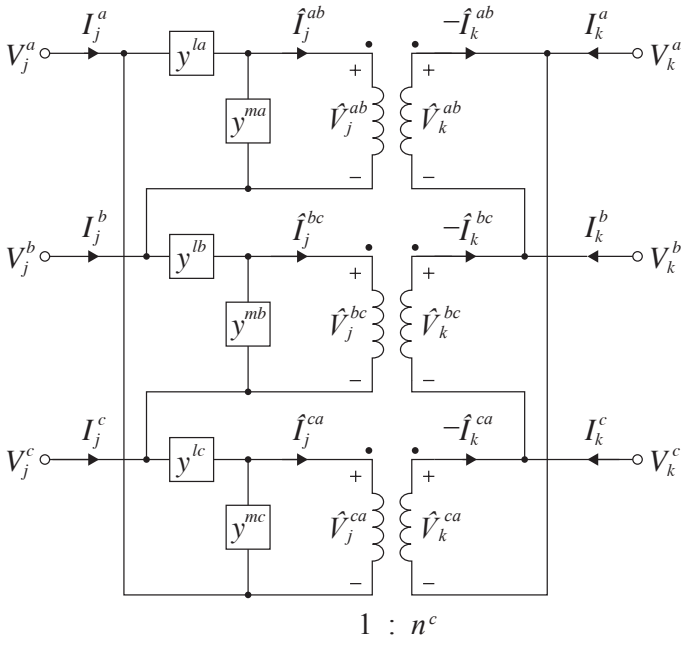


Figure 15.13 $\Delta\Delta$-configured transformer.

Substitute into the first expression in (15.20a) to eliminate $(\hat{V}_j^\Delta, \hat{I}_j^\Delta)$:

$$I_k = -\left(\Gamma^\mathsf{T} a y^l \Gamma\right) V_j + \left(\Gamma^\mathsf{T} a^2 (y^l + y^m)\Gamma\right) V_k$$

Substitute again $\hat{V}_j^\Delta$ into the first expression in (15.20a) to obtain $\hat{I}_j^\Delta = y^l \Gamma V_j - a(y^l + y^m)\Gamma V_k$. Substitute this and $\hat{V}_j^\Delta$ into the second expression in (15.20a) to eliminate $(\hat{V}_j^\Delta, \hat{I}_j^\Delta)$:

$$I_j = \left(\Gamma^\mathsf{T} y^l \Gamma\right) V_j - \left(\Gamma^\mathsf{T} a y^l \Gamma\right) V_k$$

The external model of a three-phase transformer in $\Delta\Delta$ configuration is hence

$$\begin{bmatrix} I_j \\ I_k \end{bmatrix} = \underbrace{\begin{bmatrix} \Gamma^\mathsf{T} y^l \Gamma & -\Gamma^\mathsf{T} a y^l \Gamma \\ -\Gamma^\mathsf{T} a y^l \Gamma & \Gamma^\mathsf{T} a^2 (y^l + y^m)\Gamma \end{bmatrix}}_{Y_{\Delta\Delta}} \begin{bmatrix} V_j \\ V_k \end{bmatrix} \tag{15.21a}$$

or in terms of the admittance matrix Y_{YY} in (15.19a) for a YY-configured transformer:

$$\begin{bmatrix} I_j \\ I_k \end{bmatrix} = \begin{bmatrix} \Gamma^\mathsf{T} & 0 \\ 0 & \Gamma^\mathsf{T} \end{bmatrix} \underbrace{\begin{bmatrix} y^l & -a y^l \\ -a y^l & a^2 (y^l + y^m) \end{bmatrix}}_{Y_{YY}} \begin{bmatrix} \Gamma & 0 \\ 0 & \Gamma \end{bmatrix} \begin{bmatrix} V_j \\ V_k \end{bmatrix} \tag{15.21b}$$

This is the expression (15.16). Unlike Y_{YY} the admittance matrix $Y_{\Delta\Delta}$ is not invertible (it has zero row and column sums). Since $Y_{\Delta\Delta}$ is block symmetric (as well as symmetric)

it can be represented as a three-phase Π circuit. This means that its external behavior has the same structure as that of a three-phase transmission line and can be specified by 3×3 series and shunt admittance matrices $\left(\tilde{y}^s_{jk}, \tilde{y}^m_{jk}, \tilde{y}^m_{kj}\right)$ where

$$\tilde{y}^s_{jk} := \Gamma^\mathsf{T} a y^l \Gamma, \quad \tilde{y}^m_{jk} := \Gamma^\mathsf{T}(\mathbb{I}-a)\, y^l \Gamma, \quad \tilde{y}^m_{kj} := \Gamma^\mathsf{T}\left(a(a-\mathbb{I})y^l + a^2 y^m\right)\Gamma \tag{15.21c}$$

This is the Π circuit model (15.19d) for YY-configured transformer, multiplied on both sides by Γ^T and Γ.

The submatrices in (15.21b) are (cf. Y^Δ in (14.21a)):

$$\Gamma^\mathsf{T} y^l \Gamma = \begin{bmatrix} y^{la}+y^{lc} & -y^{la} & -y^{lc} \\ -y^{la} & y^{lb}+y^{la} & -y^{lb} \\ -y^{lc} & -y^{lb} & y^{lc}+y^{lb} \end{bmatrix}, \quad \Gamma^\mathsf{T} a y^l \Gamma = \begin{bmatrix} \hat{y}^{la}+\hat{y}^{lc} & -\hat{y}^{la} & -\hat{y}^{lc} \\ -\hat{y}^{la} & \hat{y}^{lb}+\hat{y}^{la} & -\hat{y}^{lb} \\ -\hat{y}^{lc} & -\hat{y}^{lb} & \hat{y}^{lc}+\hat{y}^{lb} \end{bmatrix}$$

where $\hat{y}^{l\phi} := a^\phi y^{l\phi}$ for $\phi \in \{a,b,c\}$. In the special case where the single-phase transformers are identical, i.e., $y^l = y^{la}\mathbb{I}$ and $a := a^a\mathbb{I}$, these matrices are particularly simple:

$$\left(y^{la}\right)\Gamma^\mathsf{T}\Gamma = y^{la}\begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}, \quad \left(a^a y^{la}\right)\Gamma^\mathsf{T}\Gamma = a^a y^{la}\begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix} \tag{15.22}$$

These expressions are often used in simplified models of three-phase transformers.

15.2.6 ΔY configuration

This is a popular configuration for stepdown transformers in distribution systems. Referring to Figure 15.14, the internal voltages and currents associated with the ideal transformer are:

$$\hat{V}^\Delta_j := \begin{bmatrix} \hat{V}^{ab}_j \\ \hat{V}^{bc}_j \\ \hat{V}^{ca}_j \end{bmatrix}, \quad \hat{I}^\Delta_j := \begin{bmatrix} \hat{I}^{ab}_j \\ \hat{I}^{bc}_j \\ \hat{I}^{ca}_j \end{bmatrix}, \quad \hat{V}^Y_k := \begin{bmatrix} \hat{V}^{an}_k \\ \hat{V}^{bn}_k \\ \hat{V}^{cn}_k \end{bmatrix}, \quad \hat{I}^Y_k := \begin{bmatrix} \hat{I}^{an}_k \\ \hat{I}^{bn}_k \\ \hat{I}^{cn}_k \end{bmatrix}$$

The terminal voltages and currents are denoted by (V_j, I_j), (V_k, I_k), as before. The relation between the internal and terminal variables is given by (15.13b) for Δ configuration on the primary side and (15.14a) for Y configuration on the secondary side:

$$\hat{I}^\Delta_j = y^l \Gamma V_j - (y^l + y^m)\hat{V}^\Delta_j, \quad I_j = \Gamma^\mathsf{T}\left(\hat{I}^\Delta_j + y^m \hat{V}^\Delta_j\right) \tag{15.23a}$$

$$V_k = \hat{V}^Y_k + V^n_k \mathbf{1}, \qquad I_k = \hat{I}^Y_k, \qquad I^n_k = -\mathbf{1}^\mathsf{T} I_k \tag{15.23b}$$

The transformer gains that relate the internal variables are:

$$\hat{V}^Y_k = n\hat{V}^\Delta_j, \qquad \hat{I}^Y_k = -a\hat{I}^\Delta_j \tag{15.23c}$$

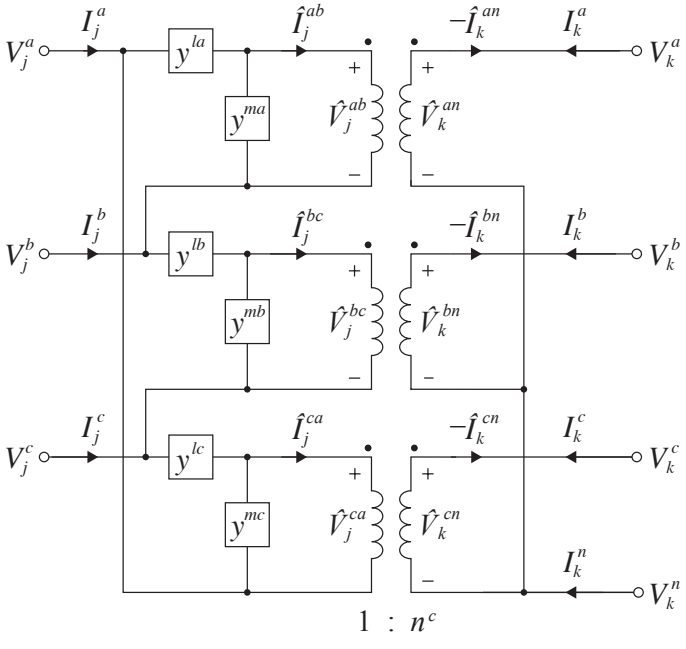


Figure 15.14 ΔY-configured transformer.

Eliminating the internal variables from (15.23), the external model of a three-phase transformer in ΔY configuration is (Exercise 15.2):

$$\begin{bmatrix} I_j \\ I_k \end{bmatrix} = \underbrace{\begin{bmatrix} \Gamma^{\mathsf{T}} y^l \Gamma & -\Gamma^{\mathsf{T}} a y^l \\ -a y^l \Gamma & a^2 (y^l + y^m) \end{bmatrix}}_{Y_{\Delta Y}} \begin{bmatrix} V_j \\ V_k \end{bmatrix} - \begin{bmatrix} -\Gamma^{\mathsf{T}} a y^l \\ a^2 (y^l + y^m) \end{bmatrix} V_k^n \mathbf{1} \tag{15.24a}$$

or in terms of the admittance matrix Y_{YY} in (15.19a):

$$\begin{bmatrix} I_j \\ I_k \end{bmatrix} = \begin{bmatrix} \Gamma^{\mathsf{T}} & 0 \\ 0 & \mathbb{I} \end{bmatrix} \underbrace{\begin{bmatrix} y^l & -a y^l \\ -a y^l & a^2 (y^l + y^m) \end{bmatrix}}_{Y_{YY}} \begin{bmatrix} \Gamma & 0 \\ 0 & \mathbb{I} \end{bmatrix} \left(\begin{bmatrix} V_j \\ V_k \end{bmatrix} - \begin{bmatrix} V_j^n \mathbf{1} \\ V_k^n \mathbf{1} \end{bmatrix} \right) \tag{15.24b}$$

It is the expression (15.16). The matrix $Y_{\Delta Y}$ in (15.24a) is not invertible. It is symmetric but not block symmetric. Therefore it cannot be represented as a three-phase Π circuit even if the neutral voltage $V_k^n = 0$.

Even though there is no neutral line on the primary side, the primary current I_j is affected by the neutral voltage V_k^n on the secondary side, unless $a = a^a \mathbb{I}$ and $y = y^a \mathbb{I}$, i.e., the single-phase transformers are identical, in which case $\Gamma^{\mathsf{T}} \mathbf{1} = 0$ and I_j becomes independent of V_k^n.

15.2.7 $Y\Delta$ configuration

Figure 15.15 shows a $Y\Delta$-configured three-phase transformer. Its external model is

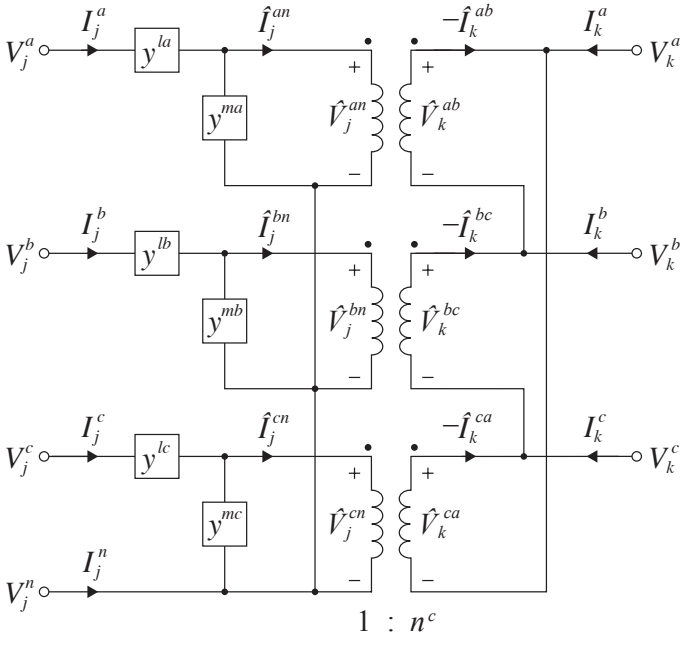


Figure 15.15 $Y\Delta$-configured transformer.

(Exercise 15.3):

$$\begin{bmatrix} I_j \\ I_k \end{bmatrix} = \underbrace{\begin{bmatrix} y^l & -ay^l\Gamma \\ -\Gamma^{\mathsf{T}}ay^l & \Gamma^{\mathsf{T}}a^2(y^l+y^m)\Gamma \end{bmatrix}}_{Y_{Y\Delta}} \begin{bmatrix} V_j \\ V_k \end{bmatrix} - \begin{bmatrix} y^l \\ -\Gamma^{\mathsf{T}}ay^l \end{bmatrix} V_j^n \mathbf{1} \tag{15.25a}$$

or in terms of the admittance matrix Y_{YY} in (15.19a):

$$\begin{bmatrix} I_j \\ I_k \end{bmatrix} = \begin{bmatrix} \mathbb{I} & 0 \\ 0 & \Gamma^{\mathsf{T}} \end{bmatrix} \underbrace{\begin{bmatrix} y^l & -ay^l \\ -ay^l & a^2(y^l+y^m) \end{bmatrix}}_{Y_{YY}} \begin{bmatrix} \mathbb{I} & 0 \\ 0 & \Gamma \end{bmatrix} \left(\begin{bmatrix} V_j \\ V_k \end{bmatrix} - \begin{bmatrix} V_j^n \mathbf{1} \\ V_k^n \mathbf{1} \end{bmatrix} \right) \tag{15.25b}$$

It is the expression (15.16). The matrix $Y_{Y\Delta}$ is singular, symmetric but not block symmetric. In particular it cannot be represented as a three-phase Π circuit even if the neutral voltage $V_j^n = 0$.

15.2.8 Open transformer

Open transformers where at least one leg of a three-phase transformer is open (not connected) are widely used in distribution systems to connect single-phase loads, e.g., a household. The analysis of a closed transformer can be adapted to that of an open transformer. Indeed their external models are identical, except that the admittance matrices are $\tilde{y}^l = \text{diag}\left(y^{la}, y^{lb}, 0\right)$ and $\tilde{y}^m = \text{diag}\left(y^{ma}, y^{mb}, 0\right)$ for an open transformer without the third leg (compare (15.21) with (15.26) for an open $\Delta\Delta$ transformer). We now derive the external model of an open $\Delta\Delta$ transformer. Other configurations, such as open YY, open ΔY, or open $Y\Delta$, can be analyzed in a similar manner. The analysis proceeds in the same manner as for its closed version, once the voltage gain expression has been modified to represent the open transformer leg where the internal voltages $\hat{V}_j^{ca}$ and $\hat{V}_k^{ca}$ are no longer related by a voltage gain.

Figure 15.16 shows an open $\Delta\Delta$-configured transformer where only two single-phase transformers are used. The leakage admittances of these transformers are $\left(y^a, y^b\right)$ and

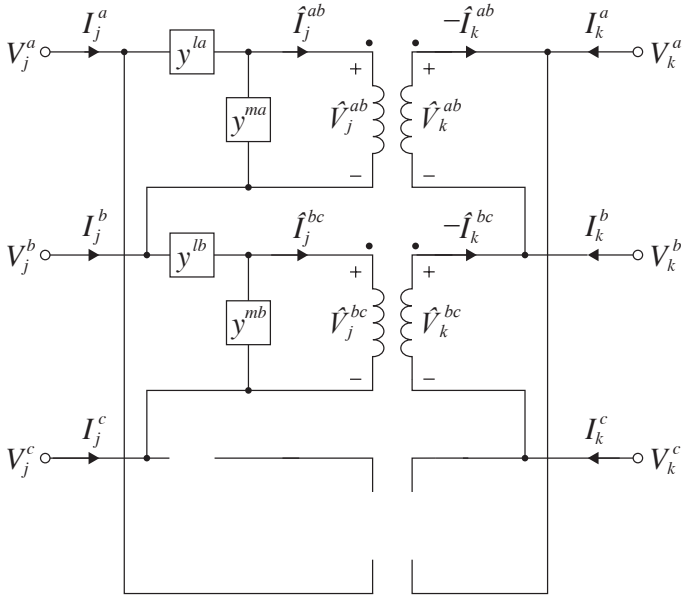


Figure 15.16 Open $\Delta\Delta$-configured transformer.

their voltage gains are $\left(n^a, n^b\right)$. The internal voltages and currents associated with the ideal transformer are:

$$\hat{V}_j^{\Delta} := \begin{bmatrix} \hat{V}_j^{ab} \\ \hat{V}_j^{bc} \\ \hat{V}_j^{ca} \end{bmatrix}, \qquad \hat{I}_j^{\Delta} := \begin{bmatrix} \hat{I}_j^{ab} \\ \hat{I}_j^{bc} \\ \hat{I}_j^{ca} \end{bmatrix}, \qquad \hat{V}_k^{\Delta} := \begin{bmatrix} \hat{V}_k^{ab} \\ \hat{V}_k^{bc} \\ \hat{V}_k^{ca} \end{bmatrix}, \qquad \hat{I}_k^{\Delta} := \begin{bmatrix} \hat{I}_k^{ab} \\ \hat{I}_k^{bc} \\ \hat{I}_k^{ca} \end{bmatrix}$$

The terminal voltages and currents are denoted by $(V_j, I_j) \in \mathbb{C}^6$, $(V_k, I_k) \in \mathbb{C}^6$, as

before. We will show that its external model is

$$\begin{bmatrix} I_j \\ I_k \end{bmatrix} = \underbrace{\begin{bmatrix} \Gamma^\mathsf{T}\tilde{y}^l\Gamma & -\Gamma^\mathsf{T}a\tilde{y}^l\Gamma \\ -\Gamma^\mathsf{T}a\tilde{y}^l\Gamma & \Gamma^\mathsf{T}a^2(\tilde{y}^l+\tilde{y}^m)\Gamma \end{bmatrix}}_{Y_{\text{open}\Delta\Delta}} \begin{bmatrix} V_j \\ V_k \end{bmatrix} \tag{15.26a}$$

or

$$\begin{bmatrix} I_j \\ I_k \end{bmatrix} = \begin{bmatrix} \Gamma^\mathsf{T} & 0 \\ 0 & \Gamma^\mathsf{T} \end{bmatrix} \begin{bmatrix} \tilde{y}^l & -a\tilde{y}^l \\ -a\tilde{y}^l & a^2(\tilde{y}^l+\tilde{y}^m) \end{bmatrix} \begin{bmatrix} \Gamma & 0 \\ 0 & \Gamma \end{bmatrix} \begin{bmatrix} V_j \\ V_k \end{bmatrix} \tag{15.26b}$$

where

$$\tilde{y}^l := \begin{bmatrix} y^{la} & 0 & 0 \\ 0 & y^{lb} & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad \tilde{y}^m := \begin{bmatrix} y^{ma} & 0 & 0 \\ 0 & y^{mb} & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{15.26c}$$

where $a := \text{diag}\left(a^a, a^b, a^c\right)$. The constant a^c is introduced for notational convenience and can take any arbitrary nonzero finite value, e.g. $a^c = 1$, as its value does not affect the external model. Hence the admittance matrix $Y_{\text{open}\Delta\Delta}$ in (15.26a)(15.26b) are the same as $Y_{\Delta\Delta}$ in (15.21a)(15.21b) for a closed $\Delta\Delta$ transformer, except that $y^{lc} = y^{mc} = 0$ on the third leg that has no transformer. It is also the same as the expression (15.16) with (y^l, y^m) in Y_{YY} replaced by $(\tilde{y}^l, \tilde{y}^m)$. The matrix $Y_{\text{open}\Delta\Delta}$ is block symmetric (as well as symmetric) and therefore has a three-phase Π circuit representation with series and shunt admittance matrices:

$$\tilde{y}^s_{jk} := \Gamma^\mathsf{T}a\tilde{y}^l\Gamma, \quad \tilde{y}^m_{jk} := \Gamma^\mathsf{T}(\mathbb{I}-a)\,\tilde{y}^l\Gamma, \quad \tilde{y}^m_{kj} := \Gamma^\mathsf{T}\left(a(a-\mathbb{I})\tilde{y}^l + a^2\tilde{y}^m\right)\Gamma \tag{15.26d}$$

which is the same as (15.21c) with (y^l, y^m) replaced by $(\tilde{y}^l, \tilde{y}^m)$.

For notational convenience, we introduce an artificial voltage gain n^c which can take any nonzero finite values, e.g., $n^c := 1$. As before let $n := \text{diag}\left(n^a, n^b, n^c\right)$ and $a := n^{-1}$. As defined above, the leakage and magnetizing admittances are $\tilde{y}^l := \text{diag}\left(y^{la}, y^{lb}, 0\right)$ and $\tilde{y}^m := \text{diag}\left(y^{ma}, y^{mb}, 0\right)$ respectively. The fact that the third leg of the transformer is open requires two adjustments to the derivation of a closed $\Delta\Delta$ transformer. These adjustments modify the internal model (the current and voltage gain on the missing leg) and the derivation then follows the same procedure, as we now explain.

1 The relation between the internal and terminal variables are still given by (15.20a)(15.20b) with the following modifications: replace (y^l, y^m) by $(\tilde{y}^l, \tilde{y}^m)$ and enforce the current on the missing leg on the secondary side to be zero (see Figure 15.16):

$$\tilde{y}^{lc} := 0, \qquad \tilde{y}^{mc} := 0, \qquad \hat{I}^{ca}_k := 0 \tag{15.27a}$$

This implies that $\hat{I}^{ca}_j = 0$ and $I^c_j = -\hat{I}^{bc}_j$ on the primary side from the last row of (15.20a).

2 For the internal model (15.20c), the current gain $\hat{I}_k^\Delta = -a\,\hat{I}_j^\Delta$ remains unchanged (given (15.27a)), but the voltage gain needs modification because the internal voltages $\hat{V}_k^{ca} := V_k^c - V_k^a$ and $\hat{V}_j^{ca} := V_j^c - V_j^a$ are no longer related by the voltage gain n, unlike in a closed transformer.

In order to follow the same derivation we will replace the voltage gain expression $\hat{V}_j^\Delta = a\hat{V}_k^\Delta$ in (15.20c), as follows. In the analysis of a closed $\Delta\Delta$ transformer, the voltage gain is used to relate $\hat{V}_j^\Delta$ to V_k through

$$\hat{V}_j^\Delta \;=\; a\hat{V}_k^\Delta \;=\; a\,\Gamma V_k$$

For an open $\Delta\Delta$ transformer, the last row of this relation is rewritten as:

$$\hat{V}_j^{ca} \;=\; a^c\hat{V}_k^{ca} + \left(\hat{V}_j^{ca} - a^c\hat{V}_k^{ca}\right)$$

leading to the voltage relation $\hat{V}_j^\Delta = a\,\hat{V}_k^\Delta + E_3\left(\hat{V}_j^\Delta - a\hat{V}_k^\Delta\right)$ where $E_3 := \mathrm{diag}(0,0,1)$. The right-hand side can then be written in terms of the *terminal* voltage V_j because $\hat{V}_j^{ca} := V_j^c - V_j^a$:

$$\hat{V}_j^\Delta \;=\; E_3\Gamma V_j + (\mathbb{I} - E_3)\,a\,\hat{V}_k^\Delta \tag{15.27b}$$

which can then be related to V_k using $\hat{V}_k^\Delta = \Gamma V_k$.

In summary, these two modifications (15.27) means that, for open $\Delta\Delta$ transformer, the conversion rules are (15.20a)(15.20b) with (y^l, y^m) replaced by $(\tilde{y}^l, \tilde{y}^m)$:

$$\hat{I}_j^\Delta \;=\; \tilde{y}^l\Gamma V_j - (\tilde{y}^l + \tilde{y}^m)\hat{V}_j^\Delta, \qquad I_j \;=\; \Gamma^\mathsf{T}\left(\hat{I}_j^\Delta + \tilde{y}^m\hat{V}_j^\Delta\right) \tag{15.28a}$$

$$\hat{V}_k^\Delta \;=\; \Gamma V_k, \qquad I_k \;=\; \Gamma^\mathsf{T}\hat{I}_k^\Delta \tag{15.28b}$$

and the internal model (15.20c) is replaced by:

$$\hat{V}_j^\Delta \;=\; E_3\Gamma V_j + (\mathbb{I} - E_3)\,a\,\hat{V}_k^\Delta, \qquad \hat{I}_k^\Delta \;=\; -a\,\hat{I}_j^\Delta \tag{15.28c}$$

We then follow the same derivation for the external model. For example we obtain from (15.28b)(15.28c):

$$\hat{V}_j^\Delta \;=\; E_3\Gamma V_j + (\mathbb{I} - E_3)\,a\,\Gamma V_k, \qquad \Gamma^\mathsf{T} a\hat{I}_j^\Delta \;=\; -I_k$$

Substitute into the first expression in (15.28a) to eliminate $(\hat{V}_j^\Delta, \hat{I}_j^\Delta)$:

$$I_k \;=\; -\left(\Gamma^\mathsf{T} a\tilde{y}^l\Gamma\right)V_j + \left(\Gamma^\mathsf{T} a^2(\tilde{y}^l + \tilde{y}^m)\Gamma\right)V_k$$

where we have used $(\tilde{y}^l + \tilde{y}^m)E_3 = 0$. Similarly we have

$$I_j \;=\; \left(\Gamma^\mathsf{T}\tilde{y}^l\Gamma\right)V_j - \left(\Gamma^\mathsf{T} a\tilde{y}^l\Gamma\right)V_k$$

verifying the external model (15.26). With $y^{lc} = y^{mc} = 0$ the matrices are explicitly:

$$\Gamma^{\mathsf{T}} \tilde{y}^l \Gamma = \begin{bmatrix} y^{la} & -y^{la} & 0 \\ -y^{la} & y^{lb} + y^{la} & -y^{lb} \\ 0 & -y^{lb} & y^{lb} \end{bmatrix}, \quad \Gamma^{\mathsf{T}} a\tilde{y}^l \Gamma = \begin{bmatrix} \hat{y}^{la} & -\hat{y}^{la} & 0 \\ -\hat{y}^{la} & \hat{y}^{lb} + \hat{y}^{la} & -\hat{y}^{lb} \\ 0 & -\hat{y}^{lb} & \hat{y}^{lb} \end{bmatrix}$$

where $\hat{y}^{l\phi} := a^{\phi} y^{l\phi}$ for $\phi \in \{a, b\}$.

Example 15.6 (Bernie Leseiutre, Allerton Conference, September 2023)**.** Bernie Leseiutre told me about an interesting circulating loop flow phenomenon in an open $\Delta\Delta$ transformer, shown in Figure He said that even if the Δ load is purely inductive,

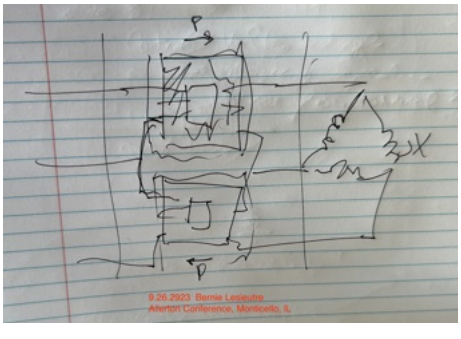


Figure 15.17 Unitary voltage network in each phase ϕ of a three-phase transformer.

there is real power P flowing between the two single-phase transformers, even if the transformers are (assumed) ideal. They have verified this experimentally. The terminal currents/powers are purely reactive, so real current/power only are in internal vars. This show be derivable from the results here. □

15.2.9 Single-phase equivalent in balanced setting

A three-phase transformer is equivalent to a YY-configured transformer if they have the same external model, i.e., their admittance matrices are equal. In general a three-phase transformer not in YY configuration does not have a YY equivalent, except in a balanced setting. In a balanced setting, not only does a three-phase transformer have a YY equivalent, there is also a single-phase transformer that can be naturally interpreted as the *single-phase equivalent* of the YY equivalent. For simplicity we assume $y^m = 0$.

Consider a $\Delta\Delta$-configured transformer whose external model is determined by the admittance matrix $Y_{\Delta\Delta}$ in (15.21b), reproduced here:

$$Y_{\Delta\Delta} := \begin{bmatrix} \Gamma^{\mathsf{T}} & 0 \\ 0 & \Gamma^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} y^l & -ay^l \\ -ay^l & a^2 y^l \end{bmatrix} \begin{bmatrix} \Gamma & 0 \\ 0 & \Gamma \end{bmatrix}$$

Recall from (15.19) that the admittance matrix $\tilde{Y}_{YY}$ of a YY-configured transformer

with turns ratio $\tilde{a}$ and leakage admittance $\tilde{y}^l$ is given by

$$\tilde{Y}_{YY} \;:=\; \begin{bmatrix} \tilde{y}^l & -\tilde{a}\tilde{y}^l \\ -\tilde{a}\tilde{y}^l & \tilde{a}^2\tilde{y}^l \end{bmatrix}$$

The $\Delta\Delta$-configured transformer has a YY equivalent if $Y_{\Delta\Delta} = \tilde{Y}_{YY}$ for some $\tilde{Y}_{YY}$. Since the submatrices of $\tilde{Y}_{YY}$ are diagonal while those of $Y_{\Delta\Delta}$ are not, there is generally no YY equivalent, even if the constituent single-phase transformers are identical, i.e., if $y^l = y^{la}\mathbb{I}$ and $a = a^a\mathbb{I}$ (see (15.22)).

The $\Delta\Delta$-configured transformer does have a YY equivalent, however, if the system is balanced, i.e., the single-phase transformers are identical and voltages and currents are positive-sequence sets. This property is used in Chapter 3.4 for per-phase analysis and can be justified using the external models derived here.

Suppose

$$y^l \;:=\; y^{la}\mathbb{I}, \qquad a \;:=\; a^a\mathbb{I}, \qquad V_j \;:=\; v_j\alpha_+, \qquad V_k \;:=\; v_k\alpha_+$$

where we recall that $\alpha_+ := (1, \alpha, \alpha^2)$ is the unit positive-sequence vector and $\alpha := e^{-\mathbf{i}2\pi/3}$. In this case Corollary 1.3 implies

$$\Gamma V_j \;=\; (1-\alpha)V_j, \qquad\qquad \Gamma^{\mathsf{T}} V_j \;=\; (1-\alpha^2)V_j$$

The external model (15.21a) of the $\Delta\Delta$-configured transformer then reduces to (with $y^m = 0$):

$$\begin{aligned} I_j \;&=\; \left(\Gamma^{\mathsf{T}} y^l \Gamma\right) V_j \;-\; \left(\Gamma^{\mathsf{T}} a y^l \Gamma\right) V_k \;=\; (1-\alpha)(1-\alpha^2) y^{la} \left(V_j - a^a V_k\right) \\ I_k \;&=\; -\left(\Gamma^{\mathsf{T}} a y^l \Gamma\right) V_j \;+\; \left(\Gamma^{\mathsf{T}} a^2 y^l \Gamma\right) V_k \;=\; (1-\alpha)(1-\alpha^2) y^{la} \left(-a^a V_j + (a^a)^2 V_k\right) \end{aligned}$$

Since $(1-\alpha)(1-\alpha^2) = 3$ we have

$$\begin{bmatrix} I_j \\ I_k \end{bmatrix} \;=\; \underbrace{\begin{bmatrix} \tilde{y} & -a\tilde{y} \\ -a\tilde{y} & a^2\tilde{y} \end{bmatrix}}_{\tilde{Y}_{YY}} \begin{bmatrix} V_j \\ V_k \end{bmatrix}$$

where $\tilde{y}^l = 3y^{la}\mathbb{I}$ and $a = a^a\mathbb{I}$. Hence when the system is balanced a $\Delta\Delta$-configured transformer has a YY equivalent with the same turns ratio a but a leakage admittance $\tilde{y}^l$ three times the original admittance y^l. Since the admittance matrix of the YY equivalent is

$$\tilde{Y}_{YY} \;:=\; \left(3y^{la} \begin{bmatrix} 1 & -a^a \\ -a^a & (a^a)^2 \end{bmatrix}\right) \otimes \mathbb{I}$$

we can interpret

$$\tilde{Y}_{1\phi} \;:=\; 3y^{la} \begin{bmatrix} 1 & -a^a \\ -a^a & (a^a)^2 \end{bmatrix}$$

as the admittance matrix of the *single-phase equivalent* of the $\Delta\Delta$ transformer in balanced setting.

In a balanced system a ΔY-configured transformer also has a YY equivalent when $V_k^n = 0$ and hence a single-phase equivalent, but the YY equivalent requires complex, rather than real, turns ratios. This is explained in the next example.

Example 15.7 (Single-phase equivalent of ΔY configuration with $V_k^n = 0$)**.** Consider a ΔY-configured transformer. Suppose, not only is the system balanced, i.e.,

$$y^l := y^{la}\mathbb{I}, \qquad a := a^a\mathbb{I}, \qquad V_j := v_j\alpha_+, \qquad V_k := v_k\alpha_+$$

but the neutral on the secondary side is also grounded with zero grounding impedance, i.e., $V_k^n = 0$. Show that its YY equivalent and single-phase equivalent are respectively

$$\tilde{Y}_{YY} := \tilde{Y}_{1\phi} \otimes \mathbb{I}, \qquad \tilde{Y}_{1\phi} := \tilde{y}^{la}\begin{bmatrix} 1 & -\tilde{a}^a \\ -\tilde{a}^{a\mathsf{H}} & |\tilde{a}^a|^2 \end{bmatrix}$$

where

$$\tilde{y}^{la} := 3y^{la}, \qquad \tilde{a}^a := \frac{a^a}{1-\alpha} = \frac{a^a}{\sqrt{3}e^{\mathbf{i}\pi/6}}$$

Solution. The external model of a ΔY-configured transformer is given by (15.24a). Applying Corollary 1.3 ($\Gamma V_j = (1-\alpha)V_j$, $\Gamma^{\mathsf{T}}V_j = (1-\alpha^2)V_j$), $(1-\alpha)(1-\alpha^2) = 3$ and $\Gamma^{\mathsf{T}}\mathbf{1} = 0$, we have[3]

$$I_j = \left(\Gamma^{\mathsf{T}}y^l\Gamma\right)V_j - \left(\Gamma^{\mathsf{T}}ay^l\right)\left(V_k - V_k^n\mathbf{1}\right) = 3y^{la}\left(V_j - \frac{a^a}{1-\alpha}V_k\right)$$

$$I_k = \left(-ay^l\Gamma\right)V_j + \left(a^2y^l\right)\left(V_k - V_k^n\mathbf{1}\right) = 3y^{la}\left(-\frac{a^a}{1-\alpha^2}V_j + \left(\frac{a^a}{\sqrt{3}}\right)^2\left(V_k - V_k^n\mathbf{1}\right)\right)$$

Since $a^a \in \mathbb{R}$ we have

$$\left(\frac{a^a}{1-\alpha^2}\right)^{\mathsf{H}} = \frac{a^a}{1-\alpha} = \frac{a^a}{\sqrt{3}e^{\mathbf{i}\pi/6}}$$

Define the matrices

$$\tilde{y}^l := 3y^{la}\mathbb{I}, \qquad \tilde{a} := \frac{a^a}{1-\alpha}\mathbb{I}, \qquad |\tilde{a}|^2 := \frac{(a^a)^2}{3}\mathbb{I} \tag{15.29a}$$

The external model of the ΔY-configured transformer is then

$$\begin{bmatrix} I_j \\ I_k \end{bmatrix} = \begin{bmatrix} \tilde{y}^l & -\tilde{a}\tilde{y}^l \\ -\tilde{a}^{\mathsf{H}}\tilde{y}^l & |\tilde{a}|^2\tilde{y}^l \end{bmatrix}\begin{bmatrix} V_j \\ V_k \end{bmatrix} - \begin{bmatrix} 0 \\ |\tilde{a}|^2\tilde{y}^l V_k^n\mathbf{1} \end{bmatrix} \tag{15.29b}$$

To derive its YY equivalent, consider a YY-configured transformer with a complex voltage gain (matrix) $\hat{n} := \mathrm{diag}\left(\hat{n}^a, \hat{n}^b, \hat{n}^c\right) \in \mathbb{C}^{3\times 3}$ and its turns ratio (matrix) $\hat{a} := \hat{n}^{-1}$. Instead of (15.18c) for real transformer gains, the transformer gains when $\hat{n}$ and $\hat{a}$ are complex are given by

$$\hat{V}_k^Y = \hat{n}\hat{V}_j^Y, \qquad \hat{I}_k^Y = \hat{a}^{\mathsf{H}}\hat{I}_j^Y \tag{15.30a}$$

[3] To illustrate the effect of V_k^n on YY equivalent we do not substitute $V_k^n = 0$ until the last step.

Let $\hat{y} \in \mathbb{C}^{3\times 3}$ denote its leakage admittance matrix. Then its external model can be shown to be (Exercise 15.5):

$$\begin{bmatrix} I_j \\ I_k \end{bmatrix} = \underbrace{\begin{bmatrix} \hat{y}^l & -\hat{a}\hat{y}^l \\ -\hat{a}^{\mathsf{H}}\hat{y}^l & |\hat{a}|^2\hat{y}^l \end{bmatrix}}_{\tilde{Y}_{YY}} \left(\begin{bmatrix} V_j \\ V_k \end{bmatrix} - \begin{bmatrix} V_j^n \mathbf{1} \\ V_k^n \mathbf{1} \end{bmatrix} \right) \tag{15.30b}$$

$$I_j^n = -\mathbf{1}^{\mathsf{T}} I_j, \qquad I_k^n = -\mathbf{1}^{\mathsf{T}} I_k \tag{15.30c}$$

where $|\hat{a}|^2$ is the matrix $|\hat{a}|^2 := \mathrm{diag}\left(1/|\hat{n}^a|^2, 1/|\hat{n}^b|^2, 1/|\hat{n}^c|^2\right)$. Note that the matrix $\tilde{Y}_{YY}$ is not complex symmetric and therefore does not have a three-phase Π circuit representation when $\hat{a}$ is complex.

Comparing (15.29b) and (15.30b) we see that, if $V_k^n = 0$, then the ΔY-configured transformer has a YY equivalent whose neutrals are grounded with zero grounding impedances on both sides and whose admittance matrix $\hat{y} = \tilde{y}$ and complex turns ratio matrix $\hat{a} = \tilde{a}$ are given by (15.29a). This completes the proof. □

15.3 Three-phase transformer models: unitary voltage network

In this section we extend the single-phase model in Chapter 3.1.5 with unitary voltage network to three-phase transformers. Multiple copies of the single-phase circuit in Figure 3.8(b) can be connected in Δ or Y configuration on each side of the unitary voltage network, per phase, to create three-phase transformers. The derivation of their external models follows a similar method as that in Chapter 15.2.2: (i) define internal variables for the unitary voltage network in each phase; (ii) derive the internal model that relate these internal variables; (iii) the transformer gains across the two ideal transformers define the conversion between the internal and terminal variables; and finally (iv) eliminate the internal variables to arrive at the external models.

15.3.1 Internal model: UVN per phase

The internal variables on the unitary voltage network in each phase $\phi \in \{a, b, c\}$ are defined in Figure 15.18. Note that the voltages $(\hat{V}_0^\phi, \hat{V}_j^\phi, \hat{V}_k^\phi)$ are defined to be the voltage drops, whether the unitary voltage network is grounded or not. These variables satisfy (3.10) for each phase ϕ:

$$\hat{I}_j^\phi = y_j^\phi(\hat{V}_j^\phi - \hat{V}_0^\phi), \quad \hat{I}_k^\phi = y_k^\phi(\hat{V}_k^\phi - \hat{V}_0^\phi), \quad \hat{I}_0^\phi + \hat{I}_j^\phi + \hat{I}_k^\phi = y_0^\phi \hat{V}_0^\phi, \quad \phi \in \{a, b, c\} \tag{15.31}$$

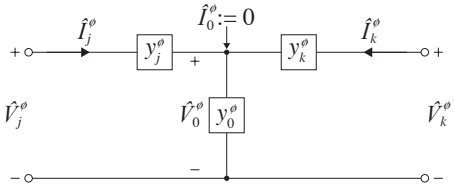


Figure 15.18 Unitary voltage network in each phase ϕ of a three-phase transformer.

Define the internal variables and admittance matrices:

$$\hat{I}_i := \begin{bmatrix} \hat{I}_i^a \\ \hat{I}_i^b \\ \hat{I}_i^c \end{bmatrix}, \qquad \hat{V}_i := \begin{bmatrix} \hat{V}_i^a \\ \hat{V}_i^b \\ \hat{V}_i^c \end{bmatrix}, \qquad y_i := \operatorname{diag}\left(y_i^a, y_i^b, y_i^c\right), \qquad i = 0, j, k$$

Then (15.31) is in vector form:

$$\hat{I}_j = y_j(\hat{V}_j - \hat{V}_0), \qquad \hat{I}_k = y_k(\hat{V}_k - \hat{V}_0), \qquad \hat{I}_0 + \hat{I}_j + \hat{I}_k = y_0\hat{V}_0$$

or in terms of a 9×9 admittance matrix:

$$\begin{bmatrix} \hat{I}_0 \\ \hat{I}_j \\ \hat{I}_k \end{bmatrix} = \begin{bmatrix} \sum_i y_i & -y_j & -y_k \\ -y_j & y_j & 0 \\ -y_k & 0 & y_k \end{bmatrix} \begin{bmatrix} \hat{V}_0 \\ \hat{V}_j \\ \hat{V}_k \end{bmatrix} \tag{15.32}$$

where $\sum_i y_i = y_0 + y_j + y_k$ is a diagonal matrix of all admittances. Since $\hat{I}_0 = 0 \in \mathbb{C}^3$ we can eliminate $\hat{V}_0$ and derive the 6×6 Kron-reduced admittance matrix Y_{uvn} that maps $\hat{V} := (\hat{V}_j, \hat{V}_k) \in \mathbb{C}^6$ to $\hat{I} := (\hat{I}_j, \hat{I}_k) \in \mathbb{C}^6$ (Exercise 15.6):

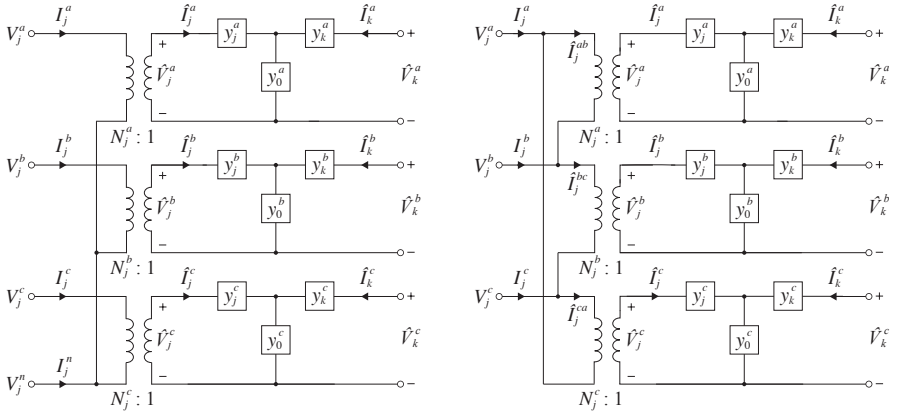
$$\hat{I} = Y_{\text{uvn}}\hat{V} \quad \text{where} \quad Y_{\text{uvn}} := \left(\mathbb{I}_2 \otimes \left(\sum_i y_i\right)^{-1}\right) \begin{bmatrix} y_j(y_0 + y_k) & -y_j y_k \\ -y_j y_k & y_k(y_0 + y_j) \end{bmatrix} \tag{15.33}$$

and $\mathbb{I}_2$ is the identity matrix of size 2. This defines the *internal model* that relates $\hat{I}$ and $\hat{V}$. Note that the phases of these internal variables are decoupled in (15.33) since the admittance matrices $y_i \in \mathbb{C}^{3\times 3}$ are diagonal. The phases will be coupled in the terminal variables (V_j, V_k) and (I_j, I_k) through Y or Δ configuration, as we now explain.

15.3.2 Conversion rules

Let the terminal currents of the three-phase transformer be $I_i := (I_i^a, I_i^b, I_i^c)$, its terminal voltages be $V_i := (V_i^a, V_i^b, V_i^c)$, and the terminal neutral voltage of Y configuration be V_i^n, $i = j, k$. The primary side is illustrated in Figure 15.19. These voltages are defined respect to an arbitrary and common reference point, e.g., the ground. Let $M_j := \operatorname{diag}\left(1/N_j^a, 1/N_j^b, 1/N_j^c\right)$ and $M_k := \operatorname{diag}\left(1/N_k^a, 1/N_k^b, 1/N_k^c\right)$ be the transformer gain matrices of the ideal transformers on each side of the unitary voltage network.

To derive the conversion between internal and terminal variables, consider first the primary side where three single-phase ideal transformers are connected to the left end of the unitary voltage network in Figure 15.18. Figure 15.19(a) shows the primary side in Y configuration. The *conversion rule* between the internal variables $(\hat{V}_j, \hat{I}_j)$ and the



(a) Y configuration (b) Δ configuration

Figure 15.19 Primary side of a three-phase transformer with unitary voltage networks.

terminal variables (V_j, I_j, V_j^n) is:

$$Y \text{ configuration:} \qquad \hat{V}_j \;=\; M_j \left(V_j - V_j^n \mathbf{1} \right), \qquad\qquad \hat{I}_j \;=\; M_j^{-1} I_j \tag{15.34a}$$

where $\mathbf{1} := (1,1,1)$. Figure 15.19(b) shows the primary side in Δ configuration. Let $\hat{I}_j^{\Delta} := (\hat{I}_j^{ab}, \hat{I}_j^{bc}, \hat{I}_j^{ca})$ denote the internal currents entering the primary side of the ideal transformer as indicated in Figure 15.19(b). From (14.9a) the internal variables $(\hat{V}_j, \hat{I}_j, \hat{I}_j^{\Delta})$ are related to the terminal variables (V_j, I_j) according to the *conversion rule*:

$$\Delta \text{ configuration:} \qquad \hat{V}_j \;=\; M_j \Gamma V_j, \qquad \hat{I}_j \;=\; M_j^{-1} \hat{I}_j^{\Delta}, \qquad I_j \;=\; \Gamma^{\mathsf{T}} \hat{I}_j^{\Delta} \tag{15.34b}$$

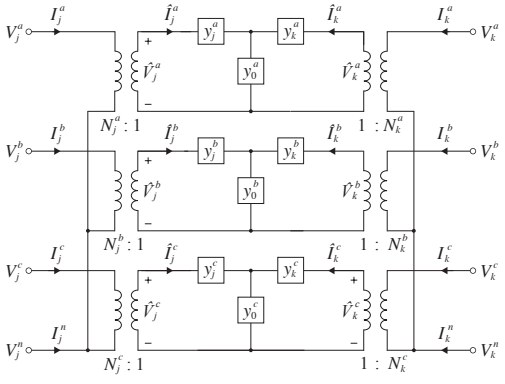
where $\Gamma, \Gamma^{\mathsf{T}}$ are conversion matrices. Similarly on the secondary side we have the conversion rule (see Figure 15.20):

$$Y \text{ configuration:} \quad \hat{V}_k \;=\; M_k \left(V_k - V_k^n \mathbf{1} \right), \quad \hat{I}_k \;=\; M_k^{-1} I_k \tag{15.34c}$$

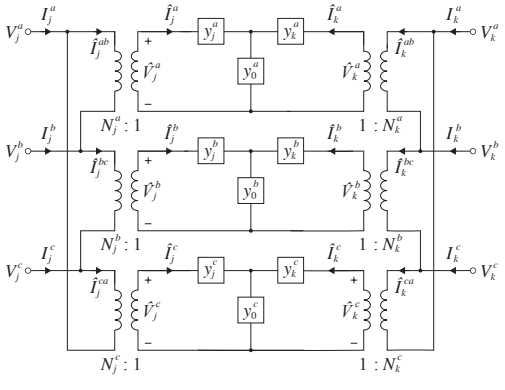
$$\Delta \text{ configuration:} \quad \hat{V}_k \;=\; M_k \Gamma V_k, \qquad\quad \hat{I}_k \;=\; M_k^{-1} \hat{I}_k^{\Delta}, \quad I_k \;=\; \Gamma^{\mathsf{T}} \hat{I}_k^{\Delta} \tag{15.34d}$$

15.3.3 External model

We can derive an external model by eliminating the internal variables $(\hat{V}, \hat{I}, \hat{I}^{\Delta})$ from the internal model (15.33) and the conversion rules (15.34). Specifically substitute



(a) YY configuration



(b) $\Delta\Delta$ configuration

Figure 15.20 Three-phase transformer models with unitary voltage networks.

(15.34) into (15.33) to get

$$YY: \quad \begin{bmatrix} M_j^{-1} I_j \\ M_k^{-1} I_k \end{bmatrix} = Y_{\text{uvn}} \begin{bmatrix} M_j (V_j - V_j^n \mathbf{1}) \\ M_k (V_k - V_k^n \mathbf{1}) \end{bmatrix}, \qquad \Delta\Delta: \quad \begin{bmatrix} M_j^{-1} \hat{I}_j^{\Delta} \\ M_k^{-1} \hat{I}_k^{\Delta} \end{bmatrix} = Y_{\text{uvn}} \begin{bmatrix} M_j \Gamma V_j \\ M_k \Gamma V_k \end{bmatrix} \tag{15.35a}$$

$$\Delta Y: \quad \begin{bmatrix} M_j^{-1} \hat{I}_j^{\Delta} \\ M_k^{-1} I_k \end{bmatrix} = Y_{\text{uvn}} \begin{bmatrix} M_j \Gamma V_j \\ M_k (V_k - V_k^n \mathbf{1}) \end{bmatrix}, \qquad Y\Delta: \quad \begin{bmatrix} M_j^{-1} I_j \\ M_k^{-1} \hat{I}_k^{\Delta} \end{bmatrix} = Y_{\text{uvn}} \begin{bmatrix} M_j (V_j - V_j^n \mathbf{1}) \\ M_k \Gamma V_k \end{bmatrix} \tag{15.35b}$$

Let $V := (V_j, V_k) \in \mathbb{C}^6$ and $I := (I_j, I_k) \in \mathbb{C}^6$ denote the vectors of terminal voltages and currents respectively. Let $M := \text{diag}(M_j, M_k) \in \mathbb{R}^{6\times 6}$ be the transformer gain matrices. Then the *external model* of a three-phase transformer is (Exercise 15.7)

$$I = D^{\mathsf{T}} (M Y_{\text{uvn}} M) D (V - \gamma) \tag{15.36a}$$

where Y_{uvn} is defined in (15.33), $D \in \mathbb{C}^{6\times 6}$ and $\gamma \in \mathbb{C}^6$ are defined in (15.16).

We often do not know the numbers N_j^ϕ, N_k^ϕ of turns of the primary and secondary windings respectively and hence cannot determine the matrices M_j, M_k, but we can always determine the turns ratio matrix $a := M_j^{-1} M_k = \operatorname{diag}\left(N_j^a/N_k^a, N_j^b/N_k^b, N_j^c/N_k^c\right)$ from the specified rated voltages. The 3×3 admittance matrices y_0, y_1, y_2 are assembled from their per-phase admittances and recall from (3.9) (see Figure 3.8):

$$\begin{aligned} y_0 &:= N_j^2\, y^m := N_j^2 \operatorname{diag}\left(y^{ma}, y^{mb}, y^{mc}\right) \\ y_j &:= N_j^2\, y^p := N_j^2 \operatorname{diag}\left(y^{pa}, y^{pb}, y^{pc}\right), \qquad y^{p\phi} := \frac{1}{z^{p\phi}}, \quad \phi \in \{a,b,c\} \\ y_k &:= N_k^2\, y^s := N_j^2 \operatorname{diag}\left(y^{sa}, y^{sb}, y^{sc}\right), \qquad y^{s\phi} := \frac{1}{z^{s\phi}}, \quad \phi \in \{a,b,c\} \end{aligned}$$

Then the matrix $MY_{\text{uvn}}M$ in (15.36a) can also be written in terms of the 3×3 turns ratio and admittance matrices a, y^p, y^s, y^m (Exercise 15.8):

$$Y_{YY} := MY_{\text{uvn}}M = y^p y^s \left(a^2 y^m + a^2 y^p + y^s\right)^{-1} \begin{bmatrix} \mathbb{I} + a^2 y^m (y^s)^{-1} & -a \\ -a & a^2\left(\mathbb{I} + y^m (y^p)^{-1}\right) \end{bmatrix} \tag{15.36b}$$

Hence the external model of a standard three-phase transformer is

$$I = D^{\mathsf{T}} Y_{YY} D\,(V - \gamma) \tag{15.36c}$$

where Y_{YY} is defined in (15.36b), $D \in \mathbb{C}^{6\times 6}$ and $\gamma \in \mathbb{C}^6$ are defined in (15.16), reproduced here: $\gamma := \left(V_j^n \mathbf{1}, V_k^n \mathbf{1}\right)$ are neutral voltages for Y configuration and D is a 6×6 block diagonal matrix that depends on configuration:

$$\begin{aligned} &YY \text{ configuration:} && D := \begin{bmatrix} \mathbb{I} & 0 \\ 0 & \mathbb{I} \end{bmatrix} \\ &\Delta\Delta \text{ configuration:} && D := \begin{bmatrix} \Gamma & 0 \\ 0 & \Gamma \end{bmatrix} \\ &\Delta Y \text{ configuration:} && D := \begin{bmatrix} \Gamma & 0 \\ 0 & \mathbb{I} \end{bmatrix} \\ &Y\Delta \text{ configuration:} && D := \begin{bmatrix} \mathbb{I} & 0 \\ 0 & \Gamma \end{bmatrix} \end{aligned}$$

For $\Delta\Delta$ configuration, $D\gamma = 0 \in \mathbb{C}^6$ in (15.36), reflecting that a Δ configuration contains no neutral voltage; similarly for other configurations.

Remark 15.3. 1 As explained in Chapter 3.1.5, the transformer model with unitary voltage networks is equivalent to the T equivalent circuit. This holds in both single-phase and three-phase settings.

2 This model is generally different from the simplified model of Chapter 15.2 which is the three-phase extension of the model in Chapter 3.1.4. From (15.36) and (15.16), these models however have the same structure. They differ only in the admittance matrix Y_{YY} for the YY configuration and the difference is due to different models for single-phase nonideal transformers.

3 When the shunt admittances are assumed zero in both models, i.e., $y_0^{\phi} = y^{m\phi} = 0$ for $\phi \in \{a,b,c\}$, these two models are equivalent, as in the single-phase case. To see this, recall that per-phase $\phi \in \{a,b,c\}$, the leakage impedances in the simplified model are $z^{l\phi} = z^{p\phi} + (a^{\phi})^2 z^{s\phi}$ and hence the leakage admittances per phase are

$$y^{l\phi} = \left(z^{la}\right)^{-1} = \left(1/y^{p\phi} + (a^{\phi})^2\, y^{s\phi}\right)^{-1} = \frac{y^{p\phi} y^{s\phi}}{(a^{\phi})^2 y^{p\phi} + y^{s\phi}}, \qquad \phi \in \{a,b,c\}$$

Since all matrices are diagonal we have $y^l = y^p y^s \left(a^2 y^p + y^s\right)^{-1}$. Substituting this and $y^m = 0$ into (15.36b), Y_{YY} for the transformer model based on the unitary voltage network reduces to

$$Y_{YY} = y^l \begin{bmatrix} \mathbb{I} & -a \\ -a & a^2 \end{bmatrix}$$

which is the same as Y_{YY} in (15.16a) for the simplified model. (See Exercise 15.9 for another proof).

4 The model (15.36) generalizes the single-phase model (3.11) in three ways. First the 6×6 admittance matrix $MY_{\text{uvn}}M$ in (15.36) has the same structure as the 2×2 matrix in (3.11). Second the neutrals of the three-phase transformer in Y configuration may not be grounded, i.e., V_j^n, V_k^n may be nonzero whereas V in (3.11) is assumed to be the voltage drop across the windings. Finally the admittance matrix of a three-phase transformer in YY configuration is $Y_{YY} := MY_{\text{unv}}M$, and a Δ configuration in either the primary or the secondary circuit is represented by conversion matrices Γ^{T} and Γ.

15.3.4 Split-phase transformer

15.4 Parameter identification: examples

15.4.1 Simplified circuit

Example 15.8 (Parameter identification)**.** Consider a three-phase transformer in ΔY configuration. Its simplified circuit model is shown in Figure 15.21. Suppose the single-phase transformers are identical, i.e. their turns ratios $a := a^a\mathbb{I}$ and leakage admittances $y^l := y^{la}\mathbb{I}$ are the same across phases. Suppose the shunt admittances are zero. We discuss parameter identification in two steps.

1 Suppose the following measurements are given:
- Terminal currents $I_j = i_j \in \mathbb{C}^3$ and $I_k = i_k \in \mathbb{C}^3$.
- Terminal voltages $V_j = v_j \in \mathbb{C}^3$ (with respect to ground) on the primary (Δ) side.
- Line-to-line voltages $\Gamma V_k = u_k \in \mathbb{C}^3$ on the secondary (Y) side.
- The neutral is grounded with zero grounding impedance so that $V_k^n := 0$.

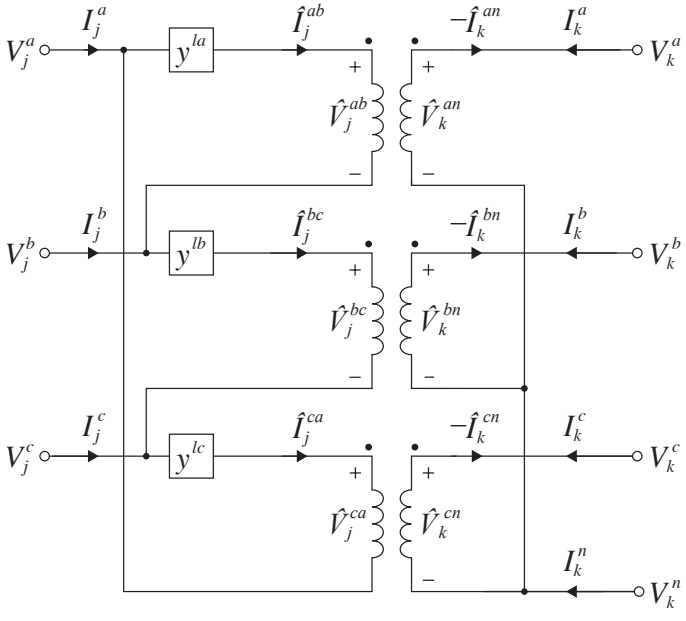


Figure 15.21 ΔY-configured transformer with zero shunt admittances.

Assume the measurements are error free and let $x := (i_j, i_k, v_j, u_k)$ be the measurement vector. Calculate:

- The turns ratio a^a and the leakage admittance y^{la}.
- The terminal voltage V_k with respect to the ground.
- The internal voltage and current $(\hat{V}_k^Y, \hat{I}_k^Y)$ on the secondary side.
- The internal voltage and current $(\hat{V}_j^\Delta, \hat{I}_j^\Delta)$ on the primary side and hence the loop flow β_j within the Δ configuration.

2 Repeat part 1 when T measurements $(x_1, \ldots, x_T)$ are given and measurement errors may be nonzero.

Solution. Under the assumption of zero measurement error, the measurement $x := (i_j, i_k, v_j, u_k) \in \mathbb{C}^{12}$, the parameter $\theta := (a^a, y^{la}) \in \mathbb{C}^2$, and the variable $V_k \in \mathbb{C}$ satisfy (15.24a) with $y^l := y^{la}\mathbb{I}, a := a^a\mathbb{I}$:

$$\begin{bmatrix} i_j \\ i_k \end{bmatrix} = y^{la} \begin{bmatrix} \Gamma^\mathsf{T}\Gamma & -a^a\Gamma^\mathsf{T} \\ -a^a\Gamma & (a^a)^2\mathbb{I} \end{bmatrix} \begin{bmatrix} v_j \\ V_k \end{bmatrix} \tag{15.37}$$

We can obtain $\Gamma^\mathsf{T} V_k$ from the line-to-line voltage measurement $\Gamma V_k = u_k$ by shifting

the values of u_k:

$$\Gamma^{\mathsf{T}} V_k \;=\; \begin{bmatrix} V^a - V^c \\ V^b - V^a \\ V^c - V^b \end{bmatrix} \;=\; -\underbrace{\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}}_{\text{permutation } P} \begin{bmatrix} V^a - V^b \\ V^b - V^c \\ V^c - V^a \end{bmatrix} \;=\; -P\Gamma V_k \;=\; -Pu_k$$

Hence the first row of (15.37) becomes

$$i_j \;=\; y^{la}\left(\Gamma^{\mathsf{T}}\Gamma v_j + a^a P u_k\right) \tag{15.38a}$$

where P is the permutation matrix

$$P \;:=\; \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \tag{15.38b}$$

This is a set of 3 quadratic equations in a positive real variable $a^a \in \mathbb{R}_+$ and a complex variables $y^{la} \in \mathbb{C}$. Under appropriate conditions a solution of (15.38) exists and can be computed numerically. Let $\theta := (a^a, y^{la})$ denote such a solution. All other variables can then be derived in terms of the parameter θ and the measurement $x := (i_j, i_k, v_j, u_k)$, as follows.

The terminal voltage V_k can be calculated from the second row of (15.37):

$$V_k \;=\; \frac{1}{(a^a)^2 y^{la}}\, i_k + \frac{1}{a^a}\Gamma v_j \tag{15.39a}$$

On the secondary side the internal voltage and current $(\hat{V}_k^Y, \hat{I}_k^Y)$ are given by the conversion rule in (15.23b) for Y configuration on the secondary side:

$$\hat{V}_k^Y \;=\; V_k - V_k^n \mathbf{1} \;=\; V_k, \qquad\qquad \hat{I}_k^Y \;=\; i_k \tag{15.39b}$$

On the primary side the internal voltage $\hat{V}_j^\Delta$ *across the ideal transformers* is given by (15.13d) with $z^m := 0$ (no shunt admittance):

$$\hat{V}_j^\Delta \;=\; \Gamma v_j - \frac{1}{y^{la}}\hat{I}_j^\Delta$$

Instead of expressing $\hat{I}_j^\Delta$ in terms of the measurement i_j using $y^m = 0$ and the conversion rule $i_j = \Gamma^{\mathsf{T}}\hat{I}_j^\Delta$, we will use the transformer current gain in (15.23c) for ΔY configuration to express $\hat{I}_j^\Delta$ in terms of the measurement i_k, yielding

$$\hat{V}_j^\Delta \;=\; \Gamma v_j + \frac{1}{a^a\, y^{la}}\, i_k, \quad \hat{I}_j^\Delta \;=\; -\frac{1}{a^a}\hat{I}_k^Y \;=\; -\frac{1}{a^a}\, i_k, \quad \beta_j \;:=\; \frac{1}{3}\mathbf{1}^{\mathsf{T}}\hat{I}_j^\Delta \;=\; -\frac{1}{3a^a}\mathbf{1}^{\mathsf{T}} i_k \tag{15.39c}$$

Even though we cannot determine the loop flow β_j from the terminal current i_j, we can from the measurement i_k on the secondary side.

When the measurement error is zero, the measurement vector $x := (i_j, i_k, v_j, u_k)$ and the parameter vector $\theta := (a^a, y^{la})$ satisfy (15.38). This can be represented as

$f(x;\theta) = 0$ for some function f. Given T measurements $x := (x_1, \ldots, x_T)$, there may not be any choice of θ such that $f(x_t;\theta) = 0$ for all $t = 1, \ldots, T$ when measurement errors are nonzero. A popular estimate of θ is one that minimizes error subject to certain constraints:

$$\hat{\theta} \;:=\; \arg\min_{\theta} \sum_{t} \|f(x_t;\theta)\| \qquad \text{s.t.} \qquad g(x_t;\theta) \;\leq\; 0, \quad t = 1, \ldots, T$$

for some appropriate norm $\|\cdot\|$. Here $g(x_t;\theta) \leq 0$ expresses some known relations that must hold, e.g., $a^a \geq 0$ is real. Let $\hat{\theta}$ denote an estimate of the parameter. Then other variables

$$\hat{y}_t \;:=\; (V_k(t), \hat{V}_j^{\Delta}(t), \hat{I}_j^{\Delta}(t), \hat{V}_k^{Y}(t), \hat{I}_k^{Y}(t)), \qquad t = 1, \ldots, T$$

can be derived from (15.39) in terms of $\hat{\theta}$ and the measurements x_t.

It is possible that the estimate $\hat{y}_t$ derived in this way may violate some known constraints, e.g., $v_k^{\min} \leq \|V_k(t)\|^2 \leq v_k^{\max}$ for some t given voltage limits. An alternative identification method is to estimate the parameter θ and the variables $y := (y_1, \ldots, y_T)$ jointly from the measurements $x := (x_1, \ldots, x_T)$, i.e., solve

$$(\hat{\theta}, \hat{y}) \;:=\; \arg\min_{(\theta, y)} \sum_{t} \|f(x_t, y_t;\theta)\| \qquad \text{s.t.} \qquad g(x_t, y_t;\theta) \;\leq\; 0, \quad t = 1, \ldots, T$$

where f represents (15.38)(15.39) and $g(x_t, y_t;\theta) \leq 0$ express some known constraints on $(\hat{\theta}, \hat{y})$. □

From Figure 15.21 the terminal powers s_j and s_k are powers injected into the transformer at terminals j and k respectively. Hence $\mathbf{1}^{\mathsf{T}}(s_j + s_k)$ is the total power loss in the three-phase transformer due to the leakage impedance $1/y^l$, as the next example shows.

Example 15.9 (Total power loss)**.** For the three-phase transformer in Example 15.8 show that the total power loss $\mathbf{1}^{\mathsf{T}}(s_j + s_k)$ in the transformer is equal to (assuming zero measurement error):

$$\mathbf{1}^{\mathsf{T}}(s_j + s_k) \;=\; \frac{1}{y^{la}} \|n^a i_k\|_2^2$$

where $n^a := 1/a^a$ is the voltage gain. Even though the transformer gain n^a relates the internal currents $(\hat{I}_j^{\Delta}, \hat{I}_k^{Y})$, not terminal currents (I_j, I_k), we can interpret $n^a i_k$ as the "effective" terminal current on the primary side.

Solution. The terminal powers are, from (15.39),

$$\begin{aligned} s_j \;&:=\; \operatorname{diag}\left(V_j I_j^{\mathsf{H}}\right) \;=\; -n^a \operatorname{diag}\left(v_j i_k^{\mathsf{H}} \Gamma\right) \\ s_k \;&:=\; \operatorname{diag}\left(V_k I_k^{\mathsf{H}}\right) \;=\; n^a \operatorname{diag}\left(\Gamma v_j i_k^{\mathsf{H}}\right) + \frac{(n^a)^2}{y^{la}} \operatorname{diag}\left(i_k i_k^{\mathsf{H}}\right) \end{aligned}$$

where $n^a := 1/a^a$, the second equality follows from $y^m = 0$ and hence $i_j = \Gamma^{\mathsf{T}} \hat{I}_j^{\Delta} =$

$-n^a \Gamma^{\mathsf{T}} i_k$, and the last equality follows from (15.39a). Hence

$$s_j + s_k = n^a \left(\operatorname{diag}\left(\Gamma v_j i_k^{\mathsf{H}}\right) - \operatorname{diag}\left(v_j i_k^{\mathsf{H}} \Gamma\right)\right) + \frac{(n^a)^2}{y^{la}} \operatorname{diag}\left(i_k i_k^{\mathsf{H}}\right)$$

Now

$$\operatorname{diag}\left(\Gamma v_j i_k^{\mathsf{H}}\right) - \operatorname{diag}\left(v_j i_k^{\mathsf{H}} \Gamma\right) = \begin{bmatrix} (v_j^a - v_j^b)\bar{i}_k^a \\ (v_j^b - v_j^c)\bar{i}_k^b \\ (v_j^c - v_j^a)\bar{i}_k^c \end{bmatrix} - \begin{bmatrix} v_j^a(\bar{i}_k^a - \bar{i}_k^c) \\ v_j^b(\bar{i}_k^b - \bar{i}_k^a) \\ v_j^c(\bar{i}_k^c - \bar{i}_k^b) \end{bmatrix} = \begin{bmatrix} v_j^a \bar{i}_k^c \\ v_j^b \bar{i}_k^a \\ v_j^c \bar{i}_k^b \end{bmatrix} - \begin{bmatrix} v_j^b \bar{i}_k^a \\ v_j^c \bar{i}_k^b \\ v_j^a \bar{i}_k^c \end{bmatrix}$$

$$\operatorname{diag}\left(i_k i_k^{\mathsf{H}}\right) = \begin{bmatrix} |i_k^a|^2 \\ |i_k^b|^2 \\ |i_k^c|^2 \end{bmatrix}$$

where P is the permutation matrix in (15.38b). The total power loss in the three-phase transformer is then

$$\mathbf{1}^{\mathsf{T}}(s_j + s_k) = n^a \left((P i_k)^{\mathsf{H}} v_j - i_k^{\mathsf{H}} (P^{\mathsf{T}} v_j) \right) + \frac{(n^a)^2}{y^{la}} \|i_k\|_2^2 = \frac{1}{y^{la}} \|n^a i_k\|_2^2$$

where the last equality follows from $(P i_k)^{\mathsf{H}} v_j = i_k^{\mathsf{H}} (P^{\mathsf{T}} v_j)$. □

15.4.2 Unitary voltage network

15.5 Bibliographical notes

The modeling of transmission lines with earth return is presented in the seminal paper [157] by J. R. Carson. Circuit models of three-phase line models studied in Chapter 15.1 are developed in e.g. [158, 159, 43]. See e.g. [156, Chapter 3] for comprehensive models of three-phase components including distribution lines, transformers and switches. For the simplified model of Chapter 15.2 see [160, 161, 162, 163] for early work and [43, Ch 8][5, Ch 7.4][164] for recent summary. The idea of decomposing a nonideal transformer into two ideal transformers connected by a unitary voltage network as in Chapter 15.3 is first mentioned, but not explored, in [160]. It is developed in detail in [165] where the unitary network is a Π circuit with a leakage (series) admittance and two shunt admittances. The unitary voltage network in [166] uses a T circuit model, as Chapter 15.3 does. The unitary voltage network that models leakage fluxes and core losses can be quite general e.g. [167, 168].

15.6 Problems

Chapter 15.1.

Exercise 15.1 (Symmetric y_{jk})**.** Let z_{jk} be a phase impedance matrix of a three-phase line (j,k). Assume z_{jk} is symmetric invertible and $z_{jk} = z_{kj}$ (A0). Show that its inverse $y_{jk} := z_{jk}^{-1}$ is symmetric. Moreover $y_{jk} = y_{kj}$.

Chapter 15.2.

Exercise 15.2 (ΔY-configured transformer)**.** Derive the external model (15.24) of the ΔY-configured three-phase transformer in Figure 15.14.

Exercise 15.3 ($Y\Delta$-configured transformer)**.** Derive the external model (15.25) of the $Y\Delta$-configured three-phase transformer in Figure 15.15.

Exercise 15.4 (Open transformers)**.**

Exercise 15.5 (Complex voltage gain)**.** Consider a YY-configured transformer with a *complex* voltage gain (matrix) $n := \operatorname{diag}\left(n^a, n^b, n^c\right) \in \mathbb{C}^{3\times 3}$. Let its turns ratio be $a := n^{-1} \in \mathbb{C}^{3\times 3}$. Let $y^l \in \mathbb{C}^{3\times 3}$ denote its series admittance and assume its shunt admittance $y^m = 0$. Show that its external model is

$$\begin{bmatrix} I_j \\ I_k \end{bmatrix} = \underbrace{\begin{bmatrix} y^l & -ay^l \\ -a^{\mathsf{H}} y^l & |a|^2 y^l \end{bmatrix}}_{Y_{YY}} \left(\begin{bmatrix} V_j \\ V_k \end{bmatrix} - \begin{bmatrix} V_j^n \mathbf{1} \\ V_k^n \mathbf{1} \end{bmatrix} \right)$$

$$I_j^n = -\mathbf{1}^{\mathsf{T}} I_j, \qquad I_k^n = -\mathbf{1}^{\mathsf{T}} I_k$$

where $|a|^2$ is the matrix $|a|^2 := \operatorname{diag}\left(1/|n^a|^2, 1/|n^b|^2, 1/|n^c|^2\right)$.

Exercise 15.6 (Unitary voltage network: 3ϕ transformers)**.** Derive (15.33), reproduced here:

$$\hat{I} = Y_{\text{uvn}} \hat{V}$$

where

$$Y_{\text{uvn}} := \left(\mathbb{I}_2 \otimes \left(\sum_i y_i \right)^{-1} \right) \begin{bmatrix} y_j (y_0 + y_k) & -y_j y_k \\ -y_j y_k & y_k (y_0 + y_j) \end{bmatrix}$$

$\mathbb{I}_2$ is the identity matrix of size 2, and $\sum_i y_i = y_0 + y_j + y_k$ is a diagonal matrix of all admittances.

Exercise 15.7 (Unitary voltage network: 3ϕ transformers)**.** Show that, for the transformer model in Chapter 15.3 with unitary voltage network, the admittance matrices of standard three-phase transformers are given by

$$I \;=\; D^{\mathsf{T}}(MY_{\text{uvn}}M)D\,(V-\gamma)$$

where Y_{uvn} is defined in (15.33), and $D \in \mathbb{C}^{6\times 6}$ and $\gamma \in \mathbb{C}^6$ are defined in (15.16).

Exercise 15.8 (Unitary voltage network: turns ratio a)**.** Prove (15.36b): the matrix $MY_{\text{uvn}}M$ in (15.36) can be written in terms of the 3×3 turns ratio and admittance matrices a, y^p, y^s, y^m:

$$Y_{YY} \;:=\; MY_{\text{uvn}}M \;=\; y^p y^s \left(a^2 y^m + a^2 y^p + y^s\right)^{-1} \begin{bmatrix} \mathbb{I} + a^2 y^m (y^s)^{-1} & -a \\ -a & a^2\left(\mathbb{I} + y^m (y^p)^{-1}\right) \end{bmatrix}$$

Exercise 15.9 (3ϕ transformer: $y^m = y_0 = 0$)**.** Suppose shunt admittances $y_0 = y^m = \text{diag}(0,0,0)$. Then the admittance matrices Y_{uvn} defined in (15.33) and Y_{YY} defined in (15.16a) become

$$Y_{\text{uvn}} \;:=\; \left(\mathbb{I}_2 \otimes \left(y_j + y_k\right)^{-1}\right) \begin{bmatrix} y_j y_k & -y_j y_k \\ -y_j y_k & y_j y_k \end{bmatrix}, \qquad Y_{YY} \;:=\; \begin{bmatrix} y^l & -a y^l \\ -a y^l & a^2 y^l \end{bmatrix}$$

Show that $MY_{\text{uvn}}M = Y_{YY}$.

Exercise 15.10 (Split-phase transformer)**.** Consider a split-phase $\Delta\Delta$ transformer in Figure **??**. Suppose $\sum_{\phi\in\{a,b,c\}} \left(I_k^{\phi} + I_k^{\phi'}\right) = 0$. Derive (**??**).

16 Bus injection models

In this chapter we use the component models in Chapters 14 and 15 to construct network models and study unbalanced three-phase analysis. In Chapter 16.1 we extend the relation between terminal voltage, current and power (V, I, s) in the single-phase bus injection model of Chapter 4.3 to the unbalanced three-phase setting. In Chapter 16.2 we formulate a general three-phase analysis problem. In Chapter 16.3 we study the analysis problem when the network is balanced. We prove formally that a general balanced network is equivalent to per-phase networks and its analysis can be solved by per-phase analysis. In Chapter 16.4 we explain that, when an unbalanced system has a certain symmetry, we can transform it to a sequence coordinate in which the system becomes decoupled even if the phases are coupled in the original coordinate. Single-phase analysis can then be applied to individual sequence networks.

16.1 Network models

In this section we develop a model for a network of three-phase devices connected by three-phase lines and transformers studied in Chapters 14 and 15. We start in Chapter 16.1.1 with a line model that models a three-phase transmission or distribution line or a three-phase transformer. The line model linearly relates the sending-end line currents $\left(I_{jk}, I_{kj}\right) \in \mathbb{C}^6$ and the nodal voltages $\left(V_j, V_k\right) \in \mathbb{C}^6$ by an admittance matrix Y_{jk} which may or may not have a three-phase Π circuit representation. The line model induces a network model through nodal current balance equations. This is derived in Chapter 16.1.2 and it linearly relates the nodal (terminal) current injections I_j and voltages V_j through a network admittance matrix Y. The admittance matrix Y also implies a single-phase equivalent circuit of the three-phase network. We then use Y to derive in Chapter 16.1.4 nonlinear power flow equations that relate nodal (terminal) power injections s_j and voltages V_j. Finally we explain in Chapter 16.1.5 that the overall model consists of the network equations of Chapters 16.1.2 and 16.1.4 and the three-phase device models of Chapter 14.3. A device model can either be specified as an internal model with conversion rules or an external model relating the terminal variables (V_j, I_j, s_j).

16.1.1 Line model

Consider a network with $N+1$ three-phase devices connected by three-phase lines represented as an undirected graph $G := (\overline{N}, E)$ where every bus $j \in \overline{N}$ and every line $(j,k) \in E$ has 3 phases. A bus is where the terminals of three-phase devices are connected. A line may model a transmission or distribution line, a transformer, or a combination. We will hence refer to $j \in \overline{N}$ interchangeably as a bus, a node, or a terminal, and $(j,k) \in E$ interchangeably as a line, a branch, a link, or an edge. The formulation can be generalized to the case where a bus or a line has a single, two, or three phases.

For simplicity of exposition we assume, by default, we can use three-wire models for these lines and their characterization includes the effects of neutral and earth return on the phase variables. This assumption is reasonable if, e.g., neutral wires are absent, the line connects devices in Δ configuration, or the neutrals are directly grounded with equal spacing along a line and at both ends of the line so that all neutrals have $V_j^n = 0$. Otherwise, the line model in this section needs to be augmented with neutral lines with variables in $\mathbb{C}^4$ instead of $\mathbb{C}^3$ and line admittance matrices in $\mathbb{C}^{4\times4}$ instead of $\mathbb{C}^{3\times3}$; see Example 16.5 and Exercise 16.7. As we will see, even though lines are assumed to be three-wired, Y-configured devices such as voltage, current and power sources and impedances do have neutral lines in our models and their neutral voltages $\gamma_j := V_j^n$ may be nonzero.

For each line $(j,k) \in E$ let $\left(V_j, V_k\right) \in \mathbb{C}^6$ denote the terminal voltages at each end of the line and $\left(I_{jk}, I_{kj}\right) \in \mathbb{C}^6$ denote the sending-end line currents in both directions. In general each line $(j,k) \in E$ is characterized by four 3×3 series and shunt admittance matrices, $\left(y_{jk}^s, y_{jk}^m\right)$ from j to k and $\left(y_{kj}^s, y_{kj}^m\right)$ from k to j. See Figure 16.1. They

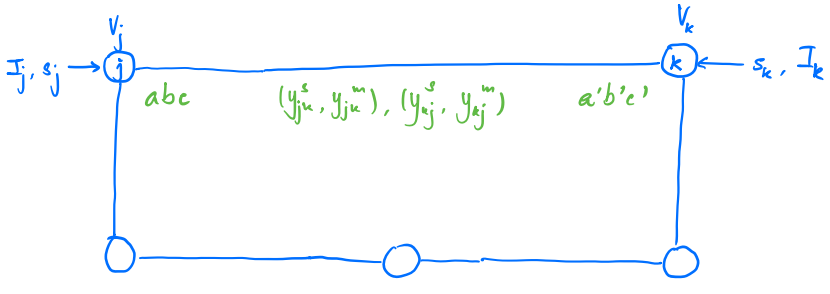


Figure 16.1 A model of three-phase system. Correction: Remove abc and $a'b'c'$.

define the relation between $\left(V_j, V_k\right)$ and $\left(I_{jk}, I_{kj}\right)$:

$$I_{jk} \;=\; y_{jk}^s(V_j - V_k) \;+\; y_{jk}^m V_j, \qquad I_{kj} \;=\; y_{kj}^s(V_k - V_j) \;+\; y_{kj}^m V_k \tag{16.1a}$$

or in matrix form:

$$\begin{bmatrix} I_{jk} \\ I_{kj} \end{bmatrix} = \underbrace{\begin{bmatrix} y^s_{jk} + y^m_{jk} & -y^s_{jk} \\ -y^s_{kj} & y^s_{kj} + y^m_{kj} \end{bmatrix}}_{Y_{jk}} \begin{bmatrix} V_j \\ V_k \end{bmatrix} \tag{16.1b}$$

We emphasize that y^s_{jk} and y^s_{kj} may be different matrices and therefore this general model Y_{jk} may not have a three-phase Π circuit representation. When $y^s_{jk} = y^s_{kj}$, it can model:

- A transmission or distribution line where, from (15.8a), $y^s_{jk} = y^s_{kj}$ is its series admittance and $\left(y^m_{jk}, y^m_{kj}\right)$ are its shunt admittances.
- A transformer in YY configuration where neutral voltages are zero ($V^n_j = V^n_k = 0$), from (15.19d),

  $$y^s_{jk} = y^s_{kj} := a\hat{y}^l, \quad y^m_{jk} := (\mathbb{I} - a)\,\hat{y}^l, \quad y^m_{kj} := a(a - \mathbb{I})\hat{y}^l + a^2\hat{y}^m \tag{16.2a}$$

  with $a := \mathrm{diag}(a^a, a^b, a^c)$ being the turns ratios of the transformer, $\hat{y}^l := \mathrm{diag}(y^{la}, y^{lb}, y^{lc})$ and $\hat{y}^m := \mathrm{diag}(y^{ma}, y^{mb}, y^{mc})$ its leakage and shunt admittances respectively.
- A transformer in $\Delta\Delta$ configuration where, from (15.21c),

  $$y^s_{jk} = y^s_{kj} := \Gamma^\mathsf{T} a \hat{y}^l \Gamma, \quad y^m_{jk} := \Gamma^\mathsf{T}(\mathbb{I} - a)\,\hat{y}^l \Gamma, \quad y^m_{kj} := \Gamma^\mathsf{T}\left(a(a - \mathbb{I})\hat{y}^l + a^2\hat{y}^m\right)\Gamma \tag{16.2b}$$

  Or a transformer in open $\Delta\Delta$ configuration where, from (15.26d),

  $$y^s_{jk} = y^s_{kj} := \Gamma^\mathsf{T} a \tilde{y}^l \Gamma, \quad y^m_{jk} := \Gamma^\mathsf{T}(\mathbb{I} - a)\,\tilde{y}^l \Gamma, \quad y^m_{kj} := \Gamma^\mathsf{T}\left(a(a - \mathbb{I})\tilde{y}^l + a^2\tilde{y}^m\right)\Gamma \tag{16.2c}$$

  which is the same as (16.2b) with $\hat{y}^l$ and $\hat{y}^m$ replaced by the leakage and shunt admittances $\tilde{y}^l := \mathrm{diag}(y^{la}, y^{lb}, 0)$ and $\tilde{y}^m := \mathrm{diag}(y^{ma}, y^{mb}, 0)$ respectively of the open transformer.

When $y^s_{jk} \neq y^s_{kj}$ is allowed, this model can also model transformers in other configurations:

- A transformer in ΔY configuration with zero neutral voltage ($V^n_k = 0$) where, from (15.24a),

  $$y^s_{jk} := \Gamma^\mathsf{T} a \hat{y}^l, \quad y^s_{kj} := a\hat{y}^l\Gamma, \quad y^m_{jk} := \Gamma^\mathsf{T}\hat{y}^l(\Gamma - a), \quad y^m_{kj} := a\hat{y}^l(a - \Gamma) + a^2\hat{y}^m \tag{16.3a}$$

- A transformer in $Y\Delta$ configuration with zero neutral voltage ($V^n_j = 0$) where, from (15.25a),

  $$y^s_{jk} := a\hat{y}^l\Gamma, \quad y^s_{kj} := \Gamma^\mathsf{T} a\hat{y}^l, \quad y^m_{jk} := \hat{y}^l(\mathbb{I} - a\Gamma), \quad y^m_{kj} := \Gamma^\mathsf{T}\left(a\hat{y}^l(a\Gamma - \mathbb{I}) + a^2\hat{y}^m\Gamma\right) \tag{16.3b}$$

Remark 16.1 (Transformer models)**.** 1 We emphasize that the models (16.2) (16.3) assume that, for three-phase transformers with Y configuration either in the primary or secondary side, their neutrals are directly grounded so the neutral voltages $V_j^n = 0$ or $V_k^n = 0$.

2 While the shunt admittances y_{jk}^m and y_{kj}^m are typically equal for a transmission or distribution line, they are typically different for a transformer. Moreover the shunt admittances (y_{jk}^m, y_{kj}^m) of the line model of a transformer are generally nonzero *even if* the shunt admittances $\hat{y}^m := \mathrm{diag}(y^{ma}, y^{mb}, y^{mc})$ (or $\tilde{y}^m := \mathrm{diag}(y^{ma}, y^{mb}, 0)$ for open $\Delta\Delta$ transformer) of the constituent single-phase transformers are assumed zero.

3 The series and shunt admittance matrices $\left(y_{jk}^s, y_{jk}^m\right)$ and $\left(y_{kj}^s, y_{kj}^m\right)$ in (16.2) are all complex symmetric. None of them are symmetric for series and shunt admittances in (16.3). Moreover the admittance matrices corresponding to Δ configuration in the primary or secondary side are singular, i.e., unlike for single-phase transformers, none of the admittances $y_{jk}^s, y_{jk}^m, y_{kj}^s, y_{kj}^m$ may have an inverse.

For simplicity we often restrict ourselves to the special case where $y_{jk}^s = y_{kj}^s$. In this case we characterize a line (j,k) by three 3×3 series and shunt admittance matrices $\left(y_{jk}^s, y_{jk}^m, y_{kj}^m\right)$. With $y_{jk}^s = y_{kj}^s$, (16.1) reduces to

$$I_{jk} = y_{jk}^s(V_j - V_k) + y_{jk}^m V_j, \qquad I_{kj} = y_{jk}^s(V_k - V_j) + y_{kj}^m V_k \tag{16.4a}$$

or in terms Y_{jk}:

$$\begin{bmatrix} I_{jk} \\ I_{kj} \end{bmatrix} = \underbrace{\begin{bmatrix} y_{jk}^s + y_{jk}^m & -y_{jk}^s \\ -y_{jk}^s & y_{jk}^s + y_{kj}^m \end{bmatrix}}_{Y_{jk}} \begin{bmatrix} V_j \\ V_k \end{bmatrix} \tag{16.4b}$$

which is now block symmetric (see Definition 16.1). We say Y_{jk} has a *three-phase* Π *circuit representation* in the sense that its external behavior is the same as the external behavior (15.8a) of a three-phase transmission line; see Figure 15.2.

From (16.3) this more restrictive Y_{jk} cannot be used to model transformers in ΔY and $Y\Delta$ configurations. It is however still widely used. We therefore often adopt this model and will explicitly state it as assumption C16.1 below when we use it.

16.1.2 *IV* relation

Associated with each bus j are three nodal variables $(V_j, I_j, s_j) \in \mathbb{C}^9$ representing the nodal voltage, current injection, and power injection respectively at the terminal of the device connected to bus j. To simplify notation we assume, without loss of generality, that at most one single-terminal device (source or load) is connected to a bus but one

or more lines can be connected to a bus.[1] The bus current and power injection (I_j, s_j) at bus j therefore refers unambiguously to the injection from the unique device at bus j. As explained in Chapters 14.3.3 and 14.3.4, the external behavior of a three-phase device is described by the relation between (V_j, I_j) or that between (V_j, s_j). We can assume without loss of generality that these three-phase devices are ideal (see Chapter 15.1.4) and their behavior is summarized in Tables 14.3 and 14.4.

Let $(V, I, s) := \left(V_j, I_j, s_j, j \in \overline{N}\right) \in \mathbb{C}^{3(N+1)}$ be nodal variables over the entire network. As for a single-phase network, a three-phase network model is a relation between the *terminal* voltage and current (V, I) or a relation between the *terminal* voltage and power (V, s), independent of the internal Y or Δ configurations of the three-phase devices that are connected by the lines. In this subsection we derive the linear IV relation defined by an admittance matrix Y and show that Y defines a single-phase equivalent circuit of the three-phase network. In the next subsection we derive the sV relation in the form of nonlinear power flow equations. In both cases the extension of the line model (16.1) to a network is the nodal current or power balance equations:

$$I_j = \sum_{k:j\sim k} I_{jk}, \qquad s_j = \sum_{k:j\sim k} \operatorname{diag}\left(S_{jk}\right), \qquad j \in \overline{N}$$

where $S_{jk} := V_j I_{jk}^{\mathsf{H}}$ are matrices defined in (15.8b).

Network admittance matrix Y.

Substitute the line currents (16.1) into the current balance equation to get

$$I_j = \sum_{k:j\sim k} I_{jk} = \sum_{k:j\sim k} y_{jk}^s (V_j - V_k) + \left(\sum_{k:j\sim k} y_{jk}^m\right) V_j$$

Therefore

$$I_j = \left(\left(\sum_{k:j\sim k} y_{jk}^s\right) + y_{jj}^m\right) V_j - \sum_{k:j\sim k} y_{jk}^s V_k, \qquad j \in \overline{N} \tag{16.5a}$$

where

$$y_{jj}^m := \sum_{k:j\sim k} y_{jk}^m \tag{16.5b}$$

Note that I_j is the net current injection.[2] In vector form, this relates the bus current vector $I := (I_0, \ldots, I_N)$ to the bus voltage vector $V := (V_0, \ldots, V_N)$:

$$I = YV \tag{16.6a}$$

[1] If K three-phase devices with terminal current injections $I_{j1}, \ldots, I_{jK}$ are connected to bus j then the net bus injection is $I_j := \sum_k I_{jk}$. Unless otherwise specified we assume $K = 1$.

[2] If there is a nodal shunt admittance load y_j^{sh}, e.g., a capacitor bank, in addition to a device whose terminal injection is $\tilde{I}_j$, then the net injection from bus j to the rest of the network is $I_j = \tilde{I}_j - y_j^{\mathrm{sh}} V_j$. This assumes that y_j^{sh} connects bus j to the ground and the terminal voltage V_j is defined with respect to the ground.

through a $3(N+1)\times 3(N+1)$ admittance matrix Y where its 3×3 submatrices $Y_{jk} \in \mathbb{C}^{3\times 3}$ are given by

$$Y_{jk} := \begin{cases} -y^s_{jk}, & j \sim k \ \ (j \neq k) \\ \sum_{l:j\sim l} y^s_{jl} + y^m_{jj}, & j = k \\ 0 & \text{otherwise} \end{cases} \tag{16.6b}$$

The submatrices Y_{jk} and Y_{kj} may be different if (j,k) models a three-phase transformer in ΔY or $Y\Delta$ configuration.

Definition 16.1 (Block symmetry and block row sum)**.** Given a matrix $A \in \mathbb{C}^{3n\times 3n}$, partition it into $n\times n$ blocks of 3×3 submatrices. Denote by $A_{jk} \in \mathbb{C}^{3\times 3}$ its jkth submatrix.

1 A is called *block symmetric* if $A_{jk} = A_{kj}$ for all $j,k = 1,\ldots,n$.
2 A is said to have zero *block row sums* if $\sum_k A_{jk} = 0$ for all $j = 1,\ldots,n$.

As discussed in Chapter 15.2.3 a matrix can be symmetric but not block symmetric, and vice versa. Symmetry of a matrix is determined only by its off-diagonal entries but its diagonal entries can be arbitrary. Block symmetry is determined only by its off-diagonal blocks but its diagonal blocks can be arbitrary. A symmetric matrix A is block symmetric if, in addition, all its off-diagonal blocks are themselves symmetric, i.e., $A^{\mathsf{T}}_{jk} = A_{jk}$, for all $j \neq k$. A block symmetric A is symmetric if, in addition, all blocks A_{jk}, including the diagonal blocks, are symmetric (Exercise 16.1). We will remark on zero block row sums below after introducing single-phase equivalent circuit.

In general an admittance matrix Y defined by (16.6) may neither be block symmetric nor symmetric. If the series admittances $y^s_{jk} = y^s_{kj}$ for all lines $(j,k) \in E$ then the admittance matrix Y is block symmetric and hence has a three-phase Π circuit representation. As in Chapter 4 we label the following assumption and will explicitly state it when it is required:

> C16.1: The series admittance matrices $y^s_{jk} = y^s_{kj}$ for every line $(j,k) \in E$, so that the admittance matrix Y is block symmetric.

If every $(j,k) \in E$ models a transmission or distribution line or a transformer described by (16.2), then Y is block symmetric with a three-phase Π circuit representation. If some $(j,k) \in E$ model transformers described by (16.3), however, then Y is not.

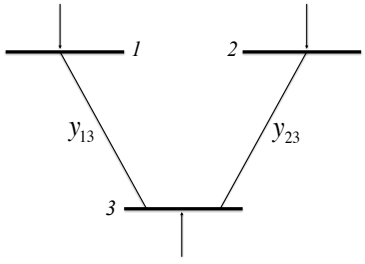
The expression (4.12) for Y for a single-phase network generalizes directly to the three-phase setting. Let $C \in \{-\mathbb{I},0,\mathbb{I}\}^{|N|\times|E|}$ be the bus-by-line incidence matrix defined by:

$$C_{jl} = \begin{cases} \mathbb{I} & \text{if } l = j \to k \text{ for some bus } k \\ -\mathbb{I} & \text{if } l = i \to j \text{ for some bus } i \\ 0 & \text{otherwise} \end{cases}$$

where $\mathbb{I}$ is the identity matrix of size 3. Let $Y^s := \text{diag}\left(y_l^s, l \in E\right)$ be the $3|E| \times 3|E|$ block diagonal matrix with the series admittance matrices $y_l^s \in \mathbb{C}^{3\times 3}$ as its diagonal submatrices. Let $Y^m := \text{diag}\left(y_{jj}^m, j \in \overline{N}\right)$ be the $|\overline{N}| \times |\overline{N}|$ block diagonal matrix with the total shunt admittances $y_{jj}^m \in \mathbb{C}^{3\times 3}$ in (16.5b) as its diagonal submatrices. Then the admittance matrix in (16.6b) is, when $y_{jk}^s = y_{kj}^s$,

$$Y = C Y^s C^{\mathsf{T}} \; + \; Y^m$$

Example 16.1. The admittance matrix Y for a 3-terminal network with zero shunt admittances is shown in Figure 16.2. □



(a) 3-bus example.

$$Y = \begin{bmatrix} \begin{bmatrix} y_{13}^{aa} & y_{13}^{ab} & y_{13}^{ac} \\ y_{13}^{ba} & y_{13}^{bb} & y_{13}^{bc} \\ y_{13}^{ca} & y_{13}^{cb} & y_{13}^{cc} \end{bmatrix} & 0 & -\begin{bmatrix} y_{13}^{aa} & y_{13}^{ab} & y_{13}^{ac} \\ y_{13}^{ba} & y_{13}^{bb} & y_{13}^{bc} \\ y_{13}^{ca} & y_{13}^{cb} & y_{13}^{cc} \end{bmatrix} \\ 0 & \begin{bmatrix} y_{23}^{aa} & y_{23}^{ab} & y_{23}^{ac} \\ y_{23}^{ba} & y_{23}^{bb} & y_{23}^{bc} \\ y_{23}^{ca} & y_{23}^{cb} & y_{23}^{cc} \end{bmatrix} & -\begin{bmatrix} y_{23}^{aa} & y_{23}^{ab} & y_{23}^{ac} \\ y_{23}^{ba} & y_{23}^{bb} & y_{23}^{bc} \\ y_{23}^{ca} & y_{23}^{cb} & y_{23}^{cc} \end{bmatrix} \\ -\begin{bmatrix} y_{13}^{aa} & y_{13}^{ab} & y_{13}^{ac} \\ y_{13}^{ba} & y_{13}^{bb} & y_{13}^{bc} \\ y_{13}^{ca} & y_{13}^{cb} & y_{13}^{cc} \end{bmatrix} & -\begin{bmatrix} y_{23}^{aa} & y_{23}^{ab} & y_{23}^{ac} \\ y_{23}^{ba} & y_{23}^{bb} & y_{23}^{bc} \\ y_{23}^{ca} & y_{23}^{cb} & y_{23}^{cc} \end{bmatrix} & \begin{bmatrix} y_{13}^{aa}+y_{23}^{aa} & y_{13}^{ab}+y_{23}^{ab} & y_{13}^{ac}+y_{23}^{ac} \\ y_{13}^{ba}+y_{23}^{ba} & y_{13}^{bb}+y_{23}^{bb} & y_{13}^{bc}+y_{23}^{bc} \\ y_{13}^{ca}+y_{23}^{ca} & y_{13}^{cb}+y_{23}^{cb} & y_{13}^{cc}+y_{23}^{cc} \end{bmatrix} \end{bmatrix}$$

(b) Admittance matrix Y.

Figure 16.2 The admittance matrix Y for a 3-terminal network with no shunt admittances.

Single-phase equivalent circuit.

The $3(N+1)\times 3(N+1)$ admittance matrix Y in (16.6) defines a single-phase equivalent circuit of the three-phase network. Recall that a three-phase network can be represented

by a graph $G := (\overline{N}, E)$ where $\overline{N}$ is a set of $N+1$ three-phase buses and E is a set of three-phase lines. The admittance matrix Y induces a network graph $G^{3\phi} := (\overline{N}^{3\phi}, E^{3\phi})$ where $\overline{N}^{3\phi}$ has $3(N+1)$ buses. Each bus in $\overline{N}^{3\phi}$ is indexed by $j\phi$ with $j \in \overline{N}, \phi \in \{a,b,c\}$ in the original network G. Each line in $E^{3\phi}$ is indexed by $(j\phi, k\phi')$. There is a line between bus $j\phi$ and another distinct bus $k\phi'$ in $G^{3\phi}$ if and only if $Y_{jk}^{\phi\phi'}$ is nonzero. We call this graph $G^{3\phi}$ the *single-phase equivalent* (circuit) of the three-phase network G. All the single-phase modeling and analysis developed in earlier chapters can be directly applied to this single-phase equivalent.

When shunt admittances are assumed zero, $y_{jk}^m = y_{kj}^m = 0$ for all $(j,k) \in E$, the $3(N+1) \times 3(N+1)$ admittance matrix Y has zero block row sums (Definition 16.1), because

$$Y_{jj} = \sum_{k:(j,k)\in E} y_{jk}^s = \sum_k -Y_{jk}, \qquad j \in \overline{N}$$

so that $\sum_k Y_{jk} = 0$ for all j. Suppose Y has zero block row sums. Then Y also has zero block column sums if and only if Y is block symmetric. The matrix has zero row sums if $\sum_{k,\phi'} Y_{j\phi,k\phi'} = 0$ for all $j\phi$. This is equivalent to

$$\sum_{k,\phi'} Y_{j\phi,k\phi'} = \sum_{\phi'\in\{a,b,c\}} y_{jj}^{\phi\phi'} - \sum_{\substack{k:(j,k)\in E \\ \phi'\in\{a,b,c\}}} y_{jk}^{\phi\phi'} = 0, \qquad j\phi \in \overline{N} \times \{a,b,c\}$$

i.e., zero row sums requires only that the 3×3 matrix $\sum_k Y_{jk}$ has zero row sums, whereas zero block row sums requires that $\sum_k Y_{jk}$ is a zero matrix. Hence if a matrix has zero block row sums, then all its row sums are zero, but the converse does not necessarily hold.

In general Y is not symmetric (nor block symmetric), i.e., it may not satisfy C4.1 as the admittance matrix of a single-phase network. It is symmetric, and block symmetric, under the following condition:

> C16.2: In addition to C16.1, all series and shunt admittance matrices $y_{jk}^s, y_{jk}^m, y_{kj}^m$ are complex symmetric, so that the admittance matrix Y is both symmetric and block symmetric.

Suppose all transmission and distribution line models satisfy C16.2 (in particular, it satisfies assumptions C15.1 and C15.2). If every $(j,k) \in E$ models a transmission or distribution line or a transformer described by (16.2), then Y is not only block symmetric, but also symmetric (hence satisfying C4.1). Therefore Y has a three-phase Π circuit representation and the admittance matrix of its single-phase equivalent is complex symmetric. If some $(j,k) \in E$ models transformers described by (16.3), however, then Y is neither symmetric nor block symmetric.

Radial network.

Even when the multiphase network G is radial (i.e., with tree topology), its single-phase equivalent $G^{3\phi}$ is a meshed network (i.e., has cycles), but in that case, $G^{3\phi}$ has a radial macro-structure in which each line is represented as a clique (complete subgraph). Specifically $G^{3\phi}$ has a maximal clique consisting of the set $\left\{j\phi, k\phi' \in \overline{N}^{3\phi} : \phi, \phi' \in \{a,b,c\}\right\}$ of buses if and only if (j,k) is a line in G; see Figure 16.3. The corresponding principal submatrix $Y_{G^{3\phi}}(j,k) \in \mathbb{C}^{6\times 6}$ of Y is:

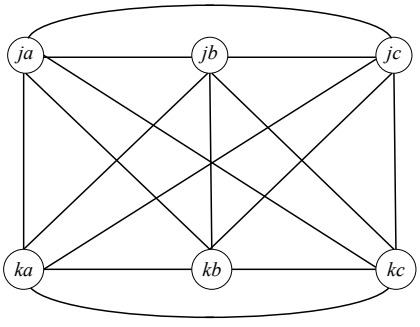


Figure 16.3 A clique of $G^{3\phi}$ corresponding to line (j,k) in G.

$$Y_{G^{3\phi}}(j,k) = \begin{bmatrix} Y_{jj} & Y_{jk} \\ Y_{kj} & Y_{kk} \end{bmatrix}$$

We will explain in Chapter **??** that $G^{3\phi}$ is a chordal graph which can be exploited to simplify the semidefinite relaxation of optimal power problems.

16.1.3 Invertibility of Y, Y_{22} and Y/Y_{22}

In this subsection we study the invertibility and properties of Y, Y_{22} and its Schur complement Y/Y_{22}. These results extend those in Chapter 4.2.3 from single-phase to three-phase networks.

Invertibility of Y.

Recall that a real matrix G is positive semidefinite (or positive definite), denoted $G \succeq 0$ (or $G \succ 0$), if G is symmetric and $v^{\mathsf{T}} G v \geq 0$ (or $v^{\mathsf{T}} G v > 0$) for all real vectors v (see Remark A.1 in Appendix A.5). Under assumption C16.2 ($y^s_{jk} = y^s_{kj}$, y^m_{jk} and y^m_{kj} are complex symmetric) the admittance matrix $Y \in \mathbb{C}^{3(N+1)\times 3(N+1)}$ is both symmetric and block symmetric. Write admittances in terms of their real and imaginary parts, $y^s_{jk} =$

$g^s_{jk} + \mathbf{i}b^s_{jk}$, $y^m_{jk} = g^m_{jk} + \mathbf{i}b^m_{jk}$, and $y^m_{kj} = g^m_{kj} + \mathbf{i}b^m_{kj}$. Consider the following conditions on the conductances $g^s_{jk}, g^m_{jk}, g^m_{kj} \in \mathbb{R}^{3\times 3}$:

C16.3: For all lines $(j,k) \in E$, $g^s_{jk} \succeq 0$, $g^m_{jk} \succeq 0$, $g^m_{kj} \succeq 0$.
C16.4a: For all buses $j \in \overline{N}$, $g^m_{jj} := \sum_{k:k\sim j} g^m_{jk} \succ 0$, i.e., for all j, there exists a line $(j,k) \in E$ such that $g^m_{jk} \succ 0$
C16.4b: For all lines $(j,k) \in E$, $g^s_{jk} \succ 0$. Furthermore there exists a line $(j',k') \in E$ such that $g^m_{j'k'} \succ 0$.
C16.4c: For all lines $(j,k) \in E$, $g^s_{jk} \succ 0$. Furthermore there exists a line $(j',k') \in E$ such that the intersection of the null spaces of $g^m_{j'k'}$ and $g^m_{k'j'}$ is $\{0\}$.

Condition C16.4b is a special case of C16.4c which does not require positive definiteness of g^m_{jk}. The next result extends Theorems 4.2, 4.3, and 4.9 in Chapter 4.2.3 from single-phase to three-phase networks.

Theorem 16.1. Suppose the network is connected and the admittance matrix $Y \in \mathbb{C}^{3(N+1)\times 3(N+1)}$ satisfies C16.2. If the conductance matrices $g^s_{jk}, g^m_{jk}, g^m_{kj} \in \mathbb{R}^{3\times 3}$ satisfy conditions C16.3 and one of C16.4a, C16.4b, C16.4c, then

1 The admittance matrix $Y^{-1} \in \mathbb{C}^{3(N+1)\times 3(N+1)}$ exists and is symmetric. Moreover both $\mathrm{Re}(Y) \succ 0$ and $\mathrm{Re}(Y^{-1}) \succ 0$.

In addition if $Y =: \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{12}^\mathsf{T} & Y_{22} \end{bmatrix}$ with invertible Y_{22}, then

2. The Schur complement $Y/Y_{22} := Y_{11} - Y_{12}Y_{22}^{-1}Y_{12}^\mathsf{T}$ of Y_{22} is symmetric and invertible. Moreover both $\mathrm{Re}(Y/Y_{22}) \succ 0$ and $\mathrm{Re}\big((Y/Y_{22})^{-1}\big) \succ 0$.

Proof Let $G := \mathrm{Re}(Y) \in \mathbb{R}^{3(N+1)\times 3(N+1)}$. We will show that $G \succ 0$. The claims then follow from Theorems 4.2 and 4.9.

Fix any real vector $\rho \in \mathbb{R}^{3(N+1)}$ and decompose it into $\rho =: (\rho_j, j \in \overline{N})$ with $\rho_j \in \mathbb{R}^3$. We have using (16.6b) and (16.5b)

$$\rho^\mathsf{T} G \rho = \sum_j \sum_{k:k\sim j} \left(\rho_j^\mathsf{T} g^s_{jk} \rho_j - \rho_j^\mathsf{T} g^s_{jk} \rho_k \right) + \sum_{j\in\overline{N}} \rho_j^\mathsf{T} g^m_{jj} \rho_j \tag{16.7a}$$

$$= \sum_{(j,k)\in E} \left(\rho_j^\mathsf{T} g^s_{jk} \rho_j - \rho_j^\mathsf{T} g^s_{jk} \rho_k - \rho_k^\mathsf{T} g^s_{kj} \rho_j + \rho_k^\mathsf{T} g^s_{kj} \rho_k \right) + \sum_j \sum_{k:k\sim j} \rho_j^\mathsf{T} g^m_{jk} \rho_j \tag{16.7b}$$

$$= \sum_{(j,k)\in E} (\rho_j - \rho_k)^\mathsf{T} g^s_{jk} (\rho_j - \rho_k) + \sum_{(j,k)\in E} \left(\rho_j^\mathsf{T} g^m_{jk} \rho_j + \rho_k^\mathsf{T} g^m_{kj} \rho_k \right) \tag{16.7c}$$

where the last equality follows because $g^s_{jk} = g^s_{kj}$ for all $(j,k) \in E$ by C16.2. Since $g^s_{jk}, g^m_{jk}, g^m_{kj} \in \mathbb{R}^{3\times 3}$ are positive semidefinite for all lines $(j,k) \in E$ by C16.3, every

summand is nonnegative and hence $\rho^{\mathsf{T}} G \rho = 0$ if and only if every summand is zero. We examine each of the three cases:

- *C16.4a holds*: Then for all buses $j \in \overline{N}$, $\rho_j^{\mathsf{T}} g_{jj}^m \rho_j > 0$ unless $\rho_j = 0$. Therefore for the second summation in (16.7a) to be zero we must have $\rho_j = 0$ for all $j \in \overline{N}$. This implies that $G \succ 0$.
- *C16.4b holds*: For the first summation in (16.7c) to be zero we must have $\rho_j = \rho_k$ for all $(j,k) \in E$. Since the network is connected, this implies that $\rho_j = \rho_1$ for all $j \in \overline{N}$. The second summation in (16.7b) then becomes, if $\rho_1 \neq 0$,

$$\sum_j \sum_{k:k\sim j} \rho_j^{\mathsf{T}} g_{jk}^m \rho_j \;=\; \rho_1^{\mathsf{T}} \left(\sum_j \sum_{k:k\sim j} g_{jk}^m \right) \rho_1 \;\geq\; \rho_1^{\mathsf{T}} g_{j'k'}^m \rho_1 \;>\; 0$$

  Therefore $\rho^{\mathsf{T}} G \rho > 0$ unless $\rho = 0$, i.e., $G \succ 0$.
- *C16.4c holds*: As for the case of C16.4b, we must have $\rho_j = \rho_1$ for all $j \in \overline{N}$. Then the second summation in (16.7c) becomes, if $\rho_1 \neq 0$,

$$\sum_{(j,k)\in E} \left(\rho_j^{\mathsf{T}} g_{jk}^m \rho_j + \rho_k^{\mathsf{T}} g_{kj}^m \rho_k \right) \;\geq\; \rho_1^{\mathsf{T}} \left(g_{j'k'}^m + g_{k'j'}^m \right) \rho_1 \;>\; 0$$

  where the last inequality follows because $g_{j'k'}^m$ and $g_{k'j'}^m$ are positive semidefinite and their null spaces intersect only at the origin. Therefore $\rho^{\mathsf{T}} G \rho > 0$ unless $\rho = 0$, i.e., $G \succ 0$.

Hence in all three cases G is positive definite. Since Y is complex symmetric and Y_{22} is nonsingular by assumption, Theorems 4.2 and 4.9 complete the proof. □

Consider the following conditions on the conductances $b_{jk}^s, b_{jk}^m, b_{kj}^m \in \mathbb{R}^{3\times 3}$:

C16.5: For all lines $(j,k) \in E$, $b_{jk}^s \preceq 0$, $b_{jk}^m \preceq 0$, $b_{kj}^m \preceq 0$.
C16.6a: For all buses $j \in \overline{N}$, $b_{jj}^m := \sum_{k:k\sim j} b_{jk}^m \prec 0$, i.e., for all j, there exists a line $(j,k) \in E$ such that $b_{jk}^m \prec 0$
C16.6b: For all lines $(j,k) \in E$, $b_{jk}^s \prec 0$. Furthermore there exists a line $(j',k') \in E$ such that $b_{j'k'}^m \prec 0$.
C16.6c: For all lines $(j,k) \in E$, $b_{jk}^s \prec 0$. Furthermore there exists a line $(j',k') \in E$ such that the intersection of the null spaces of $b_{j'k'}^m$ and $b_{k'j'}^m$ is $\{0\}$.

Condition C16.6b is a special case of C16.6c which does not require negative definiteness of b_{jk}^m. The next result extends Theorems 4.2, 4.4, and 4.9 in Chapter 4.2.3 from single-phase to three-phase networks. Its proof is left as Exercise 16.2.

Theorem 16.2. Suppose the network is connected and the admittance matrix $Y \in \mathbb{C}^{3(N+1)\times 3(N+1)}$ satisfies C16.2. If the susceptance matrices $b_{jk}^s, b_{jk}^m, b_{kj}^m \in \mathbb{R}^{3\times 3}$ satisfy conditions C16.5 and one of C16.6a, C16.6b, C16.6c, then

1 The admittance matrix $Y^{-1} \in \mathbb{C}^{3(N+1)\times 3(N+1)}$ exists and is symmetric. Moreover $\mathrm{Im}(Y) \prec 0$ and $\mathrm{Im}(Y^{-1}) \succ 0$.

In addition if $Y =: \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{12}^{\mathsf{T}} & Y_{22} \end{bmatrix}$ with invertible Y_{22}, then

2. The Schur complement $Y/Y_{22} := Y_{11} - Y_{12}Y_{22}^{-1}Y_{12}^{\mathsf{T}}$ of Y_{22} is symmetric and invertible. Moreover $\mathrm{Im}(Y/Y_{22}) \prec 0$ but $\mathrm{Im}\big((Y/Y_{22})^{-1}\big) \succ 0$.

□

The conditions in Theorem 16.1 not only ensure $\mathrm{Re}(Y) \succ 0$ and those in Theorem 16.2 not only ensure $\mathrm{Im}(Y) \prec 0$. Each set of conditions also ensures $\alpha^{\mathsf{H}} Y \alpha \neq 0$ for any nonzero $\alpha \in \mathbb{C}^{3(N+1)}$ (Exercise 16.3). Since a necessary condition for Y to be singular is the existence of a nonzero α with $\alpha^{\mathsf{H}} Y \alpha = 0$, these conditions imply the invertibility of Y, as expected, and extend the sufficient conditions in Theorems 4.3 and 4.4 to three-phase networks.

Remark 16.2. The admittance matrix of a three-phase transformer involving Δ configuration is singular (see (15.16) or (15.36)). This causes the admittance matrix Y of a network that contains such transformers to be singular. A proposal in the literature is to add a small shunt admittance (diagonal entries) to the admittance matrix of such a transformer to make it nonsingular. □

Invertibility of Y_{22} when $y^m_{jk} = y^m_{kj} = 0$.

Let $A \subsetneq \overline{N}$ and Y_A be the $3|A| \times 3|A|$ principal submatrix of Y consisting of row and column blocks Y_{jk} with $j, k \in A$. Suppose the shunt admittances are zero, $y^m_{jk} = y^m_{kj} = 0$ so that the admittance matrix Y has zero block row sums and is not invertible. The next result provides a set of simple sufficient conditions for a principal submatrix Y_A to be invertible when A is a *strict* subset of $\overline{N}$. Its proof is similar to those of Theorems 4.5 and 4.6 and left as Exercise 16.4.

Theorem 16.3. Suppose the network is connected and the admittance matrix $Y \in \mathbb{C}^{3(N+1)\times 3(N+1)}$ satisfies C16.2. Suppose $y^m_{jk} = y^m_{kj} = 0$ for all lines $(j,k) \in E$. Consider the principal submatrix $Y_A \in \mathbb{C}^{3|A|\times 3|A|}$ for a strict subset $A \subsetneq \overline{N}$.

1. If $g^s_{jk} \succ 0$ for all lines $(j,k) \in E$ then Y_A^{-1} exists and is symmetric. Moreover both $\mathrm{Re}(Y_A) \succ 0$ and $\mathrm{Re}(Y_A^{-1}) \succ 0$.
2. If $b^s_{jk} \prec 0$ for all lines $(j,k) \in E$ then Y_A^{-1} exists and is symmetric. Moreover $\mathrm{Im}(Y_A) \prec 0$ but $\mathrm{Im}(Y_A^{-1}) \succ 0$.

Even when not all g^s_{jk} are positive definite and not all b^s_{jk} are negative definite the

admittance matrix Y can still be invertible because they cannot be zero simultaneously. The next result extends Theorem 4.8 from single-phase to three-phase setting.

Theorem 16.4. Suppose the network is connected and the admittance matrix $Y \in \mathbb{C}^{3(N+1)\times 3(N+1)}$ satisfies C16.2. Suppose $y^m_{jk} = y^m_{kj} = 0$ for all lines $(j,k) \in E$. If $g^s_{jk} \geq 0$ and $b^s_{jk} \leq 0$ for all lines $(j,k) \in E$ then the principal submatrix $Y_A \in \mathbb{C}^{3|A|\times 3|A|}$ for a strict subset $A \subsetneq \overline{N}$ satisfies:

1 $\mathrm{Re}(Y_A) \succeq 0$, $\mathrm{Im}(Y_A) \preceq 0$.
2 Moreover $\mathrm{Re}(Y_A) - \mathrm{Im}(Y_A) \succ 0$.
3 Y_A^{-1} exists and is symmetric.

Proof The proof of Theorem 4.8 for single-phase network shows that G_A is diagonally dominant since $g^s_{jk} \in \mathbb{R}$ are nonnegative and hence its eigenvalues are nonnegative by the the Geršgorin disc theorem. In the three-phase case, we cannot use this argument since not every element of the 3×3 conductance matrix g^s_{jk} is nonnegative. We will use the argument in the proof of Theorem 16.1 (see (16.7)): for any real vector $\rho =: (\rho_j, j \in A)$ with $\rho_j \in \mathbb{R}^3$ we have, using $G_A := \mathrm{Re}(Y_A)$,

$$\begin{aligned}
\rho^{\mathsf{T}} G_A \rho &= \sum_j \sum_{k:k\sim j} \left(\rho_j^{\mathsf{T}} g^s_{jk} \rho_j - \rho_j^{\mathsf{T}} g^s_{jk} \rho_k \right) \\
&= \sum_{(j,k)\in E} \left(\rho_j^{\mathsf{T}} g^s_{jk} \rho_j - \rho_j^{\mathsf{T}} g^s_{jk} \rho_k - \rho_k^{\mathsf{T}} g^s_{kj} \rho_j + \rho_k^{\mathsf{T}} g^s_{kj} \rho_k \right) \\
&= \sum_{(j,k)\in E} (\rho_j - \rho_k)^{\mathsf{T}} g^s_{jk} (\rho_j - \rho_k)
\end{aligned}$$

where the last equality has used $g^s_{jk} = g^s_{kj}$ for all $(j,k) \in E$ from C16.2. Since $g^s_{jk} \geq 0$, $\rho^{\mathsf{T}} G_A \rho \geq 0$ for any ρ, i.e., $G_A \succeq 0$. Similar, using $B_A := \mathrm{Im}(Y_A)$, we have

$$\rho^{\mathsf{T}} B_A \rho = \sum_j \sum_{k:k\sim j} \left(\rho_j^{\mathsf{T}} b^s_{jk} \rho_j - \rho_j^{\mathsf{T}} b^s_{jk} \rho_k \right) = \sum_{(j,k)\in E} (\rho_j - \rho_k)^{\mathsf{T}} b^s_{jk} (\rho_j - \rho_k)$$

Therefore $\rho^{\mathsf{T}} B_A \rho \leq 0$ since $b^s_{jk} \leq 0$, i.e., $B_A \preceq 0$. This implies that $G_A - B_A \succeq 0$.

We now show that, indeed, $G_A - B_A \succ 0$ because the network is connected and $A \subset \overline{N}$ is a strict subset. The argument is the same as that for Theorem 4.8 for single-phase networks. For a $3n \times 3n$ matrix M, let $M[j,k]$ denote the 3×3 submatrix of M consisting of the jth row block and the kth row column. Since $G_A - B_A$ is real

symmetric, consider, for any nonzero real vector $\rho \in \mathbb{R}^{3|A|}$,

$$\begin{aligned}\rho^{\mathsf{T}}(G_A - B_A)\rho &= \sum_{j\in A}\sum_{k\in A}\rho_j^{\mathsf{T}}(G_A[j,k] - B_A[j,k])\rho_k \\ &= \sum_{j\in A}\sum_{\substack{k\in A:\\(j,k)\in E}}\rho_j^{\mathsf{T}}(-g^s_{jk}+b^s_{jk})\rho_k + \sum_{j\in A}\rho_j^{\mathsf{T}}\left(\sum_{\substack{k\in A:\\(j,k)\in E}}(g^s_{jk}-b^s_{jk}) + \sum_{\substack{k\notin A:\\(j,k)\in E}}(g^s_{jk}-b^s_{jk}\right. \\ &= \sum_{\substack{j,k\in A:\\(j,k)\in E}}(\rho_j-\rho_k)^{\mathsf{T}}(g^s_{jk}-b^s_{jk})(\rho_j-\rho_k) + \sum_{j\in A}\rho_j G_j \rho^{\mathsf{T}}\end{aligned}$$

where the third equality has used $g^s_{jk} = g^s_{kj}$ for all $(j,k) \in E$ from C16.2. Here $G_j := \sum_{k\notin A:(j,k)\in E}(g^s_{jk} - b^s_{jk})$ for $j \in A$ and the summation is not vacuous because the network is connected and $A \subsetneq \overline{N}$. For every line $(j,k) \in E$, $y^s_{jk} \neq 0$ and hence $g^s_{jk} - b^s_{jk} > 0$ since $g^s_{jk} \geq 0$ and $b^s_{jk} \geq 0$. This implies $G_j > 0$ as well for all $j \in A$. Therefore for $\rho^{\mathsf{T}}(G_A - B_A)\rho > 0$ for any real vector $\rho \neq 0$, i.e., $G_A - B_A \succ 0$.

Finally $G_A - B_A \succ 0$ implies that Y_A is nonsingular (it is clear that Y_A^{-1} is symmetric if it exists). The argument is exactly the same as that for Theorem 4.8 for single-phase networks. □

Application: admittance matrix Y identification.

Uniform lines.

Suppose all lines are of the same type specified by an impedance matrix y^{-1} per unit length. These lines differ only in their lengths. We will call y the *unit admittance*.[3] We show that this property is preserved under Schur complement. It means that the effective line admittances of the Kron-reduced admittance matrix Y/Y_A are also specified by the unit admittance y. This assumption makes the iterative construction of the Schur complement particularly simple.

Consider any $3(N+1)\times 3(N+1)$ complex symmetric matrix Y on a graph $G := (N,E)$ where its 3×3 (i,j)th blocks $Y[i,j]$ are given by:

$$Y[i,j] = \begin{cases} -\mu_{ij}\, y & (i,j)\in E \\ \left(\sum_{k:(i,k)\in E}\mu_{ik}\right) y & i = j \\ 0 & \text{otherwise}\end{cases} \tag{16.8}$$

where $y \in \mathbb{C}^{3\times 3}$ is complex symmetric. Suppose $\mathrm{Re}(y) \succ 0$ and $\mu_{ij} > 0$ for all $(i,j) \in E^0$. Then Theorem 4.2 implies that y^{-1} exists, is symmetric, and $\mathrm{Re}(y^{-1}) \succ 0$. Kron reduction preserves this structure.

Theorem 16.5. Suppose $\mathrm{Re}(y) \succ 0$ and $\mu_{ij} > 0$ for all $(i,j) \in E^0$ in the complex

[3]

symmetric matrix Y defined in (16.8). Let $Y =: \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{12}^{\mathsf{T}} & Y_{22} \end{bmatrix}$ with a $3n \times 3n$ nonsingular submatrix Y_{22}, $1 \le n \le N$.

1. The 3×3 (i,j)th blocks $(Y/Y_{22})[i,j]$ of the Schur complement Y/Y_{22} of Y_{22} of Y are given by

$$(Y/Y)[i,j] \;=\; \begin{cases} -\tilde{\mu}_{ij}\, y & i \rightsquigarrow j \\ \left(\sum_{k:i\rightsquigarrow k} \tilde{\mu}_{ik}\right) y & i = j \\ 0 & \text{otherwise} \end{cases} \tag{16.9}$$

   for some $\tilde{\mu}_{ij} = \tilde{\mu}_{ji} > 0$. Here $i \rightsquigarrow j$ if and only if there is a path in the underlying graph G connecting nodes i and j.
2. If the network is connected and the admittance matrix Y satisfies C16.2, then $(Y/Y_{22})^{-1}$ exists and is symmetric, and both $\mathrm{Re}(Y/Y_{22}) \succ 0$ and $\mathrm{Re}(Y/Y_{22})^{-1} \succ 0$.

Proof The Schur complement Y/Y_{22} is the admittance matrix describing the effective connectivity between nodes $1,\ldots,N-n+1$ obtained by eliminating interior nodes $N-n+2,\ldots,N+1$ by Kron reduction. We follow the approach of [169] to prove the theorem by induction on the interior nodes to be Kron reduced one by one. Define

$$A^0 := Y, \quad A^1 := A^0/A^0[n,n], \quad \cdots \quad A^n := A^{n-1}/A^{n-1}[N-n+2, N-n+2] = Y/Y_{22}$$

i.e., A^{l+1} is the admittance matrix for the graph after the last node in A^l has been Kron reduced, and hence $Y/Y_{22} = A^n$. Define the set of lines in the graph underlying A^0, $A^1,\ldots,A^n$ by

$$E^0 := E, \qquad E^l := \left\{(i,j) : A^l[i,j] \neq 0\right\}, \qquad l = 1,\ldots,k$$

Hence these sets are well-defined given the matrices $A^0, A^1,\ldots,A^n$. For $0 < l < n$, let the induction hypothesis be

$$A^l[i,j] \;=\; \begin{cases} -\mu_{ij}^l\, y & (i,j) \in E^l \\ \left(\sum_{k:(i,k)\in E^l} \mu_{ik}^l\right) y & i = j \\ 0 & \text{otherwise} \end{cases} \tag{16.10}$$

for some $\mu_{ij}^l = \mu_{ji}^l > 0$. Clearly A^0 satisfies (16.10). Suppose A^l satisfies (16.10). We now prove that $A^{l+1} := A^l/A^l[N-l+1, N-l+1]$ satisfies (16.10).

The 3×3 (i,j)th block $A^{l+1}[i,j]$ is given by

$$A^{l+1}[i,j] \;=\; A^l[i,j] - A^l[i,N-l+1]\left(A^l[N-l+1,N-l+1]\right)^{-1} A^l[j,N-l+1] \tag{16.11}$$

We consider 6 cases by substituting the induction hypothesis (16.10) into (16.11):

1. If $(i,j) \in E^l$ but either $(i,N-l+1) \notin E^l$ or $(j,N-l+1) \notin E^l$ then, substituting the induction hypothesis (16.10) into (16.11), we have $A^{l+1}[i,j] = -\mu_{ij}^{l+1}\, y$ where $\mu_{ij}^{l+1} := \mu_{ij}^l > 0$.

2. If $(i,j) \notin E^l$ but both $(i,N-l+1) \in E^l$ and $(j,N-l+1) \in E^l$ then

$$A^{l+1}[i,j] \;=\; -\mu^l_{i(N-l+1)}\, y \left(\sum_{k:(k,N-l+1)\in E^l} \mu^l_{k(N-l+1)}\, y \right)^{-1} \mu^l_{j(N-l+1)}\, y \;=\; -\mu^{l+1}_{ij}\, y$$

where

$$\mu^{l+1}_{ij} \;:=\; \mu^l_{i(N-l+1)}\, \mu^l_{j(N-l+1)} \left(\sum_{k:(k,N-l+1)\in E^l} \mu^l_{k(N-l+1)} \right)^{-1} \;>\; 0$$

3. If $(i,j) \in E^l$, $(i,N-l+1) \in E^l$ and $(j,N-l+1) \in E^l$ then

$$\begin{aligned} A^{l+1}[i,j] \;&:=\; -\mu^l_{ij}\, y - \mu^l_{i(N-l+1)}\, y \left(\sum_{k:(k,N-l+1)\in E^l} \mu^l_{k(N-l+1)}\, y \right)^{-1} \mu^l_{j(N-l+1)}\, y \\ &=\; -\mu^{l+1}_{ij}\, y \end{aligned}$$

where

$$\mu^{l+1}_{ij} \;:=\; \mu^l_{ij} + \mu^l_{i(N-l+1)}\, \mu^l_{j(N-l+1)} \left(\sum_{k:(k,N-l+1)\in E^l} \mu^l_{k(N-l+1)} \right)^{-1} \;>\; 0$$

4. If $i = j$ but $(i,N-l+1) \notin E^l$ then $A^{l+1}[i,i] = \left(\sum_{k:(i,k)\in E^{l+1}} \mu^{l+1}_{ik} \right) y$ where $\mu^{l+1}_{ik} := \mu^l_{ik} > 0$.
5. If $i = j$ and $(i,N-l+1) \in E^l$ then

$$\begin{aligned} A^{l+1}[i,i] \;&:=\; \left(\sum_{k:(i,k)\in E^l} \mu^l_{ik} \right) y - \mu^l_{i(N-l+1)}\, y \left(\sum_{k:(k,N-l+1)\in E^l} \mu^l_{k(N-l+1)}\, y \right)^{-1} \mu^l_{i(N-l+1)}\, y \\ &=\; \left(\sum_{k:(i,k)\in E^{l+1}} \mu^{l+1}_{ik} \right) y \end{aligned}$$

where $\mu^{l+1}_{ik} := \mu^l_{ik} > 0$ for $(i,k) \in E^l$ and $k = 1,\ldots,N-l+1$, and

$$\mu^{l+1}_{i(N-l+1)} \;:=\; \mu^l_{i(N-l+1)} \left(1 - \frac{\mu^l_{i(N-l+1)}}{\sum_{k:(k,N-l+1)\in E^l} \mu^l_{k(N-l+1)}} \right) \;>\; 0$$

6. Otherwise, $i \neq j$ and $(i,j) \notin E^l$ and $A^{l+1}[i,j] = 0$.

This completes the induction and the proof of part 1. Part 2 follows from $\mathrm{Re}(y) > 0$ and Theorem 16.1. □

16.1.4 sV relation

Power flow equations.

The power flow equations that relate bus injections $s := \left(s_j, j \in \overline{N}\right)$ and voltages $V := \left(V_j, j \in \overline{N}\right)$ can be obtained by applying the derivation for single-phase systems to the single-phase equivalent network $G^{3\phi}$. In particular the bus injection model in complex form is defined by the following power flow equation that expresses power balance at each bus $j\phi$ in terms of the elements $Y_{j\phi,k\phi'}$ of the $3(N+1)\times 3(N+1)$ admittance matrix Y defined in (16.6):

$$s_j^{\phi} = \sum_{\substack{k\in\overline{N}\\ \phi'\in\{a,b,c\}}} Y^{\mathsf{H}}_{j\phi,k\phi'} V_j^{\phi} \left(V_k^{\phi'}\right)^{\mathsf{H}}, \qquad j \in \overline{N},\ \phi \in \{a,b,c\} \tag{16.12a}$$

This directly generalizes (4.26b) from the single-phase setting to the three-phase setting. To generalize (4.26a) to the three-phase setting note that

$$s_j = \sum_{k:j\sim k} \operatorname{diag}\left(V_j I_{jk}^{\mathsf{H}}\right), \qquad j \in \overline{N}$$

where $s_j, V_j, I_{jk} \in \mathbb{C}^3$ are power injections, voltages, and line currents in all phases. We then have from (15.8)

$$s_j = \sum_{k:j\sim k} \operatorname{diag}\left(V_j (V_j - V_k)^{\mathsf{H}} \left(y^s_{jk}\right)^{\mathsf{H}} + V_j V_j^{\mathsf{H}} \left(y^m_{jk}\right)^{\mathsf{H}}\right), \qquad j \in \overline{N} \tag{16.12b}$$

Power flow analysis and optimization for unbalanced three-phase networks can be conducted using both forms of the bus injection model (16.12). In particular (16.12b) will be used in Chapter **??** to prove the equivalence of the branch flow model and the bus injection model (Theorem 17.1). The model (16.12) does not require condition C16.1 nor C16.2.

16.1.5 Overall model

Most power flow analysis or optimization applications involve three-phase devices, either in Y or Δ configuration, connected by three-phase lines. The lines may not be phase-decoupled and the sources and loads may not be balanced. In this subsection we compose an overall model consisting of the device modes of Chapter 14.3 and the network equations of this section. We use this overall model to formulate a general three-phase analysis problem in the next section.

The overall model consists of:

1 A network model that relates terminal voltage, current, and power (V, I, s). Any

equivalent model can be used, whichever is convenient for the problem under study, including:
- The (linear) current balance equation (16.5)(16.6).
- The (quadratic) power flow equation that defines the BIM model (16.12).

2 A device model for each three-phase device j. For ideal devices, this can either be:
- Its internal model (14.29) and the conversion rules (14.8) and (14.9)(14.10); or
- Its external model summarized in Tables 14.3 and 14.4 when only terminal quantities are needed.

For non-ideal devices, this can either be:
- Its internal model summarized in Table 14.2 and the conversion rules (14.8) and (14.9)(14.10); or
- Its external model summarized in Table 14.2 when only terminal quantities are needed.

If only voltage sources, current sources and impedances are involved then the overall model is linear, consisting of the nodal current balance equation (16.5)(16.6) and linear device models. If power sources are also involved then, even though (16.5)(16.6) can still be used as the network model, the overall model will be nonlinear because of nonlinear power source models.

16.2 Three-phase analysis

A device model relates its internal and terminal variables. A network equation relates the terminal variables of these devices. A typical three-phase analysis problem is: given a collection of voltage sources, current sources, power sources and impedances connected by three-phase lines, compute a certain set of external and internal variables. We first illustrate this in Chapter 16.2.1 using examples. We then formulate in Chapter 16.2.2 a general three-phase analysis problem and outline in Chapter 16.2.3 a solution strategy based on intuitions from these examples.

16.2.1 Examples

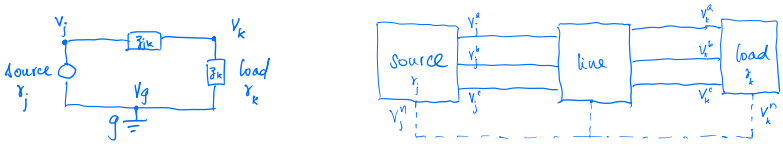
Three-phase analysis or optimization problems in practice are large-scale and can only be solved numerically. The goal of analyzing small examples is to gain intuition on how to specify these problems using the models developed in this chapter and illustrate their structure.

Consider a network of three-phase sources and impedances connected by three-phase lines. Assume without loss of generality that there is exactly one device at each

bus j. The quantities of interest include the internal variables $\left(V_j^{Y/\Delta}, I_j^{Y/\Delta}, s_j^{Y/\Delta}, \beta_j\right)$ and the terminal variables $(V_j, I_j, s_j, \gamma_j)$ at each bus j. The first set of examples is driven by voltage and current sources and the second set by power sources as well. In these examples we specify the parameters of a set of (ideal) devices and our objective is to compute the remaining internal and terminal voltages, currents, and powers.

The general analysis problem we formulate in Chapter 16.2.3 will specify γ_j for all voltage sources. The first example shows how γ_j arises in a circuit.

Example 16.2 (Reference voltage and γ_j)**.** We start with a single-phase circuit shown in Figure 16.4(a) where the source can be a voltage, current, or power source, the



(a) Single-phase network (b) Three-phase network

Figure 16.4 Reference voltage and constant γ_j

load is an impedance z_k, and the line is a series impedance z_{jk}. The terminal voltages (V_j, V_k, V_g) are defined with respect to an arbitrary but fixed reference point. The defining equations are

$$V_j - V_k = z_{jk} I_{jk}, \qquad V_k - V_g = z_k I_{jk}$$

Suppose the source is a current source with a given J_j from g to terminal j. Then the solution is:

$$I_{jk} = J_j, \qquad V_j = \left(z_{jk} + z_k\right) J_j + V_g \qquad V_k = z_k J_j + V_g$$

The terminal voltages depend on the choice of the reference point through the ground voltage V_g. For this example, $\gamma_j = \gamma_k = V_g$. In particular, if γ_j at the source is specified then γ_k at the load is fixed and the voltages (V_j, V_k) are uniquely determined. If we choose the reference point to be the ground then $\gamma_j = \gamma_k = V_g = 0$.

Consider now a three-phase system shown in Figure 16.4(b) where a device may or may not have a neutral line and the neutrals may or may not be grounded, directly or through an impedance. The voltage conversion rule between internal and terminal voltages for Y and Δ configured devices is (14.8)(14.9), reproduced here:

$$V_j = V_j^Y + \gamma_j \mathbf{1}, \qquad V_j^\Delta = \Gamma V_j \quad \text{or equivalently} \quad V_j = \Gamma^\dagger V_j^\Delta + \gamma_j \mathbf{1}$$

For Y-configured devices, $\gamma_j = V_j^n$, i.e., their neutral voltages with respect to the reference point. In general we need two of $\left(V_j, V_j^Y, \gamma_j\right)$ to determine the third. For Δ-configured devices, γ_j can be determined by specifying one of $\left(V_j^a, V_j^b, V_j^c\right)$ for each

device j. Knowing the vector V_j is sufficient to determine both the internal voltage V_j^Δ and γ_j. Knowing V_j^Δ however is not sufficient to determine V_j without γ_j. This is studied in detail in the next few examples. □

Voltage and current sources.

For a network driven by constant voltage and current sources without power sources, both the device models and the network equation $I = YV$ are linear. We will therefore focus on linear analysis to compute terminal and internal voltages and currents. Given (V_j, I_j) and $\left(V_j^{Y/\Delta}, I_j^{Y/\Delta}\right)$, external and internal powers can be computed. As we will see, the key step in our analysis is to solve for the *internal* currents $I_k^{Y/\Delta}$ of all impedances k, together with other quantities such as the terminal voltages V_j of current sources j, using the network equation, internal models of impedances and the voltage and current conversion rules. All other variables can then be derived. This solution strategy is extended in Chapter 16.2.3 to general three-phase networks.

Example 16.3 (Generator/load in Y configuration)**.** Consider the system in Figure 16.5 where an (ideal) voltage source is connected through a three-phase line to an impedance, both in Y configuration. We assume the neutrals are not grounded and there is not a neutral line. Suppose the following are specified:

- Voltage source $\left(E_j^Y, \gamma_j := V_j^n\right)$.
- Impedance $\left(z_k^Y, \gamma_k := V_k^n\right)$.
- Line parameters $\left(y_{jk}^s, y_{jk}^m, y_{kj}^m\right)$. In particular assumption C16.1 is satisfied.

Derive the terminal and internal voltages and currents $\left(V_k, I_k, V_k^Y, I_k^Y\right)$ of the impedance.

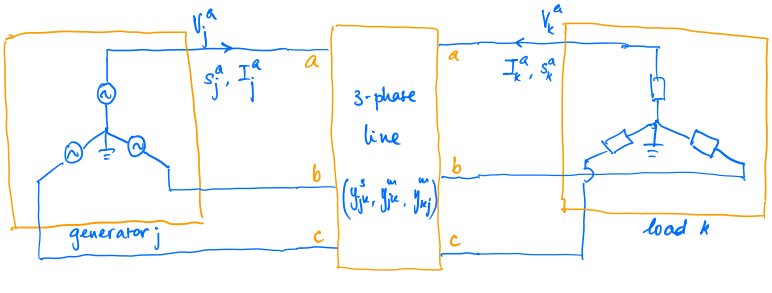


Figure 16.5 Example 16.3: A Y-configured generator connected through a three-phase line to a Y-configured impedance load.

Solution. The terminal voltages (V_j, V_k) and current injections (I_j, I_k) are related

according to (16.5):

$$I_j = y^s_{jk}\left(V_j - V_k\right) + y^m_{jk} V_j \tag{16.13a}$$

$$I_k = y^s_{jk}\left(V_k - V_j\right) + y^m_{kj} V_k \tag{16.13b}$$

From Table 14.3, the external models for the ideal voltage source and impedance in Y configuration are

$$V_j = E^Y_j + \gamma_j, \qquad V_k = -z^Y_k I_k + \gamma_k \mathbf{1} \tag{16.13c}$$

This is a system of 12 linear equations in 12 unknowns $\left(V_j, I_j\right)$ and $\left(V_k, I_k\right)$.

Substituting V_j from (16.13c) and the current conversion rule $I^Y_k = -I_k$ into (16.13b) we have

$$-I^Y_k = -y^s_{jk}\left(E^Y_j + \gamma_j\right) + \left(y^s_{jk} + y^m_{kj}\right) V_k \tag{16.14a}$$

Substituting V_k from (16.13c) we have

$$\left(\left(y^s_{jk} + y^m_{kj}\right)^{-1} + z^Y_k\right) I^Y_k = \left(y^s_{jk} + y^m_{kj}\right)^{-1} y^s_{jk} V_j - \gamma_k \mathbf{1} \tag{16.14b}$$

Hence

$$\begin{aligned} I^Y_k = -I_k &= \left(\hat{z}_{jk} + z^Y_k\right)^{-1} \hat{z}_{jk}\, y^s_{jk} V_j - \gamma_k \left(\hat{z}_{jk} + z^Y_k\right)^{-1} \mathbf{1} \\ &= \left(z^Y_k + z^s_{jk} + z^s_{jk} y^m_{kj} z^Y_k\right)^{-1} V_j - \gamma_k \left(\hat{z}_{jk} + z^Y_k\right)^{-1} \mathbf{1} \end{aligned}$$

where $z^s_{jk} := \left(y^s_{jk}\right)^{-1}$, $\hat{z}_{jk} := \left(y^s_{jk} + y^m_{kj}\right)^{-1}$ and $V_j = E^Y_j + \gamma_j$. From (16.13c)

$$\begin{aligned} V^Y_k &= z^Y_k I^Y_k = z^Y_k \left(z^Y_k + z^s_{jk} + z^s_{jk} y^m_{kj} z^Y_k\right)^{-1} V_j - \gamma_k z^Y_k \left(\hat{z}_{jk} + z^Y_k\right)^{-1} \mathbf{1} \\ V_k &= V^Y_k + \gamma_k \mathbf{1} = z^Y_k \left(z^Y_k + z^s_{jk} + z^s_{jk} y^m_{kj} z^Y_k\right)^{-1} V_j + \gamma_k \left(\mathbb{I} - \left(\hat{z}_{jk} + z^Y_k\right)^{-1}\right) \mathbf{1} \end{aligned}$$

□

In Example 16.3 the neutral voltages γ_j, γ_k are given explicitly. Often some of them are not explicitly given but additional information is available to indirectly specify them, i.e., to either compute their values, provide additional equations, or eliminate them in terms of other variables. For instance, if a neutral at bus j is grounded with zero grounding impedance and voltages are defined with respect to the ground then $\gamma_j = 0$. The next two examples study this in more detail. In Example 16.4, γ_k of the impedance z^Y_k is not explicitly given, but the additional information shows that its terminal voltage and current satisfy $V_k = -Z^Y_k I_k$; see (16.15). This means that the external model of the impedance is equivalent to that of an impedance with an effective internal impedance Z^Y_k with a known neutral voltage $\gamma_k = 0$. (See also Exercise 16.7 for another four-wire example).

Example 16.4 (Indirect specification of $\gamma_k = V_k^n$)**.** Repeat Example 16.3 with the modification that the impedance is specified only by z_k^Y (i.e., γ_k is not specified), and that the neutral of the impedance is connected through a given impedance z_k^n to the ground and not to the voltage source.

Solution. The equations (16.13) in Example 16.3 is now a system of 4 vector linear equations in 4 vector unknowns (V_j, V_k, I_j, I_k) and a scalar unknown, the unspecified neutral voltage $\gamma_k := V_k^n$ of the impedance, one more unknown than in Example 16.3. Since the neutral of the impedance is connected only to the ground (and not to the voltage source) through the impedance z_k^n, KCL and Ohm's law provide the additional equation

$$\gamma_k \;:=\; V_k^n \;=\; -z_k^n \left(\mathbf{1}^\mathsf{T} I_k\right)$$

Substituting into γ_k in (16.13c) we have $V_k = -z_k^Y I_k - z_k^n \mathbf{1}\mathbf{1}^\mathsf{T} I_k$. Hence the external device model (16.13c) in Example 16.3 can be replaced by

$$V_j \;=\; E_j^Y + \gamma_j \mathbf{1}, \qquad\qquad V_k \;=\; -\underbrace{\left(z_k^Y I_k + z_k^n \mathbf{1}\mathbf{1}^\mathsf{T}\right)}_{Z_k^Y} I_k \tag{16.15}$$

It says that the external behavior of the impedance z_k^Y when its neutral is grounded through z_k^n is equivalent to an impedance with an effective admittance Z_k^Y that is grounded directly so that $\gamma_k := V_k^n = 0$. The same computation leads to the same solution for (V_k, I_k) with the following replacement:

$$z_k^Y \;\to\; Z_k^Y, \qquad\qquad \gamma_k \;\to\; 0$$

□

The next example illustrates the case where the neutrals are not grounded but connected directly to each end of a four-wire line (also see Exercise 16.7). In this case, neither γ_j nor γ_k needs to be explicitly specified and can be determined from the network equation $I = YV$. This is an example where γ_j of a voltage source cannot be specified arbitrarily but is constrained by the network equation, in contrast to the three-wire models of Examples 16.3 and 16.4. This is because, when the neutral of the voltage source j is not grounded nor connected to bus k, the current I_j is determined only by (V_j, V_k) through (16.13a) and γ_j can be arbitrary. With the neutral wire, the additional constraint $I_j^{n'} = \mathbf{1}^\mathsf{T} I_j$ determines γ_j uniquely. Similarly for γ_k for the impedance.

Example 16.5 (Four-wire model)**.** Repeat Example 16.3 with the modification that the neutrals of both devices are ungrounded and are connected to the neutral wires at each end of a 4-wire line; see Figure 16.6. Suppose the following are specified:

- Voltage source E_j^Y.

- Impedance z_k^Y.
- Line parameters $\left(y_{jk}^s, y_{jk}^m, y_{kj}^m\right)$. In particular assumption C16.1 is satisfied.

Note that neither γ_j nor γ_k is explicitly specified.

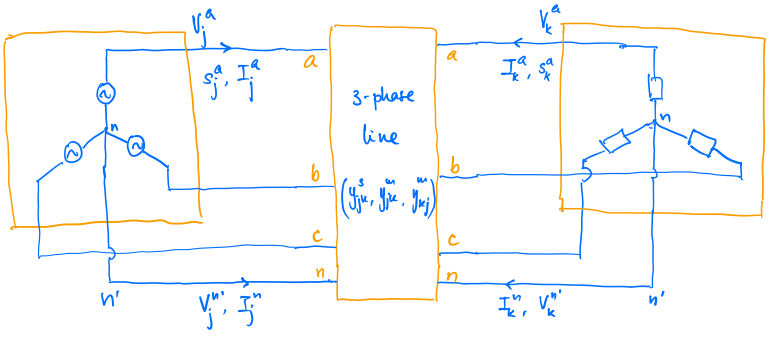


Figure 16.6 Example 16.5: A Y-configured generator connected through a four-wire line to a Y-configured impedance load.

Solution. To indicate the direction of internal currents on the neutral lines, we will use n to denote the internal neutral of a device and n' to denote the external terminal of the neutral line. In this example, $n' = n$ in the sense that $V_j^{n'n} = V_k^{n'n} = 0$. See Exercise 16.7 for the case where the neutrals of the voltage source and the load are connected through internal impedances $\left(z_j^n, z_k^n\right)$ to each end of the four-wire line, so $\left(V_j^{n'n}, V_k^{n'n}\right)$ may not be zero.

Define the terminal voltages (with respect to a common reference point) and currents in $\mathbb{C}^4$:

$$\hat{V}_j := \begin{bmatrix} V_j^a \\ V_j^b \\ V_j^c \\ V_j^{n'} \end{bmatrix}, \qquad \hat{V}_k := \begin{bmatrix} V_k^a \\ V_k^b \\ V_k^c \\ V_k^{n'} \end{bmatrix}, \qquad \hat{I}_j := \begin{bmatrix} I_j^a \\ I_j^b \\ I_j^c \\ I_j^n \end{bmatrix}, \qquad \hat{I}_k := \begin{bmatrix} I_k^a \\ I_k^b \\ I_k^c \\ I_k^n \end{bmatrix}$$

As noted above, $z_k^n = 0$ implies that $\gamma_j := V_j^n = V_j^{n'}$ and $\gamma_k := V_k^n = V_k^{n'}$ are variables to be determined. These terminal variables are related by $\hat{I} = \hat{Y}\hat{V}$ as in (16.13a) (16.13b), except that the admittance matrices are replaced by their four-wire counterparts:

$$\hat{I}_j = \hat{y}_{jk}^s \left(\hat{V}_j - \hat{V}_k\right) + \hat{y}_{jk}^m \hat{V}_j, \qquad \hat{I}_k = \hat{y}_{jk}^s \left(\hat{V}_k - \hat{V}_j\right) + \hat{y}_{kj}^m \hat{V}_k \tag{16.16a}$$

The external model of a four-wire voltage source in Y configuration is, since the neutrals

are ungrounded and connected to each other,

$$\hat{V}_j = \begin{bmatrix} E_j^{an}+V_j^n \\ E_j^{bn}+V_j^n \\ E_j^{cn}+V_j^n \\ V_j^n \end{bmatrix} = \underbrace{\begin{bmatrix} E_j^Y \\ 0 \end{bmatrix}}_{\hat{E}_j^Y} + \gamma_j \hat{\mathbf{1}} =: \hat{E}_j^Y + \gamma_j \hat{\mathbf{1}}, \quad I_j^n = \mathbf{1}^\mathsf{T} I_j \tag{16.16b}$$

where $\hat{\mathbf{1}}$ is the vector of all 1s of size 4 and $I_j := (I_j^a, I_j^b, I_j^c)$. Similarly the internal model of a four-wire impedance in Y configuration is, since the neutrals are ungrounded and connected to each other,

$$\hat{V}_k = \begin{bmatrix} z_k^{an} I_k^{an} \\ z_k^{bn} I_k^{bn} \\ z_k^{cn} I_k^{cn} \\ 0 \end{bmatrix} + \gamma_k \hat{\mathbf{1}} = -\begin{bmatrix} z_k & 0 \\ 0 & 0 \end{bmatrix} \hat{I}_k + \gamma_k \hat{\mathbf{1}}, \quad I_k^n = \mathbf{1}^\mathsf{T} I_k \tag{16.16c}$$

This is a set of 18 linear equations in 18 unknowns $(\hat{V}_j, \hat{I}_j, \gamma_j)$ and $(\hat{V}_k, \hat{I}_k, \gamma_k)$. It replaces (16.13) when neutrals are ungrounded and unconnected to each other and γ_j, γ_k must be given explicitly. It can be solved as in Example 16.3.

Exercise 16.6 expresses (γ_j, γ_k) in terms of the phase voltages and currents (V_j, V_k, I_j, I_k). □

The next example considers the setup of Example 16.3 in Δ configuration when the load is supplied by a voltage source. Exercise 16.8 considers the Δ configuration when the load is supplied by a current source. A voltage source $\left(E_j^\Delta, \gamma_j, \beta_j\right)$ is fully specified. A current source only needs to specify its internal current J_j^Δ if shunt admittances of the line are nonzero. Otherwise its zero-sequence voltage γ_j also needs to be specified (see Exercise 16.8 and Remark 16.8). Neither the zero-sequence voltage nor the zero-sequence current (γ_k, β_k) of the load need to be specified. They will be derived from network equations. A more detailed comparison between Example 16.3 (voltage source) and Exercise 16.8 (current source) is given in Tables 16.1 and 16.2 and in Remark 16.3. We will also explain in Remark 16.6 in Chapter 16.2.2 the asymmetry in the specification of voltage and current sources in Δ configuration.

Example 16.6 (Generator/load in Δ configuration)**.** Repeat Example 16.3 when the devices are in Δ configuration as shown in Figure 16.7, Suppose the following are specified:

- Voltage source $\left(E_j^\Delta, \gamma_j, \beta_j\right)$.
- Impedance z_k^Δ. (Note that the internal current β_k need not be specified and can be derived.)
- Line admittances $\left(\left(z_{jk}^s\right)^{-1}, y_{jk}^m = y_{kj}^m := 0\right)$. We have assumed assumption C16.1 and that shunt admittances are zero.

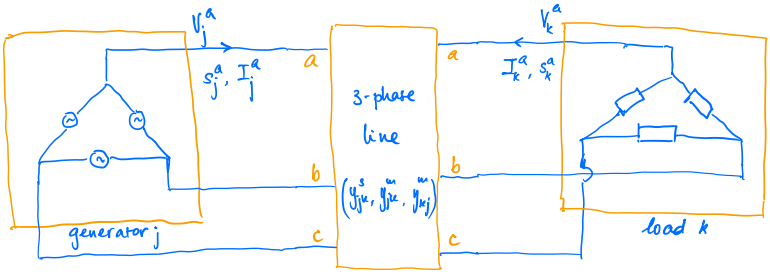


Figure 16.7 Example 16.6: Three-phase generator in Δ configuration connected through a three-phase line to an impedance load in Δ configuration.

1 Compute all the other quantities in Table 16.1. In particular show that the internal voltage and current (V_k^Δ, I_k^Δ) of the impedance depends only on E_j^Δ, but not on (γ_j, β_j).
2 Show that $I_j^\Delta - I_k^\Delta = \delta \mathbf{1}$ for some $\delta \in \mathbb{C}$ when $\mathbf{1}\mathbf{1}^\mathsf{T} Z_{\text{Th}}^{-1} E_j^\Delta$ is in span$(\mathbf{1})$ where $Z_{\text{Th}} := \Gamma z_{jk}^s \Gamma^\mathsf{T} + z_k^\Delta$.
3 Show that $\gamma_k = \gamma_j$ when the three-phase line is symmetric of the form in (15.9) with $z_{jk}^1 + 2z_{jk}^2 \neq 0$.
4 In deriving the impedance model (14.27b), we have shown that its internal variable β_k and terminal current I_k must satisfy $\beta_k = \frac{1}{\zeta_k}\left(\tilde{z}_k^{\Delta\mathsf{T}} \Gamma^{\mathsf{T}\dagger}\right) I_k$, where $\tilde{z}_k^\Delta := z_k^\Delta \mathbf{1}$ and $\zeta_k := \mathbf{1}^\mathsf{T} z_k^\Delta \mathbf{1}$. Verify this expressions using the answer to part 1.

Solution. We will derive the quantities in the following order: $E_j^\Delta \Rightarrow I_k^\Delta, V_k^\Delta \Rightarrow \beta_k, I_k, I_j$. Then $E_j^\Delta, \gamma_j \Rightarrow V_j, V_k, \gamma_k \Rightarrow I_j^\Delta$.

The current balance equation (16.5) with $y_{jk}^m = y_{kj}^m = 0$ is:

$$V_k \;=\; V_j - z_{jk}^s I_j$$

Multiplying both sides by Γ and substituting the conversion rule $V_k^\Delta = \Gamma V_k$, $E_j^\Delta = \Gamma V_j$, and $I_j = -I_k$, we have

$$V_k^\Delta \;=\; \Gamma V_k \;=\; E_j^\Delta + \Gamma z_{jk}^s I_k \tag{16.17}$$

Substitute the internal model $V_k^\Delta = z_k^\Delta I_k^\Delta$ of impedance and the conversion rule $I_k = -\Gamma^\mathsf{T} I_k^\Delta$ to get

$$\left(\Gamma z_{jk}^s \Gamma^\mathsf{T} + z_k^\Delta\right) I_k^\Delta \;=\; E_j^\Delta \tag{16.18}$$

Hence

$$I_k^\Delta \;=\; Z_{\text{Th}}^{-1} E_j^\Delta, \qquad\qquad V_k^\Delta \;=\; z_k^\Delta Z_{\text{Th}}^{-1} E_j^\Delta$$

where $Z_{\text{Th}} := \Gamma z_{jk}^s \Gamma^\mathsf{T} + z_k^\Delta$ is the Thévenin equivalent of the three-phase line and the three-phase impedance. The expression for V_k^Δ is the three-phase version of the voltage

divider rule. Note that the internal variables $(V_k^\Delta, I_k^\Delta, \beta_j)$ of the impedance does not depend on γ_j.

We now calculate the other variables $\left(V_j, I_j, I_j^\Delta\right)$ and $(V_k, I_k, \gamma_k, \beta_k)$. The zero-sequence current and the terminal current of the impedance are

$$I_k \;=\; -\Gamma^\mathsf{T} I_k^\Delta \;=\; -\Gamma^\mathsf{T} Z_{\text{Th}}^{-1} E_j^\Delta, \qquad \beta_k \;:=\; \frac{1}{3}\mathbf{1}^\mathsf{T} I_k^\Delta \;=\; \frac{1}{3}\mathbf{1}^\mathsf{T} Z_{\text{Th}}^{-1} E_j^\Delta$$

Using the external model of an ideal voltage source from Table 14.4 we have

$$V_j \;=\; \frac{1}{3}\Gamma^\mathsf{T} E_j^\Delta + \gamma_j \mathbf{1}, \qquad I_j \;=\; -I_k \;=\; \Gamma^\mathsf{T} Z_{\text{Th}}^{-1} E_j^\Delta \tag{16.19}$$

Hence

$$\begin{aligned} V_k \;&=\; V_j - z_{jk}^s I_j \;=\; \left(\frac{1}{3}\Gamma^\mathsf{T} - z_{jk}^s \Gamma^\mathsf{T} Z_{\text{Th}}^{-1}\right) E_j^\Delta + \gamma_j \mathbf{1} \\ \gamma_k \;&=\; \frac{1}{3}\mathbf{1}^\mathsf{T} V_k \;=\; \gamma_j - \frac{1}{3}\left(\mathbf{1}^\mathsf{T} z_{jk}^s \Gamma^\mathsf{T}\right) Z_{\text{Th}}^{-1} E_j^\Delta \end{aligned}$$

Since $-\Gamma^\mathsf{T} I_j^\Delta = I_j = \Gamma^\mathsf{T} Z_{\text{Th}}^{-1} E_j^\Delta$ from (16.19) we have $\Gamma^\mathsf{T}\left(I_j^\Delta + Z_{\text{Th}}^{-1} E_j^\Delta\right) = 0$. Therefore (since the null space of Γ^T is span$(\mathbf{1})$)

$$I_j^\Delta \;=\; -Z_{\text{Th}}^{-1} E_j^\Delta + \beta_j' \mathbf{1}$$

where $\beta_j' \in \mathbb{C}$ is related to the given $\beta_j := \frac{1}{3}\mathbf{1}^\mathsf{T} I_j^\Delta$ by $\beta_j' = \beta_j + \frac{1}{3}\mathbf{1}^\mathsf{T} Z_{\text{Th}}^{-1} E_j^\Delta$. Hence[4]

$$I_j^\Delta \;=\; -Z_{\text{Th}}^{-1} E_j^\Delta + \left(\frac{1}{3}\mathbf{1}^\mathsf{T} Z_{\text{Th}}^{-1} E_j^\Delta + \beta_j\right)\mathbf{1}$$

From the derivation above, $\gamma_k = \gamma_j$ if $\mathbf{1}^\mathsf{T} z_{jk}^s \Gamma^\mathsf{T} Z_{\text{Th}}^{-1} E_j^\Delta = 0$. When the line is symmetric of the form in (15.9) we have

$$\mathbf{1}^\mathsf{T} z_{jk}^s \;=\; \mathbf{1}^\mathsf{T} \begin{bmatrix} z_{jk}^1 & z_{jk}^2 & z_{jk}^2 \\ z_{jk}^2 & z_{jk}^1 & z_{jk}^2 \\ z_{jk}^2 & z_{jk}^2 & z_{jk}^1 \end{bmatrix} \;=\; \left(z_{jk}^1 + 2z_{jk}^2\right)\mathbf{1}^\mathsf{T}$$

Hence (since $z_{jk}^1 + 2z_{jk}^2 \neq 0$)

$$\mathbf{1}^\mathsf{T} z_{jk}^s \Gamma^\mathsf{T} \;=\; \left(z_{jk}^1 + 2z_{jk}^2\right)\left(\mathbf{1}^\mathsf{T}\Gamma^\mathsf{T}\right) \;=\; 0$$

Finally we verify that the expressions $\beta_k = \frac{1}{3}\mathbf{1}^\mathsf{T} Z_{\text{Th}}^{-1} E_j^\Delta$ and $I_k = -\Gamma^\mathsf{T} Z_{\text{Th}}^{-1} E_j^\Delta$ satisfy

[4] Alternative derivation is: $-\Gamma^\mathsf{T} I_j^\Delta = \Gamma^\mathsf{T} Z_{\text{Th}}^{-1} E_j^\Delta$ implies

$$I_j^\Delta \;=\; -\frac{1}{3}\Gamma\Gamma^\mathsf{T} Z_{\text{Th}}^{-1} E_j^\Delta + \beta_j \mathbf{1} \;=\; -Z_{\text{Th}}^{-1} E_j^\Delta + \frac{1}{3}\mathbf{1}\mathbf{1}^\mathsf{T} Z_{\text{Th}}^{-1} E_j^\Delta + \beta_j \mathbf{1}$$

where the last equality follows from $\frac{1}{3}\Gamma\Gamma^\mathsf{T} = \mathbb{I} - \frac{1}{3}\mathbf{1}\mathbf{1}^\mathsf{T}$ by Theorem 14.2.

$\beta_k = \frac{1}{\zeta_k}\left(\tilde{z}_k^{\Delta\mathsf{T}}\Gamma^{\mathsf{T}\dagger}\right) I_k$ where $\tilde{z}_k^{\Delta} := z_k^{\Delta}\mathbf{1}$ and $\zeta_k := \mathbf{1}^{\mathsf{T}} z_k^{\Delta}\mathbf{1}$. We have

$$\left(\tilde{z}_k^{\Delta\mathsf{T}}\Gamma^{\mathsf{T}\dagger}\right) I_k \;=\; -\tilde{z}_k^{\Delta\mathsf{T}}\left(\Gamma^{\mathsf{T}\dagger}\Gamma^{\mathsf{T}}\right) Z_{\mathrm{Th}}^{-1} E_j^{\Delta} \;=\; -\tilde{z}_k^{\Delta\mathsf{T}}\left(\mathbb{I} - \frac{1}{3}\mathbf{1}\mathbf{1}^{\mathsf{T}}\right) Z_{\mathrm{Th}}^{-1} E_j^{\Delta} \;=\; -\tilde{z}_k^{\Delta\mathsf{T}} Z_{\mathrm{Th}}^{-1} E_j^{\Delta} + \frac{\zeta_k}{3}\mathbf{1}^{\mathsf{T}} Z_{\mathrm{Th}}^{-1} E_j^{\Delta}$$

where the second equality follows from Theorem 14.2. But

$$\tilde{z}_k^{\Delta\mathsf{T}} Z_{\mathrm{Th}}^{-1} E_j^{\Delta} \;=\; \mathbf{1}^{\mathsf{T}} z_k^{\Delta} Z_{\mathrm{Th}}^{-1} E_j^{\Delta} \;=\; \mathbf{1}^{\mathsf{T}} V_k^{\Delta} \;=\; 0$$

where the last equality follows from (16.17). Hence $\left(\tilde{z}_k^{\Delta\mathsf{T}}\Gamma^{\mathsf{T}\dagger}\right) I_k = \zeta_k\beta_k$ as desired. □

Voltage source j			
V_j^{Δ}	given E_j^{Δ}	V_k^{Δ}	$z_k^{\Delta} I_k^{\Delta} = z_k^{\Delta} Z_{\mathrm{Th}}^{-1} E_j^{\Delta}$
I_j^{Δ}	$-\left(\Gamma\Gamma^{\dagger}\right) Z_{\mathrm{Th}}^{-1} E_j^{\Delta} + \beta_j\mathbf{1}$	I_k^{Δ}	$Z_{\mathrm{Th}}^{-1} E_j^{\Delta}$
β_j	given	β_k	$\frac{1}{3}\mathbf{1}^{\mathsf{T}} I_k^{\Delta} = \frac{1}{3}\mathbf{1}^{\mathsf{T}} Z_{\mathrm{Th}}^{-1} E_j^{\Delta}$
V_j	$\Gamma^{\dagger} E_j^{\Delta} + \gamma_j$	V_k	$\left(\frac{1}{3}\Gamma^{\mathsf{T}} - z_{jk}^{s}\Gamma^{\mathsf{T}} Z_{\mathrm{Th}}^{-1}\right) E_j^{\Delta} + \gamma_j\mathbf{1}$
I_j	$-I_k = \Gamma^{\mathsf{T}} Z_{\mathrm{Th}}^{-1} E_j^{\Delta}$	I_k	$-\Gamma^{\mathsf{T}} I_k^{\Delta} = -\Gamma^{\mathsf{T}} Z_{\mathrm{Th}}^{-1} E_j^{\Delta}$
γ_j	given	γ_k	$\gamma_j - \frac{1}{3}\left(\mathbf{1}^{\mathsf{T}} z_{jk}^{s}\Gamma^{\mathsf{T}}\right) Z_{\mathrm{Th}}^{-1} E_j^{\Delta}$

Table 16.1 Example 16.6: parameters and variables for a voltage source j where $Z_{\mathrm{Th}} := \Gamma z_{jk}^{s}\Gamma^{\mathsf{T}} + z_k^{\Delta}$.

Current source j			
V_j^{Δ}	$V_k^{\Delta} - \left(\Gamma z_{jk}^{s}\Gamma^{\mathsf{T}}\right) I_j^{\Delta}$	V_k^{Δ}	$z_k^{\Delta} I_k^{\Delta} \;=\; z_k^{\Delta} A(z_k^{\Delta}) J_j^{\Delta}$
I_j^{Δ}	given J_j^{Δ}	I_k^{Δ}	$A(z_k^{\Delta}) J_j^{\Delta}$
β_j	$\frac{1}{3}\mathbf{1}^{\mathsf{T}} J_j^{\Delta}$	β_k	$\left(\frac{\tilde{z}_k^{\Delta}}{\zeta_k} - \frac{\mathbf{1}}{3}\right)^{\mathsf{T}} J_j^{\Delta}$
V_j	$V_k + z_{jk}^{s} I_j = V_k - z_{jk}^{s}\Gamma^{\mathsf{T}} J_j^{\Delta}$	V_k	$\frac{1}{3}\Gamma^{\mathsf{T}} V_k^{\Delta} + \gamma_k\mathbf{1}$
I_j	$-\Gamma^{\mathsf{T}} J_j^{\Delta}$	I_k	$\Gamma^{\mathsf{T}} J_j^{\Delta}$
γ_j	given	γ_k	$\gamma_j + \frac{1}{3}\mathbf{1}^{\mathsf{T}} z_{jk}^{s}\Gamma^{\mathsf{T}} J_j^{\Delta}$

Table 16.2 Exercise 16.8: parameters and variables for a current source j where $\tilde{z}_k^{\Delta} := z_k^{\Delta}\mathbf{1}$, $\zeta_k := \mathbf{1}^{\mathsf{T}} z_k^{\Delta}\mathbf{1}$, and $A(z_k^{\Delta}) := \left(\frac{1}{\zeta_k}\mathbf{1}\tilde{z}_k^{\Delta\mathsf{T}} - \mathbb{I}\right)$.

Remark 16.3 (Comprison: voltage vs current sources)**.** In both Example 16.6 and Exercise 16.8, the key to the derivation is to first calculate the internal current I_k^{Δ} of the impedance by relating it to the given source parameter E_j^{Δ} or J_j^{Δ}. Given I_k^{Δ}, all other

variables can be derived. This insight will be used in Chapter 16.2.3 for analyzing a general three-phase problem.

Compare the results in Table 16.1 from Example 16.6 for the voltage source with the results in Table 16.2 from Exercise 16.8 for the current source.

1 The internal variables $\left(V_k^\Delta, I_k^\Delta, \beta_k\right)$ of the impedance do not depend on $\left(\gamma_j, \beta_j\right)$, but only on E_j^Δ for the voltage source and I_j^Δ for the current source.
2 For the current source, $I_k^\Delta = A(z_k^\Delta) J_j^\Delta$ depends only on the impedance z_k^Δ but not on the line series admittance y_{jk}^s. This is because of the assumption $z_{jk}^m = z_{kj}^m = 0$. For the voltage source, $I_k^\Delta = Z_{\text{Th}}^{-1} E_j^\Delta$ depends on both z_{jk}^s and z_k^Δ through their Thévenin equivalent. Their values are equal if $E_j^\Delta = Z_{\text{Th}} A\left(z_k^\Delta\right) J_j^\Delta$.
3 For both the voltage and current source, $\gamma_k = \gamma_j$ if z_{jk}^s is symmetric.
4 For the current source, the loop flows β_j and β_k are related as follows (see Exercise 16.8):
   - $\beta_k = -\beta_j$ if and only if $z_k^{ab} J_j^{ab} + z_k^{bc} J_j^{bc} + z_k^{ca} J_j^{ca} = 0$.
   - $\beta_k = 0$ if and only if $z_k^{ab} J_j^{ab} + z_k^{bc} J_j^{bc} + z_k^{ca} J_j^{ca} = \zeta_k \beta_j$.
   - $\beta_k = 0$ if the impedance $z_k^\Delta = \frac{\zeta_k}{3}\mathbb{I}$ is balanced, regardless of whether J_j^Δ is balanced or whether β_j is zero. The converse does not necessarily hold.

□

Example 16.7 (Balanced system)**.** Assume the system in Example 16.6 is a balanced system, i.e., given

- The voltage source parameters $\left(E_j^\Delta, \gamma_j, \beta_j\right)$ with $E_j^\Delta := \lambda_j \alpha_+$ where $\lambda_j \in \mathbb{C}$, $\alpha_+ := (1, \alpha, \alpha^2)$, and $\alpha := e^{-\mathbf{i}2\pi/3}$,
- The impedance $z_k^\Delta := \zeta_k' \mathbb{I}$ where $\zeta_k' \in \mathbb{C}$.
- Line admittances $\left(\left(z_{jk}^s\right)^{-1}, y_{jk}^m = y_{kj}^m := 0\right)$ with $z_{jk}^s = \zeta_{jk}\mathbb{I}$, i.e., the phases are decoupled.

1 Show that $Z_{\text{Th}} = \zeta_k'\mathbb{I} + \zeta_{jk}\Gamma\Gamma^\mathsf{T}$ and $Z_{\text{Th}}^{-1} = a\left(\mathbb{I} - \frac{a\zeta_{jk}}{3a\zeta_{jk}-1}\mathbf{1}\mathbf{1}^\mathsf{T}\right)$ where $a := 1/(\zeta_k' + 3\zeta_{jk})$.
2 Show that all variables (V_j, V_k, I_j, I_k), $\left(V_k^\Delta, I_k^\Delta\right)$ are balanced positive-sequence sets.

Solution. By definition

$$Z_{\text{Th}} := z_k^\Delta + \Gamma z_{jk}^s \Gamma^\mathsf{T} = \zeta_k'\mathbb{I} + \zeta_{jk}\Gamma\Gamma^\mathsf{T}$$

Substituting $\Gamma\Gamma^\mathsf{T} = 3\mathbb{I} - \mathbf{1}\mathbf{1}^\mathsf{T}$ from Theorem 14.2 we have $Z_{\text{Th}} = (1/a)\left(\mathbb{I} - a\zeta_{jk}\mathbf{1}\mathbf{1}^\mathsf{T}\right)$.

Apply the matrix inversion formula (A.6) in Appendix A.3: given a scalar $c \in \mathbb{C}$, vectors $b, d \in \mathbb{C}^n$, and the identity matrix $\mathbb{I}_n$ of size n,

$$\left(\mathbb{I}_n + bcd^{\mathsf{T}}\right)^{-1} = \mathbb{I}_n - b\left(c^{-1} + d^{\mathsf{T}}b\right)^{-1} d^{\mathsf{T}}$$

we therefore have (with $c := -a\zeta_{jk}$, $b = d = \mathbf{1}$)

$$Z_{\text{Th}}^{-1} = a\left(\mathbb{I} - \frac{a\zeta_{jk}}{3a\zeta_{jk} - 1}\mathbf{1}\mathbf{1}^{\mathsf{T}}\right) \tag{16.20}$$

To show that all voltages and currents are balanced positive-sequence sets, i.e., in span(α_+), the key property that we will use is Corollary 1.3 which states that: For any balanced positive-sequence vector $x + a\mathbf{1} \in \mathbb{C}^3$ with $a \in \mathbb{C}$, we have

$$\Gamma(x + a\mathbf{1}) = (1-\alpha)x, \qquad \Gamma^{\mathsf{T}}(x + a\mathbf{1}) = (1-\alpha^2)x$$

We have from Table 16.1 (substituting $E_j^\Delta = \lambda_j \mathbb{I}$ and $z_k^\Delta = \zeta_k' \mathbb{I}$)

$$I_k^\Delta = Z_{\text{Th}}^{-1} E_j^\Delta = a\lambda_j\left(\mathbb{I} - \frac{a\zeta_{jk}}{3a\zeta_{jk} - 1}\mathbf{1}\mathbf{1}^{\mathsf{T}}\right)\alpha_+ = \frac{\lambda_j}{\zeta_k' + 3\zeta_{jk}}\alpha_+$$

$$V_k^\Delta = z_k^\Delta I_k^\Delta = \frac{\zeta_k'}{\zeta_k' + 3\zeta_{jk}}\lambda_j\,\alpha_+, \qquad \beta_k := \frac{1}{3}\mathbf{1}^{\mathsf{T}} I_k^\Delta = 0$$

where we have used $\mathbf{1}^{\mathsf{T}}\alpha_+ = 0$. The expression for V_k^Δ is the voltage divider rule.

We now calculate the other variables $\left(V_j, I_j, I_j^\Delta\right)$ and (V_k, I_k, γ_k). The terminal current of the impedance are

$$I_k = -\Gamma^{\mathsf{T}} I_k^\Delta = -\frac{\lambda_j}{\zeta_k' + 3\zeta_{jk}}\Gamma^{\mathsf{T}}\alpha_+ = -\frac{(1-\alpha^2)\lambda_j}{\zeta_k' + 3\zeta_{jk}}\alpha_+$$

Using the external model of an ideal voltage source from Table 14.4 we have

$$V_j = \frac{1}{3}\Gamma^{\mathsf{T}} E_j^\Delta + \gamma_j\mathbf{1} = \frac{1}{3}(1-\alpha^2)\lambda_j\,\alpha_+ + \gamma_j\mathbf{1}$$

$$I_j = -I_k = \frac{(1-\alpha^2)\lambda_j}{\zeta_k' + 3\zeta_{jk}}\alpha_+$$

Hence

$$V_k = V_j - z_{jk}^s I_j = \frac{(1-\alpha^2)\zeta_k'}{3\left(\zeta_k' + 3\zeta_{jk}\right)}\lambda_j\,\alpha_+ + \gamma_j\mathbf{1}, \qquad \gamma_k = \frac{1}{3}\mathbf{1}^{\mathsf{T}} V_k = \gamma_j$$

Finally

$$I_j^\Delta = -\frac{1}{3}\Gamma I_j + \beta_j\mathbf{1} = -\frac{(1-\alpha)(1-\alpha^2)\lambda_j}{3\left(\zeta_k' + 3\zeta_{jk}\right)}\alpha_+ + \beta_j\mathbf{1}$$

□

With power sources.

The solution strategy is the same as that for problems without power sources with the addition of quadratic device models of power sources. Specifically we first relate *internal* voltages and currents to power sources $(\sigma_j^\Delta, \gamma_j)$ to obtain a system of quadratic equations that can be solved numerically. Then all other voltages and currents can be obtained analytically in terms of a solution of the quadratic equations. Finally we can calculate internal and external power using $s_j^{Y/\Delta} := \operatorname{diag}\left(V_j^{Y/\Delta} I_j^{Y/\Delta \mathsf{H}}\right)$ and $s_j := \operatorname{diag}\left(V_j I_j^{\mathsf{H}}\right)$ respectively. This solution strategy is extended in Chapter 16.2.3 to general three-phase networks.

Example 16.8 (Power source)**.** Consider the system in Figure 16.7 where, instead of a voltage source, the generator is a three-phase power source. Suppose the following are specified:

- Power source $\left(\sigma_j^\Delta, \gamma_j\right)$.
- Impedance z_k^Δ. (Note that β_k needs not be specified for an impedance and can be derived.)
- Line admittances $\left(y_{jk}^s, y_{jk}^m, y_{kj}^m\right)$ with nonzero y_{jk}^m and y_{kj}^m. In particular assumption C16.1 is satisfied.

Find all remaining internal and external variables $\left(V_i^\Delta, I_i^\Delta, s_i^\Delta, \beta_j\right)$ and $(V_i, I_i, s_i, \gamma_k)$, $i = j, k$.

Solution. The current balance equation $I = YV$, the internal models of the power source and impedance, and the conversion rules are:

$$\begin{bmatrix} I_j \\ I_k \end{bmatrix} = \begin{bmatrix} y_{jk}^s + y_{jk}^m & -y_{jk}^s \\ -y_{jk}^s & y_{jk}^s + y_{kj}^m \end{bmatrix} \begin{bmatrix} V_j \\ V_k \end{bmatrix} \tag{16.21a}$$

$$\sigma_j^\Delta = \operatorname{diag}\left(V_j^\Delta I_j^{\Delta \mathsf{H}}\right), \qquad V_k^\Delta = z_k^\Delta I_k^\Delta \tag{16.21b}$$

$$\Gamma V_i = V_i^\Delta, \qquad I_i = -\Gamma^{\mathsf{T}} I_i^\Delta, \qquad i = j, k \tag{16.21c}$$

Assuming the admittance matrix Y is invertible (e.g., it satisfies the condition in Theorem 4.3), denote its inverse by

$$Y^{-1} := \begin{bmatrix} y_{jk}^s + y_{jk}^m & -y_{jk}^s \\ -y_{jk}^s & y_{jk}^s + y_{kj}^m \end{bmatrix}^{-1} = \begin{bmatrix} z_{jj} & z_{jk} \\ z_{kj} & z_{kk} \end{bmatrix}$$

We can then relate the internal variables $\left(V_i^\Delta, I_i^\Delta\right)$, $i = j, k$, by eliminating the external variables to get

$$\begin{aligned} \begin{bmatrix} V_j^\Delta \\ V_k^\Delta \end{bmatrix} &= -\operatorname{diag}(\Gamma, \Gamma) \begin{bmatrix} z_{jj} & z_{jk} \\ z_{kj} & z_{kk} \end{bmatrix} \operatorname{diag}\left(\Gamma^{\mathsf{T}}, \Gamma^{\mathsf{T}}\right) \begin{bmatrix} I_j^\Delta \\ I_k^\Delta \end{bmatrix} \\ \sigma_j^\Delta &= \operatorname{diag}\left(V_j^\Delta I_j^{\Delta \mathsf{H}}\right) \end{aligned} \tag{16.22}$$

Eliminating V_k^Δ using $V_k^\Delta = z_k^\Delta I_k^\Delta$ and re-arranging, we get

$$\begin{bmatrix} Z_{jj} & Z_{jk} & \mathbb{I} \\ Z_{kj} & Z_{kk}+z_k^\Delta & 0 \end{bmatrix} \begin{bmatrix} I_j^\Delta \\ I_k^\Delta \\ V_j^\Delta \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{16.23a}$$

$$\operatorname{diag}\left(V_j^\Delta I_j^{\Delta\mathsf{H}}\right) = \sigma_j^\Delta \tag{16.23b}$$

where $Z_{jj} := \Gamma z_{jj} \Gamma^\mathsf{T}$ and so on. This is a system of 9 quadratic equations in 9 variables $\left(V_j^\Delta, I_j^\Delta, I_k^\Delta\right)$. It can be solved numerically. All other variables can then be derived analytically in terms of a solution $\left(V_j^\Delta, I_j^\Delta, I_k^\Delta\right)$.

We can further reduce (16.23) by eliminating V_j^Δ and I_k^Δ to get a quadratic equation in I_j^Δ:

$$\operatorname{diag}\left(\left(-Z_{jj} + Z_{jk}\left(Z_{kk}+z_k^\Delta\right)^{-1} Z_{kj}\right) I_j^\Delta I_j^{\Delta\mathsf{H}}\right) = \sigma_j^\Delta, \qquad j \in \overline{N} \tag{16.24}$$

In summary we can first solve (16.24) numerically to obtain I_j^Δ and then derive all other variables, or first solve (16.23) numerically to obtain $\left(V_j^\Delta, I_j^\Delta, I_k^\Delta\right)$ and then all other variables. They are equivalent to solving the original system (16.21) numerically. The decentralized structure of (16.24) is quite striking: the system of power flow equations for the entire network reduces to this quadratic equation separately for each bus j that can be solved in parallel.

We now derive all other variables from I_j^Δ, by tracing back the derivation of (16.24). From (16.23a) we have

$$I_k^\Delta = -\left(Z_{kk}+z_k^\Delta\right)^{-1} Z_{kj} I_j^\Delta, \quad V_j^\Delta = -Z_{jj} I_j^\Delta - Z_{jk} I_k^\Delta = \left(-Z_{jj} + Z_{jk}\left(Z_{kk}+z_k^\Delta\right)^{-1} Z_{kj}\right) I_j^\Delta$$

From (16.21b) we have

$$V_k^\Delta = z_k^\Delta I_k^\Delta = -z_k^\Delta \left(Z_{kk}+z_k^\Delta\right)^{-1} Z_{kj} I_j^\Delta,$$

The internal zero-sequence currents are given by

$$\beta_j = \frac{1}{3}\mathbf{1}^\mathsf{T} I_j^\Delta, \qquad \beta_k = \frac{1}{3}\mathbf{1}^\mathsf{T} I_k^\Delta$$

This completes the derivation of internal voltages and currents.

The terminal currents can be obtained from the conversion rule (16.21c):

$$I_j = -\Gamma^\mathsf{T} I_j^\Delta, \qquad I_k = -\Gamma^\mathsf{T} I_k^\Delta = \Gamma^\mathsf{T}\left(Z_{kk}+z_k^\Delta\right)^{-1} Z_{kj} I_j^\Delta$$

Note that $\mathbf{1}^\mathsf{T} V_j^\Delta = \mathbf{1}^\mathsf{T} V_k^\Delta = 0$ from (16.22). Hence the conversion rule (16.21c) yields (recall that γ_j is specified)

$$V_j = \frac{1}{3} V_j^\Delta + \gamma_j \mathbf{1} \tag{16.25a}$$

Given the terminal voltage V_j of the power source, (V_k, γ_k) of the impedance can then be determined through the network equation (16.21a):

$$V_k \;=\; \left(y^s_{jk} + y^m_{kj}\right)^{-1} \left(y^s_{jk} V_j + I_k\right), \qquad \gamma_k \;=\; \frac{1}{3}\mathbf{1}^{\mathsf{T}} V_k \tag{16.25b}$$

Notice that the zero-sequence voltage γ_j of the power source uniquely determines γ_k of the impedance. □

The derivation in Example 16.8 relies on the assumption that the admittance matrix Y in (16.21a) is invertible. If the shunt admittances $y^m_{jk} = y^m_{kj} = 0$ then Y has zero block row sums (Definition 16.1), i.e., $\sum_k Y_{jk} = 0$ for all j. This implies that Y has zero row sums, i.e., $\sum_{k,\phi'} Y_{j\phi,k\phi'} = 0$ for all $j\phi$, and is therefore singular. In that case, additional information needs to be specified to obtain a unique solution, as the next example illustrates.

Example 16.9 (Power source)**.** Repeat Example 16.8 but with zero shunt admittances and given zero-sequence currents, i.e., suppose the following are specified:

- Power source $\left(\sigma^\Delta_j, \gamma_j\right)$.
- Impedance z^Δ_k.
- Line admittances $\left(y^s_{jk}, y^m_{jk} = y^m_{kj} = 0\right)$ with nonsingular y^s_{jk}. In particular assumption C16.1 is satisfied.
- $\beta_j + \beta_k := \frac{1}{3}\mathbf{1}^{\mathsf{T}}\left(I^\Delta_j + I^\Delta_k\right) = \beta'$.

Solution. When $y^m_{jk} = y^m_{kj} = 0$ the network equation (16.21a) reduces to

$$I_j \;=\; -I_k \;=\; y^s_{jk}\left(V_j - V_k\right) \tag{16.26}$$

Hence $\Gamma^{\mathsf{T}}\left(I^\Delta_j + I^\Delta_k\right) = 0$ from (16.21c), implying that

$$I^\Delta_j + I^\Delta_k \;=\; \left(\beta_j + \beta_k\right)\mathbf{1} \;=\; \beta'\mathbf{1} \tag{16.27}$$

with β' a given quantity. We will express V^Δ_j in terms of I^Δ_j in order to write $\sigma_j = \operatorname{diag}\left(V^\Delta_j I^{\Delta\mathsf{H}}_j\right)$ as a quadratic equation in I^Δ_j.

Multiplying both sides of (16.26) by $z^s_{jk} := \left(y^s_{jk}\right)^{-1}$ and using the conversion rule again (16.21b)(16.21c), we have

$$V^\Delta_j \;=\; \left(\Gamma z^s_{jk}\Gamma^{\mathsf{T}} + z^\Delta_k\right) I^\Delta_k \;=\; Z^\Delta_{jk}\left(-I^\Delta_j + \beta'\mathbf{1}\right) \;=\; -Z^\Delta_{jk} I^\Delta_j + \beta' \tilde{z}^\Delta_k \tag{16.28}$$

where the second equality follows from (16.27), $Z^\Delta_{jk} := \Gamma z^s_{jk}\Gamma^{\mathsf{T}} + z^\Delta_k$, and $\tilde{z}^\Delta_k := z^\Delta_k\mathbf{1}$. Hence we have

$$\sigma^\Delta_j \;=\; \operatorname{diag}\left(V^\Delta_j I^{\Delta\mathsf{H}}_j\right) \;=\; \operatorname{diag}\left(-Z^\Delta_{jk} I^\Delta_j I^{\Delta\mathsf{H}}_j + \beta' \tilde{z}^\Delta_k I^{\Delta\mathsf{H}}_j\right) \tag{16.29}$$

This is a system of three quadratic equations in three variables $I_j^\Delta \in \mathbb{C}^3$. Assume a solution exists and can be obtained by solving (16.29) numerically.

Given a solution I_j^Δ of (16.29), all other variables can be derived analytically in terms of I_j^Δ by tracing back the derivation of (16.29), similar to the derivation in Example 16.8. Specifically we have $I_j = -\Gamma^\mathsf{T} I_j^\Delta$ and $\beta_j := \frac{1}{3}\mathbf{1}^\mathsf{T} I_j^\Delta$. We obtain V_j^Δ from (16.28), from which we have $V_j = \frac{1}{3}\Gamma^\mathsf{T} V_j^\Delta + \gamma_j \mathbf{1}$. This computes all voltages and currents of the power source j.

The network equation (16.26) then yields $V_k = V_j - z_{jk}^s I_j$ and hence also $\gamma_k := \frac{1}{3}\mathbf{1}^\mathsf{T} V_k$. We also have $I_k = -I_j = \Gamma^\mathsf{T} I_j^\Delta$, $\beta_k = \beta' - \beta_j$, and hence $I_k^\Delta = -\frac{1}{3}\Gamma I_k + \beta_k \mathbf{1}$ and $V_k^\Delta = z_k^\Delta I_k^\Delta$. This computes all voltages and currents of the impedance k. □

The next example shows that if the power source and the impedance are balanced and the line is decoupled and balanced, then all voltages, currents, and powers will be generalized balanced vectors. This will be proved for general networks in Chapter 16.3. Furthermore the given power σ_j^Δ cannot be arbitrary but must be consistent with other parameters of the network such as line and device impedances, e.g., from (16.33), b_j/c must be real. This generalizes the single-phase case where a power source s supplies an impedance load z with a current i. Then $s = z|i|^2$ implying that s/z is a read number. This is because $\angle z = \angle s$ fixes the phase difference between the voltage v and current i across the impedance.

Example 16.10 (Balanced power source)**.** Repeat Example 16.8 when the system is balanced, i.e.,

- Power source $\left(\sigma_j^\Delta, \gamma_j\right)$ with $\sigma_j^\Delta = a_j \alpha_+ + b_j \mathbf{1}$ for given $\left(a_j, b_j\right)$, i.e., a balanced power source must be a generalized balanced vector. Moreover its voltage and current $\left(V_j^\Delta, I_j^\Delta\right)$ are generalized balanced vectors.
- Impedance $z_k^\Delta := \zeta_k^\Delta \mathbb{I}$.
- Line admittances $\left(y_{jk}^s, y_{jk}^m, y_{kj}^m\right) := \left(\eta_{jk}^s \mathbb{I}, \eta_{jk}^m \mathbb{I}, \eta_{kj}^m \mathbb{I}\right)$ with nonzero η_{jk}^s, η_{jk}^m and η_{kj}^m.

Find all remaining internal and external variables $\left(V_i^\Delta, I_i^\Delta, s_i^\Delta, \beta_i^\Delta\right)$ and $(V_i, I_i, s_i, \gamma_i)$, $i = j, k$. Show that the problem can be solved *analytically* when a reference angle is given, say, $\angle V_j^a := \theta_j^a$.

Solution. Let (recall that $\mathbf{1}^\mathsf{T} V_j^\Delta = 0$)

$$V_j^\Delta \;=:\; v_j^\Delta \alpha_+, \qquad\qquad I_j^\Delta \;=:\; i_j^\Delta \alpha_+ + \beta_j \mathbf{1} \tag{16.30}$$

giving (noting $\operatorname{diag}\left(\alpha_+ \alpha_+^\mathsf{H}\right) = \mathbf{1}$)

$$\sigma_j^\Delta \;=\; \operatorname{diag}\left(v_j^\Delta \alpha_+ \left(i_j^\Delta \alpha_+ + \beta_j \mathbf{1}\right)^\mathsf{H}\right) \;=\; \left(v_j^\Delta \overline{\beta}_j\right)\alpha_+ \;+\; \left(v_j^\Delta \overline{i}_j^\Delta\right)\mathbf{1}$$

where $\left(v_j^\Delta, i_j^\Delta, \beta_j \in \mathbb{C}^3\right)$ are to be determined. Recall that $\overline{x}$ denotes the complex conjugate of any $x \in \mathbb{C}$. Therefore, since $\sigma_j^\Delta = a_j \alpha_+ + b_j \mathbf{1}$,

$$v_j^\Delta \overline{\beta}_j \;=\; a_j, \qquad\qquad v_j^\Delta \overline{i}_j^\Delta \;=\; b_j \tag{16.31}$$

which are two quadratic equations in unknowns $\left(v_j^\Delta, i_j^\Delta, \beta_j\right) \in \mathbb{C}^3$. Note that the internal power σ_j^Δ is different in each phase (with different phase angles separated by 120°) if and only if $\beta_j \neq 0$.

We will solve this problem by substituting the given balanced system parameters into the solution of Example 16.8.

Specifically the admittance matrix is

$$Y \;:=\; \begin{bmatrix} y_{jk}^s + y_{jk}^m & -y_{jk}^s \\ -y_{jk}^s & y_{jk}^s + y_{kj}^m \end{bmatrix} \;=\; \underbrace{\begin{bmatrix} \eta_{jk}^s + \eta_{jk}^m & -\eta_{jk}^s \\ -\eta_{jk}^s & \eta_{jk}^s + y_{kj}^m \end{bmatrix}}_{Y^{1\phi}} \otimes \mathbb{I}$$

Assuming the 2×2 admittance matrix $Y^{1\phi}$ is invertible with inverse $\left(Y^{1\phi}\right)^{-1} =: \begin{bmatrix} \zeta_{jj} & \zeta_{jk} \\ \zeta_{kj} & \zeta_{kk} \end{bmatrix}$ we have

$$Y^{-1} \;=\; \left(Y^{1\phi}\right)^{-1} \otimes \mathbb{I} \;=:\; \begin{bmatrix} \zeta_{jj} & \zeta_{jk} \\ \zeta_{kj} & \zeta_{kk} \end{bmatrix} \otimes \mathbb{I}$$

where the first equality follows from $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$ in Lemma 16.6. Then (16.22) becomes

$$\begin{aligned} \begin{bmatrix} V_j^\Delta \\ V_k^\Delta \end{bmatrix} \;&=\; -\operatorname{diag}(\Gamma, \Gamma) \left(\begin{bmatrix} \zeta_{jj} & \zeta_{jk} \\ \zeta_{kj} & \zeta_{kk} \end{bmatrix} \otimes \mathbb{I} \right) \operatorname{diag}\left(\Gamma^\mathsf{T}, \Gamma^\mathsf{T}\right) \begin{bmatrix} I_j^\Delta \\ I_k^\Delta \end{bmatrix} \\ &=\; \begin{bmatrix} \zeta_{jj} & \zeta_{jk} \\ \zeta_{kj} & \zeta_{kk} \end{bmatrix} \otimes \left(\Gamma \Gamma^\mathsf{T}\right) \begin{bmatrix} I_j^\Delta \\ I_k^\Delta \end{bmatrix} \end{aligned}$$

where $\Gamma\Gamma^\mathsf{T} = 3\mathbb{I} - \mathbf{1}\mathbf{1}^\mathsf{T}$ from Theorem 14.2. Then (16.23) becomes (16.31) together with

$$\begin{bmatrix} \zeta_{jj}\left(\Gamma\Gamma^\mathsf{T}\right) & \zeta_{jk}\left(\Gamma\Gamma^\mathsf{T}\right) & \mathbb{I} \\ \zeta_{kj}\left(\Gamma\Gamma^\mathsf{T}\right) & \zeta_{kk}\left(\Gamma\Gamma^\mathsf{T}\right) + \zeta_k^\Delta \mathbb{I} & 0 \end{bmatrix} \begin{bmatrix} i_j^\Delta \alpha_+ + \beta_j \mathbf{1} \\ I_k^\Delta \\ v_j^\Delta \alpha_+ \end{bmatrix} \;=\; \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

where we have used the specification (16.30). This is a system of 8 (redundant) quadratic equations that can be solved numerically for the 6 unknowns $\left(v_j^\Delta, i_j^\Delta, \beta_j\right) \in \mathbb{C}^3$ and $I_k^\Delta \in \mathbb{C}^3$. It implies that I_k^Δ is a generalized balanced vector of the form $I_k^\Delta = i_k^\Delta \alpha_+ + \beta_k \mathbf{1}$ for some $\left(i_k^\Delta, \beta_k\right)$.

To evaluate (16.24) we have

$$I_j^\Delta I_j^{\Delta\mathsf{H}} \;=\; \left(i_j^\Delta \alpha_+ + \beta_j \mathbf{1}\right)\left(i_j^\Delta \alpha_+ + \beta_j \mathbf{1}\right)^\mathsf{H} \;=\; \left|i_j^\Delta\right|^2 \alpha_+ \alpha_+^\mathsf{H} + i_j^\Delta \overline{\beta}_j \, \alpha_+ \mathbf{1}^\mathsf{T} + \overline{i}_j^\Delta \beta_j \, \mathbf{1} \alpha_+^\mathsf{H} + \left|\beta_j\right|^2 \mathbf{1}\mathbf{1}^\mathsf{T}$$

and therefore

$$\left(\Gamma\Gamma^{\mathsf{T}}\right) I_j^{\Delta} I_j^{\Delta\mathsf{H}} \;=\; 3\left(\left|i_j^{\Delta}\right|^2 \alpha_+\alpha_+^{\mathsf{H}} + i_j^{\Delta}\overline{\beta}_j\,\alpha_+\mathbf{1}^{\mathsf{T}}\right) \tag{16.32a}$$

where we have used $\Gamma\Gamma^{\mathsf{T}}\alpha_+ = 3\alpha_+$ from Corollary 1.3 and $\Gamma^{\mathsf{T}}\mathbf{1} = 0$. Furthermore

$$\left(Z_{kk}+z_k^{\Delta}\right)^{-1} \;=\; \left(\zeta_{kk}\left(\Gamma\Gamma^{\mathsf{T}}\right)+\zeta_k^{\Delta}\mathbb{I}\right)^{-1} \;=\; \left(\left(3\zeta_{kk}+\zeta_k^{\Delta}\right)\mathbb{I}-\zeta_{kk}\mathbf{1}\mathbf{1}^{\mathsf{T}}\right)^{-1} \;=\; \frac{1}{3\zeta_{kk}+\zeta_k^{\Delta}}\left(\mathbb{I}-\frac{\zeta_{kk}}{\zeta_k^{\Delta}}\mathbf{1}\mathbf{1}^{\mathsf{T}}\right)$$

where the last equality follows from the matrix inversion formula (see Appendix A.3.2)

$$(\mathbb{I}_n + BD)^{-1} = I_n \;-\; B\,(\mathbb{I}_k + DB)^{-1}\,D$$

when $B, D^{\mathsf{T}} \in \mathbb{C}^{n\times k}$ and $\mathbb{I}_n, \mathbb{I}_k$ denote identity matrices of sizes n, k respectively. Hence

$$Z_{jk}\left(Z_{kk}+z_k^{\Delta}\right)^{-1} Z_{kj} \;=\; \frac{\zeta_{jk}\zeta_{kj}}{3\zeta_{kk}+\zeta_k^{\Delta}}\left(\Gamma\Gamma^{\mathsf{T}}\right)\left(\mathbb{I}-\frac{\zeta_{kk}}{\zeta_k^{\Delta}}\mathbf{1}\mathbf{1}^{\mathsf{T}}\right)\left(\Gamma\Gamma^{\mathsf{T}}\right) \;=\; \frac{3\zeta_{jk}\zeta_{kj}}{3\zeta_{kk}+\zeta_k^{\Delta}}\,\Gamma\Gamma^{\mathsf{T}} \tag{16.32b}$$

Together with $Z_{jj} = \zeta_{jj}\Gamma\Gamma^{\mathsf{T}}$, (16.32) implies that (16.24) is

$$\begin{aligned}\sigma_j \;=\; a_j\alpha_+ + b_j\mathbf{1} \;&=\; \left(-\zeta_{jj}+\frac{3\zeta_{jk}\zeta_{kj}}{3\zeta_{kk}+\zeta_k^{\Delta}}\right)\operatorname{diag}\left(\Gamma\Gamma^{\mathsf{H}} I_j^{\Delta} I_j^{\Delta\mathsf{H}}\right)\\ &=\; \underbrace{3\left(-\zeta_{jj}+\frac{3\zeta_{jk}\zeta_{kj}}{3\zeta_{kk}+\zeta_k^{\Delta}}\right)}_{c}\left(i_j^{\Delta}\overline{\beta}_j\,\alpha_+ + \left|i_j^{\Delta}\right|^2\mathbf{1}\right)\end{aligned}$$

where we have used $\operatorname{diag}\left(\alpha_+\alpha_+^{\mathsf{H}}\right) = \mathbf{1}$. Hence

$$c\,i_j^{\Delta}\overline{\beta}_j \;=\; a_j, \qquad\qquad c\left|i_j^{\Delta}\right|^2 \;=\; b_j \tag{16.33}$$

which is a system of 2 quadratic equations. This yields the magnitude of i_j^{Δ}:

$$\left|i_j^{\Delta}\right|^2 \;=\; \frac{b_j}{c}$$

which in particular means that the specification cannot be arbitrary, e.g., b_j/c must be real.

When the reference angle $\angle V_j^a := \theta_j^a$ is given, let $\phi_j := \angle i_j^{\Delta}$. Given $i_j^{\Delta} := \sqrt{\frac{b_j}{c}}e^{\mathbf{i}\phi_j}$, all the other variables $\left(v_j^{\Delta}, i_j^{\Delta}, \beta_j\right) \in \mathbb{C}^3$ and $I_k^{\Delta} \in \mathbb{C}^3$ can be obtained as in Example 16.8, as a function of ϕ_j which can then be determined from the given reference angle:

$$\angle V_j^a \;=\; \angle\left[\frac{1}{3}\Gamma^{\mathsf{T}} v_j^{\Delta}\alpha_+ + \gamma_j\mathbf{1}\right]^a \;=\; \theta_j^a$$

This also shows that all variables are (generalized) balanced positive-sequence sets. □

Remark 16.4 (Nonuniqueness of specification)**.** Device specification is not unique and depends on the application under study. For Example 16.8, since both internal voltages V_j^Δ and V_k^Δ are obtained in terms of I_j^Δ in (16.25), we can either specify γ_j for the power source and derive γ_k of the impedance through the network equation, as done in Example 16.8, or alternatively, we can specify γ_k and determine γ_j from the network equation instead. While Example 16.8 contains no power sources, the next example illustrates multiple ways to specify and solve the case when both the generator and the load are power sources.

Also see Remark 16.6 for discussions on the asymmetry in device specifications. □

The next example uses the internal model or an external model of power sources, depending on how the power sources are specified. Specifically the solution boils down to a system of quadratic equations that can be solved numerically. All other variables can then be derived analytically in terms of a solution of the quadratic equations. For each of the two power sources, if its zero-sequence voltage γ_i is specified, we will use the internal model for the power source to obtain the system of quadratic equations in the internal currents I_i^Δ. Then the internal voltage V_i^Δ can be derived and, with the given γ_i, the terminal voltages V_i. If its zero-sequence current β_i is specified, on the other hand, we will use an external model to obtain the quadratic equations in the terminal current I_i from which, with the given β_i, the internal current I_i^Δ can then be derived. The network equation is used to express V_i^Δ in terms of I_i^Δ in the first case and express V_i in terms of I_i in the second case in the derivation of the system of quadratic equations.

Example 16.11 (Power sources)**.** Consider the system in Figure 16.7 where both the generator and load are power sources. Suppose the line admittances $\left(y_{jk}^s, y_{jk}^m, y_{kj}^m\right)$ are specified with nonzero y_{jk}^m, y_{kj}^m and assumption C16.1, as in Example 16.8.

1 Suppose the power sources are specified as $\left(\sigma_j^\Delta, \gamma_j\right)$ and $\left(\sigma_k^\Delta, \gamma_k\right)$. Determine all variables $\left(V_i^\Delta, I_i^\Delta, \beta_i\right)$ and (V_i, I_i, s_i), $i = j, k$.
2 Suppose the power sources are specified as $\left(\sigma_j^\Delta, \beta_j\right)$ and $\left(\sigma_k^\Delta, \beta_k\right)$. Determine all variables $\left(V_i^\Delta, I_i^\Delta\right)$ and $(V_i, I_i, s_i, \gamma_i)$, $i = j, k$.
3 Suppose the power sources are specified as $\left(\sigma_j^\Delta, \gamma_j\right)$ and $\left(\sigma_k^\Delta, \beta_k\right)$. Determine all variables $\left(V_i^\Delta, I_i^\Delta\right)$ and (V_i, I_i, s_i), $i = j, k$, and β_j, γ_k.

Solution.

1 The internal model of the power sources, the conversion rules, and the current

balance equation are

$$\sigma_i^\Delta := \operatorname{diag}\left(V_i^\Delta I_i^{\Delta\mathsf{H}}\right), \qquad V_i^\Delta = \Gamma V_i, \qquad I_i = -\Gamma^\mathsf{T} I_i^\Delta, \qquad i = j,k \tag{16.34a}$$

$$\begin{bmatrix} I_j \\ I_k \end{bmatrix} = \begin{bmatrix} y_{jk}^s + y_{jk}^m & -y_{jk}^s \\ -y_{jk}^s & y_{jk}^s + y_{kj}^m \end{bmatrix} \begin{bmatrix} V_j \\ V_k \end{bmatrix} \tag{16.34b}$$

Assume the admittance matrix Y in (16.34b) is invertible and let $Y^{-1} =: \begin{bmatrix} z_{jj} & z_{jk} \\ z_{kj} & z_{kk} \end{bmatrix}$. Then substituting the conversion rules into the network equation (16.34b) yields

$$\begin{bmatrix} V_j^\Delta \\ V_k^\Delta \end{bmatrix} = \underbrace{-\operatorname{diag}(\Gamma,\Gamma) \begin{bmatrix} z_{jj} & z_{jk} \\ z_{kj} & z_{kk} \end{bmatrix} \operatorname{diag}\left(\Gamma^\mathsf{T},\Gamma^\mathsf{T}\right)}_{Z := \begin{bmatrix} Z_{jj} & Z_{jk} \\ Z_{kj} & Z_{kk} \end{bmatrix}} \begin{bmatrix} I_j^\Delta \\ I_k^\Delta \end{bmatrix} \tag{16.35}$$

Substituting V_j^Δ and V_k^Δ into the internal power source models in (16.34a) yelds

$$\sigma_j^\Delta := -\operatorname{diag}\left(\left(Z_{jj} I_j^\Delta + Z_{jk} I_k^\Delta\right) I_j^{\Delta\mathsf{H}}\right), \quad \sigma_k^\Delta := -\operatorname{diag}\left(\left(Z_{kj} I_j^\Delta + Z_{kk} I_k^\Delta\right) I_k^{\Delta\mathsf{H}}\right) \tag{16.36}$$

This is a system of 6 quadratic equations that can be solved numerically for $\left(I_j^\Delta, I_k^\Delta\right) \in \mathbb{C}^6$.

All other variables can then be derived in terms of a solution $\left(I_j^\Delta, I_k^\Delta\right)$. Specifically, the internal voltages can be obtained from the internal power source model (16.34a) (or equivalently from (16.35)), $V_i^\Delta = \left(\operatorname{diag}\left(I_i^{\Delta\mathsf{H}}\right)\right)^{-1} \sigma_i^\Delta$, $i = 1,2$. Using γ_i, the terminal voltages are determined by the conversion rule, $V_i = \frac{1}{3}\Gamma^\mathsf{T} V_i^\Delta + \gamma_i \mathbf{1}$, $i = 1,2$. In terms of I_i^Δ we have $\beta_i := \frac{1}{3}\mathbf{1}^\mathsf{T} I_i^\Delta$ and $I_i = -\Gamma^\mathsf{T} I_i^\Delta$, $i = j,k$. The terminal power is $s_i := \operatorname{diag}\left(V_i I_i^\mathsf{H}\right)$, $i = j,k$.

2 When $\left(\gamma_j, \gamma_k\right)$ are given as in part 1, we set up equation (16.36) to solve numerically for $\left(I_j^\Delta, I_k^\Delta\right)$, so that V_i^Δ and then V_i can be derived for $i = j,k$. When $\left(\beta_j, \beta_k\right)$ are given instead, we will solve numerically for (V_j, V_k) by using the external model (14.25a) of a power source, reproduced here:

$$\sigma_i^\Delta = -\frac{1}{3}\operatorname{diag}\left(\Gamma\left(V_i I_i^\mathsf{H}\right)\Gamma^\mathsf{T}\right) + \overline{\beta}_i \Gamma V_i, \qquad \mathbf{1}^\mathsf{T} I_i = 0, \qquad i = j,k$$

and the network equation (16.34b). Note that all these equations relate terminal voltages and currents.

Specifically, instead of (16.35)), obtain from the network equation (16.34b)

$$\begin{bmatrix} V_j \\ V_k \end{bmatrix} = \begin{bmatrix} z_{jj} & z_{jk} \\ z_{kj} & z_{kk} \end{bmatrix} \begin{bmatrix} I_j \\ I_k \end{bmatrix}$$

Substituting into the external models of the power sources we have

$$\sigma_j^\Delta = -\frac{1}{3}\operatorname{diag}\left(\Gamma\left(z_{jj}I_j + z_{jk}I_k\right)I_j^{\mathsf{H}}\Gamma^{\mathsf{T}}\right) + \overline{\beta}_j\Gamma\left(z_{jj}I_j + z_{jk}I_k\right), \quad \mathbf{1}^{\mathsf{T}}I_j = 0$$

$$\sigma_k^\Delta = -\frac{1}{3}\operatorname{diag}\left(\Gamma\left(z_{kj}I_j + z_{kk}I_k\right)I_k^{\mathsf{H}}\Gamma^{\mathsf{T}}\right) + \overline{\beta}_k\Gamma\left(z_{kj}I_j + z_{kk}I_k\right), \quad \mathbf{1}^{\mathsf{T}}I_k = 0$$

This is a system of 8 (redundant) quadratic equations that can be solved numerically for $(I_j, I_k) \in \mathbb{C}^6$. Given a solution (I_j, I_k), the internal currents can be determined from the conversion rule and the given (β_j, β_k) as $I_i^\Delta = -\frac{1}{3}\Gamma I_i + \beta_i \mathbf{1}$, $i = j, k$. The remaining variables can then be derived as in part 1.

3 This combines the solution approaches of parts 1 and 2. Specifically we use the internal model for power source j, the external model for k:

$$\sigma_j^\Delta := \operatorname{diag}\left(V_j^\Delta I_j^{\Delta\mathsf{H}}\right), \qquad V_j^\Delta = \Gamma V_j, \qquad I_j = -\Gamma^{\mathsf{T}} I_j^\Delta \tag{16.37a}$$

$$\sigma_k^\Delta = -\frac{1}{3}\operatorname{diag}\left(\Gamma\left(V_k I_k^{\mathsf{H}}\right)\Gamma^{\mathsf{T}}\right) + \overline{\beta}_k\Gamma V_k, \qquad \mathbf{1}^{\mathsf{T}}I_k = 0 \tag{16.37b}$$

From the network equation (16.34b) we have

$$\begin{bmatrix} V_j^\Delta \\ V_k \end{bmatrix} = \operatorname{diag}(\Gamma, \mathbb{I}) \begin{bmatrix} z_{jj} & z_{jk} \\ z_{kj} & z_{kk} \end{bmatrix} \operatorname{diag}\left(-\Gamma^{\mathsf{T}}, \mathbb{I}\right) \begin{bmatrix} I_j^\Delta \\ I_k \end{bmatrix} = \begin{bmatrix} -\Gamma z_{jj}\Gamma^{\mathsf{T}} & \Gamma z_{jk} \\ -z_{kj}\Gamma^{\mathsf{T}} & z_{kk} \end{bmatrix} \begin{bmatrix} I_j^\Delta \\ I_k \end{bmatrix}$$

Substituting V_j^Δ and V_k into the internal power source models in (16.37) yelds

$$\sigma_j^\Delta := \operatorname{diag}\left(\left(-\Gamma z_{jj}\Gamma^{\mathsf{T}} I_j^\Delta + \Gamma z_{jk} I_k\right) I_j^{\Delta\mathsf{H}}\right)$$

$$\sigma_k^\Delta = -\frac{1}{3}\operatorname{diag}\left(\Gamma\left(-z_{kj}\Gamma^{\mathsf{T}} I_j^\Delta + z_{kk} I_k\right) I_k^{\mathsf{H}}\Gamma^{\mathsf{T}}\right) + \overline{\beta}_k\Gamma\left(-z_{kj}\Gamma^{\mathsf{T}} I_j^\Delta + z_{kk} I_k\right), \quad \mathbf{1}^{\mathsf{T}}I_k = 0$$

This is a system of 7 (redundant) quadratic equations that can be solved numerically for $\left(I_j^\Delta, I_k\right) \in \mathbb{C}^6$. All other variables can then be derived analytically in terms of a solution $\left(I_j^\Delta, I_k\right)$ as done in parts 1 and 2.

□

16.2.2 General analysis problem

We now formulate a general three-phase analysis problem. Consider a three-phase network $G := (\overline{N}, E)$ where each line $(j,k) \in E$ is characterized by 3×3 series and shunt admittance matrices $\left(y_{jk}^s, y_{jk}^m, y_{kj}^m\right)$. At each bus $j \in \overline{N}$ we assume, without loss of generality, there is a single three-wire device in either Y or Δ configuration. Associated with each device j are its internal variables $\left(V_j^{Y/\Delta}, I_j^{Y/\Delta}, s_j^{Y/\Delta}, \beta_j\right) \in \mathbb{C}^{10}$ (or in $\mathbb{C}^9$ for Y-configured devices j without β_j) and terminal variables $(V_j, I_j, s_j, \gamma_j) \in \mathbb{C}^{10}$. Some of these variables will be specified in our formulation. The others are to be computed from network equations, device models and the conversion rules.

We start by describing which of these variables are specified for each type of devices using the internal and external device models in Tables 14.3 and 14.4. It is important to keep in mind that device specification is not unique and our formulation here may need to be modified depending on the details of an application, especially for problems involving power sources as discussed in Remark 16.4 and illustrated in Example 16.11. The principle of analysis described here, however, is widely applicable and can be applied to other formulations. For instance, we formulate our analysis problem in a three-wire model. If the neutrals of two Y-configured devices are not grounded and are connected to each other through a four-wire line, then a four-wire model needs to be used; see Example 16.5 and Exercise 16.7. In that case the neutral voltages of these devices may not be arbitrarily specified but must be determined through network equations and device models, even for a voltage source, unlike the formulation here.

Partition $\overline{N}$ into 8 disjoint subsets:

- $N_v^{Y/\Delta}$: buses with ideal voltage sources in Y or Δ configurations. Let $N_v := N_v^Y \cup N_v^\Delta$.
- $N_c^{Y/\Delta}$: buses with ideal current sources in Y or Δ configurations. Let $N_c := N_c^Y \cup N_c^\Delta$.
- $N_i^{Y/\Delta}$: buses with impedances in Y or Δ configurations. Let $N_i := N_i^Y \cup N_i^\Delta$.
- $N_p^{Y/\Delta}$: buses with ideal power sources in Y or Δ configurations. Let $N_p := N_p^Y \cup N_p^\Delta$.

with $\overline{N} = N_v \cup N_c \cup N_i \cup N_p$. These devices are specified as follows.

1 *Voltage source* $\left(E_j^Y, \gamma_j\right)$ *or* $\left(E_j^\Delta, \gamma_j, \beta_j\right)$: It is specified by its internal voltage $E_j^{Y/\Delta}$ and a parameter γ_j where $\gamma_j := V_j^n$ is the neutral voltage for Y configuration and $\gamma_j := \frac{1}{3}\mathbf{1}^\mathsf{T} V_j$ is the zero-sequence terminal voltage for Δ configuration. For Δ configuration, E_j^Δ should satisfy $\mathbf{1}^\mathsf{T} E_j^\Delta = 0$. The zero-sequence internal current $\beta_j := \frac{1}{3}\mathbf{1}^\mathsf{T} I_j^\Delta$ also needs to be specified in order to determine I_j^Δ from the terminal current I_j.
2 *Current source* $\left(J_j^Y, \gamma_j\right)$ *or* J_j^Δ: It is specified by its internal current $J_j^{Y/\Delta}$. For a Y-configured current source, its neutral voltage γ_j is also specified. For a Δ-configured current source, the zero-sequence voltage γ_i generally need not be specified and can be derived in terms of other quantities, but there are exceptions; see Remark 16.8.
3 *Power source* $\left(\sigma^Y, \gamma_j\right)$ *or* $\left(\sigma^\Delta, \gamma_j\right)$: It is specified by its internal power and zero-sequence voltage $\left(\sigma^{Y/\Delta}, \gamma_j\right)$. See Example 16.11 for other power source specifications and their solution methods.
4 *Impedance* $\left(z^Y, \gamma_j\right)$ *or* z^Δ: A Y-configured impedance j is specified by its internal impedance z_j^Y and the neutral voltage γ_j. A Δ-configured impedance j is specified by z_j^Δ. Its zero-sequence voltage and current $\left(\gamma_j, \beta_j\right)$ can generally be derived from network equations as we will see in Chapter 16.2.3.

A three-phase analysis problem is: given devices specified as above connected by

lines with given admittance matrices $\left(y^s_{jk}, y^m_{jk}\right)$, $\left(y^s_{kj}, y^m_{kj}\right)$, determine some or all of the internal variables $\left(V^{Y/\Delta}_j, I^{Y/\Delta}_j, s^{Y/\Delta}_j, \beta_j\right)$ and terminal variables $(V_j, I_j, s_j, \gamma_j)$ at every bus j. This is summarized in Table 16.3. Note that the analysis problem does not

Buses j	Specification	Unknowns
N^Y_v	$V^Y_j := E^Y_j$, γ_j	$\left(I^Y_j, s^Y_j\right)$, (V_j, I_j, s_j)
N^Δ_v	$V^\Delta_j := E^\Delta_j$, γ_j, β_j,	$\left(I^\Delta_j, s^\Delta_j\right)$, (V_j, I_j, s_j)
N^Y_c	$I^Y_j := J^Y_j$, γ_j	$\left(V^Y_j, s^\Delta_j\right)$, (V_j, I_j, s_j)
N^Δ_c	$I^\Delta_j := J^\Delta_j$	$\left(V^\Delta_j, s^\Delta_j, \beta_j\right)$, $(V_j, I_j, s_j, \gamma_j)$
N^Y_i	z^Y_j, γ_j	$\left(V^Y_j, I^Y_j, s^Y_j\right)$, (V_j, I_j, s_j)
N^Δ_i	z^Δ_j	$\left(V^\Delta_j, I^\Delta_j, s^\Delta_j, \beta_j\right)$, $(V_j, I_j, s_j, \gamma_j)$
N^Y_p	σ^Y_j, γ_j	$\left(V^Y_j, I^Y_j\right)$, (V_j, I_j, s_j)
N^Δ_p	σ^Δ_j, γ_j	$\left(V^\Delta_j, I^\Delta_j, \beta_j\right)$, (V_j, I_j, s_j)

Table 16.3 Three-phase analysis problem: given the specification in blue, compute the remaining unknowns in black.

assume C16.1 and therefore each line (j,k) may model a transmission or distribution line, or a three-phase transformer where its series admittance matrices y^s_{jk} and y^s_{kj} may be different.

We make a few remarks on the voltage γ_j. See Remark 16.3 on how the loop flow β_k of an impedance k may depend on β_j of a current source j.

Remark 16.5 (Voltage γ_j)**.** 1 *Y configuration.* The voltage parameter γ_j needs to be specified for every Y-configured device in our formulation here. It may be specified explicitly, or more likely, indirectly. By that, we mean information additional to generic device models is available to either compute their values, provide additional equations, or eliminate them in terms of other variables. For instance if the neutral of a Y-configured device is grounded and all voltages are defined with respect to the ground, then $\gamma_j = V^n_j = -z^n_j \left(\mathbf{1}^\mathsf{T} I_j\right)$, which allows the elimination of γ_j from the model. If the neutral is grounded directly (i.e., $z^n_j = 0$), then $\gamma_j = 0$. If the neutral is not grounded but the internal voltage V^Y_j is known to satisfy $\mathbf{1}^\mathsf{T} V^Y_j = 0$, then $\gamma_j = \frac{1}{3}\mathbf{1}^\mathsf{T} V_j$. This is studied in detail in Examples 16.3 and 16.4 for a three-wire line model as we have been assuming in almost all of our analysis.

For a Y-configured current source, γ_j is usually not needed to determine its terminal voltage V_j, but needed to compute its internal voltage $V^Y_j = V_j - \gamma_j \mathbf{1}$ from the terminal voltage V_j.

As noted above, Example 16.5 and Exercise 16.7 consider a four-wire line model where the neutrals of the voltage source and the impedance are connected

to each other. Here the (internal) neutral voltages (γ_j, γ_k) of neither device can be arbitrarily specified but must be determined through the network equation and device models.

2 Δ *configuration.* For a Δ-configured voltage source, the zero-sequence voltage $\gamma_j := \frac{1}{3}\mathbf{1}^\mathsf{T} V_j$ needs to be specified, e.g., by specifying one of its terminal voltages, say, V_j^a. For a Δ-configured current source or impedance, γ_j can be determined once its terminal voltage V_j is determined from network equations. For a Δ-configured power source, typically either γ_j or β_j can be specified; see Example 16.11.

3 *Neutral voltage γ_j and zero-sequence voltage.* For any Y-configured device, we have

$$V_j \;=\; V_j^Y \;+\; V_j^n \mathbf{1}$$

The parameter $\gamma_j := V_j^n$ may or may not equal the zero-sequence voltage $\frac{1}{3}\mathbf{1}^\mathsf{T} V_j$. They are equal if and only if the internal voltages have no zero-sequence component since $\frac{1}{3}\mathbf{1}^\mathsf{T} V_j = \frac{1}{3}\mathbf{1}^\mathsf{T} V_j^Y + V_j^n$.

□

Remark 16.6 (Asymmetry in Δ specification)**.** As summarized in Table 16.3, in our formulation, for Δ configuration, a voltage source needs to specify both (γ_j, β_j), but a power source only needs to specify its γ_j, and a current source or impedance needs to specify none. This asymmetry is because internal currents I_j^Δ contain more information (they fix β_j) than internal voltages V_j^Δ (they do not fix γ_j). Device specification and network equation determine $\left(E_j^\Delta, I_j\right)$ for voltage sources, which contains neither β_j nor γ_j. These quantities therefore need to be specified. Device specification and network equation, on the other hand, determine $\left(J_j^\Delta, V_j\right)$ for current sources, which contains both β_j and γ_j. For impedances, as we will see in Chapter 16.2.3, the network equation will determine their internal currents I_j^Δ which contain β_j. When the terminal voltages of all sources, including power sources, are specified or obtained, the terminal voltages V_j of impedances can be determined by the network equation. Therefore both (γ_j, β_j) are determined by the network equation in that case. □

16.2.3 Solution strategy

The solution strategy for the problem formulated in Chapter 16.2.2 consists of three steps:

1 Write down a network equation that relates the terminal variables (V, I, s), either the current balance equation (16.5)(16.6) $I = YV$ or the power flow equation (16.12). As discussed in Remark 16.7 we can always use the linear equation $I = YV$.

2 Write down the device models of the given collection of sources and impedances, either their internal models and conversion rules, or their external models.
3 Numerically solve this system of equations for desired variables.

Step 1 specifies, for the entire network, an equation that relates all the terminal variables. For examples, see (16.38) and (16.43) for analysis problems without and with power sources respectively. Step 2 specifies, for each device, equations relating its terminal variables to its internal variables or specified parameters. For examples, see (16.39d)(16.39d) and (16.44a) respectively.

Remark 16.7 (Nonlinearity)**.** Using the nonlinear power flow equations $s_j = \text{diag}\left(V_j(V_j - V_k)^{\mathsf{H}}\left(y^s_{jk}\right)^{\mathsf{H}}\right)$ as the network equation in Step 1 is equivalent to using the linear current balance equation $I = YV$. This is because dividing both sides of the power flow equations by V_j and taking complex conjugate yields $I = YV$. Therefore if no power sources are involved, then the device models of voltage sources, current sources and impedances are linear and therefore the overall model will be linear.

If power sources are involved, then even if we use $I = YV$ as the network equation, the device models of power sources will be quadratic and therefore the overall network will be nonlinear. In this case the power source device model is the only place where nonlinearity appears. □

In the rest of this subsection we first describe in detail Steps 1 and 2 in the general solution strategy outline above to obtain a system of equations that can be solved numerically. In light of Remark 16.7 we will use the current balance equation $I = VY$ as our network equation. Then, motivated by the examples in Chapter 16.2.1, we show how to reduce the entire system of equations obtained from Steps 1 and 2 into a smaller system with possibly much fewer variables, which must be solved numerically. All other variables can then be derived analytically in terms of the solution of the reduced system. (For problems without power sources, this reduces equations (16.38)(16.39) to (16.42).) This simpler solution strategy not only reduces the size of the system that needs numerical solution, but more importantly, it often reveals more clearly the essential structure of the problem. For instance, for problems with power sources, the reduced system is equation (16.47) which consists of a linear equation and a quadratic equation due to power sources.

We first derive the solution for the case without power sources. We then show how to extend the solution to incorporate power sources simply by adding their device models to the systems of equations. We will focus on determining terminal and internal voltages and currents. Once they are determined, internal and external powers can be calculated using $s_j^{Y/\Delta} := \text{diag}\left(V_j^{Y/\Delta} I_j^{Y/\Delta \mathsf{H}}\right)$ and $s_j := \text{diag}\left(V_j I_j^{\mathsf{H}}\right)$ respectively.

Without power sources.

Recall that $N_v := N_v^Y \cup V_v^\Delta$, $N_c := N_c^Y \cup V_c^\Delta$, and $N_i := N_i^Y \cup V_i^\Delta$ are the set of buses with, respectively, voltage sources, current sources, and impedances. With a slight abuse of notation define the following (column) vectors of terminal voltages and currents:

$$(V_v, I_v) := \left(V_j, I_j, j \in N_v\right), \quad (V_c, I_c) := \left(V_j, I_j, j \in N_c\right), \quad (V_i, I_i) := \left(V_j, I_j, j \in N_i\right)$$

Some of them will be specified and the remaining voltages and currents will be determined from the network equation and device models. Step 1 of the solution strategy is to write the network equation $I = YV$:

$$\begin{bmatrix} I_v \\ I_c \\ I_i \end{bmatrix} = \underbrace{\begin{bmatrix} Y_{vv} & Y_{vc} & Y_{vi} \\ Y_{cv} & Y_{cc} & Y_{ci} \\ Y_{iv} & Y_{ic} & Y_{ii} \end{bmatrix}}_{Y} \begin{bmatrix} V_v \\ V_c \\ V_i \end{bmatrix} \tag{16.38}$$

where the admittance matrix Y is defined in (16.6).

Step 2 is to describe the device models. The specifications for voltage sources, current sources and impedances are, from Table 16.3:

$$\begin{aligned} \left(E_v^{Y/\Delta}, \gamma_v^{Y/\Delta}, \beta_v^\Delta\right) &:= \left(E_j^{Y/\Delta}, \gamma_j, j \in N_v^{Y/\Delta};\ \beta_j, j \in N_v^\Delta\right) \\ \left(J_c^{Y/\Delta}, \gamma_c^Y\right) &:= \left(J_j^{Y/\Delta}, j \in N_c^{Y/\Delta};\ \gamma_j, j \in N_c^Y\right) \\ \left(Z_i^{Y/\Delta}, \gamma_i^Y\right) &:= \left(\operatorname{diag}\left(z_j^\Delta, j \in N_i^{Y/\Delta}\right);\ \gamma_j, j \in N_i^Y\right) \end{aligned}$$

To unify notation we define the following matrices

$$\begin{aligned} \Gamma_v^{Y\dagger} &:= \mathbb{I}_v^Y \otimes \mathbb{I}, & \Gamma_v^{\Delta\dagger} &:= \mathbb{I}_v^\Delta \otimes \Gamma^\dagger, & \Gamma_v^\dagger &:= \operatorname{diag}\left(\Gamma_v^{Y\dagger}, \Gamma_v^{\Delta\dagger}\right) \\ \Gamma_c^Y &:= \mathbb{I}_c^Y \otimes \mathbb{I}, & \Gamma_c^\Delta &:= \mathbb{I}_c^\Delta \otimes \Gamma, & \Gamma_c &:= \operatorname{diag}\left(\Gamma_c^Y, \Gamma_c^\Delta\right) \\ \Gamma_i^Y &:= \mathbb{I}_i^Y \otimes \mathbb{I}, & \Gamma_i^\Delta &:= \mathbb{I}_i^\Delta \otimes \Gamma, & \Gamma_i &:= \operatorname{diag}\left(\Gamma_i^Y, \Gamma_i^\Delta\right) \end{aligned}$$

where $\mathbb{I}_v^Y, \mathbb{I}_c^Y, \mathbb{I}_i^Y$ are the identity matrices of sizes $|N_v^Y|$, $|N_c^Y|$, $|N_i^Y|$ respectively and $\mathbb{I}_v^\Delta, \mathbb{I}_c^\Delta, \mathbb{I}_i^\Delta$ denote the identity matrices of sizes $|N_v^\Delta|$, $|N_c^\Delta|$, $|N_i^\Delta|$ respectively. Define vectors of specifications

$$E_v := \begin{bmatrix} E_v^Y \\ E_v^\Delta \end{bmatrix}, \qquad J_c := \begin{bmatrix} J_c^Y \\ J_c^\Delta \end{bmatrix}, \qquad Z_i := \operatorname{diag}\left(Z_i^Y, Z_i^\Delta\right) \tag{16.39a}$$

$$\gamma_v := \begin{bmatrix} \gamma_v^Y \\ \gamma_v^\Delta \end{bmatrix}, \qquad \gamma_c := \begin{bmatrix} \gamma_c^Y \\ 0 \end{bmatrix}, \qquad \gamma_i := \begin{bmatrix} \gamma_i^Y \\ 0 \end{bmatrix} \tag{16.39b}$$

so that $\gamma_v \in \mathbb{C}^{|N_v|}$, $\gamma_c \in \mathbb{C}^{|N_c|}$ and $\gamma_i \in \mathbb{C}^{|N_i|}$. Then the terminal voltage and current

V_v and I_c in (16.38) are given by

$$V_v := \begin{bmatrix} E_v^Y + \gamma_v^Y \otimes \mathbf{1} \\ \Gamma_v^{\Delta\dagger} E_v^\Delta + \gamma_v^\Delta \otimes \mathbf{1} \end{bmatrix} = \Gamma_v^\dagger E_v + \gamma_v \otimes \mathbf{1} \tag{16.39c}$$

$$I_c := -\begin{bmatrix} J_c^Y \\ \Gamma_c^{\Delta\mathsf{T}} J_c^\Delta \end{bmatrix} = -\Gamma_c^\mathsf{T} J_c \tag{16.39d}$$

Define the following notations for internal variables of impedances:

$$I_i^Y := \left(I_j^Y, j \in N_i^Y\right), \qquad I_i^\Delta := \left(I_j^\Delta, j \in N_i^\Delta\right), \qquad I_i^{\text{int}} := \begin{bmatrix} I_i^Y \\ I_i^\Delta \end{bmatrix}$$

$$V_i^Y := \left(V_j^Y, j \in N_i^Y\right), \qquad V_i^\Delta := \left(V_j^\Delta, j \in N_i^\Delta\right), \qquad V_i^{\text{int}} := \begin{bmatrix} V_i^Y \\ V_i^\Delta \end{bmatrix}$$

The internal model of the impedances in Y and Δ configurations is then

$$V_i^{\text{int}} = Z_i I_i^{\text{int}} \tag{16.39e}$$

where Z_i is defined in (16.39a). The conversion rule for the current and voltage (I_i, V_i) is:

$$I_i = \begin{bmatrix} -I_i^Y \\ -\Gamma_i^{\Delta\mathsf{T}} I_i^\Delta \end{bmatrix} = -\Gamma_i^\mathsf{T} I_i^{\text{int}}, \quad \Gamma_i V_i = \begin{bmatrix} V_i^Y + \gamma_i^Y \otimes \mathbf{1} \\ V_i^\Delta \end{bmatrix} = V_i^{\text{int}} + \gamma_i \otimes \mathbf{1} \tag{16.39f}$$

The analysis problem is: Solve the network equation (16.38) and the device models (16.39) for the unknown external and internal variables. This can be done by numerically solving the system of equations (16.38)(16.39). Note that the analysis problem defined by (16.38)(16.39) does not assume C16.1 and therefore each line (j,k) may model a transmission or distribution line, or a three-phase transformer where its series admittance matrices y_{jk}^s and y_{kj}^s may be different.

The intuition from Example 16.3, Example 16.6 and Exercise 16.8 suggests that, instead of numerically solving (16.38)(16.39), it is possible to reduce it to a smaller system of equations with possibly much fewer variables. Once the reduced system is solved numerically, all other variables can be derived analytically in terms of a solution of the reduced system. The key observation from the examples is to first solve for the internal currents I_i^{int} of all impedances, not their internal voltages V_i^{int} nor other terminal variables (V_i, I_i), using the network equation, the internal device models and the conversion rules. We now explain how to obtain the reduced system of equations in the internal currents I_i^{int} of all impedances and the terminal voltages V_c of all current sources.

Substituting I_i in (16.39f) into (16.38) we have

$$\begin{bmatrix} I_c \\ -\Gamma_i^\mathsf{T} I_i^{\text{int}} \end{bmatrix} = \begin{bmatrix} Y_{cv} \\ Y_{iv} \end{bmatrix} V_v + \begin{bmatrix} Y_{cc} & Y_{ci} \\ Y_{ic} & Y_{ii} \end{bmatrix} \begin{bmatrix} V_c \\ V_i \end{bmatrix} \tag{16.40}$$

To express V_i in this equation in terms of I_i^{int}, suppose the inverse

$$\begin{bmatrix} Z_{cc} & Z_{ci} \\ Z_{ic} & Z_{ii} \end{bmatrix} := \begin{bmatrix} Y_{cc} & Y_{ci} \\ Y_{ic} & Y_{ii} \end{bmatrix}^{-1} \tag{16.41}$$

exists and multiplying both sides of (16.40) by this inverse and then by $\text{diag}(\mathbb{I}_c, \Gamma_i)$ we have

$$\text{diag}(\mathbb{I}_c \otimes \mathbb{I}, \Gamma_i) \begin{bmatrix} Z_{cc} & Z_{ci} \\ Z_{ic} & Z_{ii} \end{bmatrix} \begin{bmatrix} I_c \\ -\Gamma_i^{\mathsf{T}} I_i^{\text{int}} \end{bmatrix} = \underbrace{\text{diag}(\mathbb{I}_c \otimes \mathbb{I}, \Gamma_i) \begin{bmatrix} Z_{cc} & Z_{ci} \\ Z_{ic} & Z_{ii} \end{bmatrix} \begin{bmatrix} Y_{cv} \\ Y_{iv} \end{bmatrix}}_{\begin{bmatrix} A_{cv} \\ A_{iv} \end{bmatrix}} V_v + \begin{bmatrix} V_c \\ \Gamma_i V_i \end{bmatrix}$$

where $\mathbb{I}_c$ is the identity matrix of size $|N_c|$. Substituting $\Gamma_i V_i = V_i^{\text{int}} + \gamma_i \otimes \mathbf{1} = Z_i I_i^{\text{int}} + \gamma_i \otimes \mathbf{1}$ from (16.39e) and (16.39f) and re-arranging, we have thus reduced the original system (16.38)(16.39) into the following reduced system in (V_c, I_i^{int}):

$$\begin{bmatrix} \mathbb{I}_c \otimes \mathbb{I} & Z_{ci}\Gamma_i^{\mathsf{T}} \\ 0 & \Gamma_i Z_{ii} \Gamma_i^{\mathsf{T}} + Z_i \end{bmatrix} \begin{bmatrix} V_c \\ I_i^{\text{int}} \end{bmatrix} = \begin{bmatrix} Z_{cc} \\ \Gamma_i Z_{ic} \end{bmatrix} I_c - \begin{bmatrix} A_{cv} \\ A_{iv} \end{bmatrix} V_v - \begin{bmatrix} 0 \\ \gamma_i \otimes \mathbf{1} \end{bmatrix} \tag{16.42}$$

Here V_v, I_c, Z_i and γ_i are given by (16.39), the submatrices $Z_{cc}, Z_{ci}, Z_{ic}, Z_{ii}$ are from the inverse in (16.41), and

$$\begin{bmatrix} A_{cv} \\ A_{iv} \end{bmatrix} := \text{diag}(\mathbb{I}_c \otimes \mathbb{I}, \Gamma_i) \begin{bmatrix} Z_{cc} & Z_{ci} \\ Z_{ic} & Z_{ii} \end{bmatrix} \begin{bmatrix} Y_{cv} \\ Y_{iv} \end{bmatrix}$$

All quantities on the right-hand side of (16.42) are known. This is a system of $3(|N_c| + |V_i|)$ linear equations in $3(|N_c| + |V_i|)$ unknowns (V_c, I_i^{int}). Assuming the matrix on the left-hand side is invertible, the methods described in Chapter 4.2.5 can be used to compute numerically a solution (V_c, I_i^{int}) of (16.42).

We now explain how to derive all the remaining variables.

1 For impedances, with I_i^{int}, the internal voltage $V_i^{\text{int}} = Z_i I_i^{\text{int}}$ and the terminal current $I_i = -\Gamma_i^{\mathsf{T}} I_i^{\text{int}}$ from the internal model (16.39e) and the conversion rule (16.39f). With both I_i^{int} and V_c, we can obtain V_i from (16.40). The zero-sequence voltages and currents $\left(\gamma_j = \frac{1}{3}\mathbf{1}^{\mathsf{T}} V_j, \beta_j := \frac{1}{3}\mathbf{1}^{\mathsf{T}} I_j^{\Delta}\right)$ of all Δ-configured impedances $j \in N_i^{\Delta}$ can then be derived from (V_i, I_i^{int}). This completes the derivation of all voltages and currents of impedances.
2 For voltage sources, with (V_v, V_c, V_i), the terminal current I_v can be derived from (16.38). For Y-configured voltage sources $j \in N_v^Y$, the internal currents are $I_j^Y = -I_j$. For Δ-configured voltage sources $j \in N_v^{\Delta}$, β_j are given and hence the internal currents are $I_j^{\Delta} = -\frac{1}{3}\Gamma I_j + \beta_j \mathbf{1}$. This completes the derivation of all voltages and currents of power sources.
3 For Y-configured current sources $j \in N_c^Y$, γ_j are given and hence the internal voltages are $V_j^Y = V_j - \gamma_j \mathbf{1}$. For Δ-configured current sources $j \in N_c^{\Delta}$, β_j can be

calculated from J_j^Δ and $\left(V_j^\Delta, \gamma_j\right)$ can be calculated from V_j. This completes the derivation of the voltages and currents of all current sources.

With all voltages and currents determined, the internal and external powers are then $s_j^{Y/\Delta} := \operatorname{diag}\left(V_j^{Y/\Delta} I_j^{Y/\Delta \mathsf{H}}\right)$ and $s_j = \operatorname{diag}\left(V_j I_j^{\mathsf{H}}\right)$, $j \in \overline{N}$, respectively. This completes the derivation of the variables of all devices in the network.

Remark 16.8. The derivation of the reduced system (16.42) depends critically on the assumption that the admittance matrix in (16.40) and the effective impedance matrix $\Gamma_i Z_{ii} \Gamma_i^{\mathsf{T}} + Z_i$ in (16.42) are invertible. When that is not the case, additional information will be needed to uniquely determine all the quantities.

1 If there are voltage sources then the matrix in (16.40) is a strict submatrix of an admittance matrix and therefore will be invertible if the conditions in Theorem 4.5 are satisfied, including the condition $y_{jk}^s = y_{kj}^s$.

   In Example 16.3 where a voltage source j supplies an impedance k both in Y configuration over a three-phase line, the equation (16.40) is (16.14a) for which the inverse exists. In Example 16.6 where the devices are in Δ configuration, the equation (16.40) takes the form

$$-\Gamma^{\mathsf{T}} I_k^\Delta \;=\; -y_{jk}^s V_j + y_{jk}^s V_k$$

   so the inverse $\left(y_{jk}^s\right)^{-1}$ also exists.

2 When only current sources are present, the matrix in (16.40) is the network admittance matrix and is invertible if the conditions in Theorem 4.3 are satisfied, including the condition $y_{jk}^s = y_{kj}^s$. In particular if the shunt admittances of all three-phase lines are assumed zero, then the admittance matrix is not invertible because it will have zero row sums. In that case, additional information needs to be specified to provide an additional equation to (16.40) for solving $\left(V_c, I_i^{\text{int}}\right)$ and V_i.

   In Exercise 16.8 where the voltage source is replaced by a current source j and shunt admittances $\left(y_{jk}^m, y_{kj}^m\right)$ are assumed zero, the equation (16.40) takes the form

$$\begin{bmatrix} I_j \\ -\Gamma^{\mathsf{T}} I_k^\Delta \end{bmatrix} = \begin{bmatrix} y_{jk}^s & -y_{jk}^s \\ -y_{kj}^s & y_{kj}^s \end{bmatrix} \begin{bmatrix} V_j \\ V_k \end{bmatrix}$$

   for which the inverse does not exist. As a result the zero-sequence voltage γ_j of the current source is also specified to provide the additional equation for solving $\left(V_j, I_i^{\text{int}}\right)$. If the shunt admittances $\left(y_{jk}^m, y_{kj}^m\right)$ are nonzero as in Exercise 16.9, γ_j of the current source need not be specified and can be derived because the equation above will be invertible.

3 The reduced system (16.42) generalizes (16.14b) in Example 16.3 and (16.18) in Example 16.6 to general networks and with current sources.

□

With power sources.

Analysis problems with power sources can be solved following the same procedure, but with the addition of device models of power sources. Specifically the current balance equation (16.38) is extended to

$$\begin{bmatrix} I_v \\ I_c \\ I_i \\ I_p \end{bmatrix} = \underbrace{\begin{bmatrix} Y_{vv} & Y_{vc} & Y_{vi} & Y_{vp} \\ Y_{cv} & Y_{cc} & Y_{ci} & V_{cp} \\ Y_{iv} & Y_{ic} & Y_{ii} & Y_{ip} \\ Y_{pv} & Y_{pc} & Y_{pi} & Y_{pp} \end{bmatrix}}_{Y} \begin{bmatrix} V_v \\ V_c \\ V_i \\ V_p \end{bmatrix} \tag{16.43}$$

where $(V_p, I_p) := (V_j, I_j, j \in N_p)$, with $N_p := N_p^Y \cup N_p^\Delta$, are the terminal voltages and currents of power sources.

The device model (16.39) also needs to be extended to include power sources. For a Y-configured power source, $\left(s_j^Y := \sigma_j^Y, \gamma_j := V_j^n\right)$ are specified. For a Δ-configured power source, we assume that $\left(s_j^\Delta := \sigma_j^\Delta, \gamma_j := \frac{1}{3}\mathbf{1}^\mathsf{T} V_j\right)$ are specified. Let $\sigma_p := \begin{bmatrix} \sigma_p^Y \\ \sigma_p^\Delta \end{bmatrix}$. Then the internal models of the power sources in Y and Δ configurations are

$$\sigma_p^Y = \left(\operatorname{diag}\left(V_j^Y I_j^{Y\mathsf{H}}\right), j \in N_p^Y\right), \qquad \sigma_p^\Delta := \left(\operatorname{diag}\left(V_j^\Delta I_j^{\Delta\mathsf{H}}\right), j \in N_p^\Delta\right)$$

To simplify notation define the internal currents and voltages for all power sources:

$$I_p^Y := \left(I_j^Y, j \in N_p^Y\right), \qquad I_p^\Delta := \left(I_j^\Delta, j \in N_p^\Delta\right), \qquad I_p^{\text{int}} := \begin{bmatrix} I_p^Y \\ I_p^\Delta \end{bmatrix}$$

$$V_p^Y := \left(V_j^Y, j \in N_p^Y\right), \qquad V_p^\Delta := \left(V_j^\Delta, j \in N_p^\Delta\right), \qquad V_p^{\text{int}} := \begin{bmatrix} V_p^Y \\ V_p^\Delta \end{bmatrix}$$

Then the internal models of the power sources can be written as

$$\sigma_p = \operatorname{diag}\left(V_p^{\text{int}} I_p^{\text{intH}}\right) \tag{16.44a}$$

This is a quadratic equation in the unknowns internal voltage and current $\left(V_p^{\text{int}}, I_p^{\text{int}}\right)$.[5] Define

$$\Gamma_p^\Delta := \mathbb{I}_p^\Delta \otimes \Gamma, \qquad \Gamma_p := \operatorname{diag}\left(\mathbb{I}_p^Y, \Gamma_p^\Delta\right), \qquad \gamma_p := \begin{bmatrix} \gamma_p^Y \\ 0 \end{bmatrix}$$

[5] We can also use the equivalent model $\sigma_p = \operatorname{diag}\left(\left(\Gamma_p V_p - \gamma_p\right) I_p^{\text{intH}}\right)$ of power sources in terms of the terminal voltage V_p, the internal current I_p^{int}, and the neutral voltage γ_p^Y. The network equation however will only allow us to solve for $\Gamma_j V_j = V_j^\Delta$ for Δ-configured power sources j. Specifically V_p^{int} in (16.47a) will be replaced by $\Gamma_p V_p$ and (16.47b) by $\sigma_p = \operatorname{diag}\left(\left(\Gamma_p V_p - \gamma_p\right) I_p^{\text{intH}}\right)$. Therefore it is simpler to solve for the internal voltage V_p^{int} and then use γ_j to obtain the terminal voltages V_j of Δ-configured power sources j.

where $\mathbb{I}_p^{\Delta}$ denotes the identity matrix of size $|N_p^{\Delta}|$, $\mathbb{I}_p^{Y}$ the identity matrices of size $3|N_p^{Y}|$, and $\gamma_p^{Y} := \left(\gamma_j := V_j^n, j \in N_p^{Y}\right)$ are neutral voltages of all Y-configured power sources j. The current and voltage conversion rule is (similar to (16.39f))

$$I_p \;=\; -\Gamma_p\, I_p^{\text{int}}, \qquad\qquad \Gamma_p V_p \;=\; V_p^{\text{int}} + \gamma_p \otimes \mathbf{1} \tag{16.44b}$$

The analysis problem can be stated as: Solve the network equation (16.43) and the device models (16.39) (16.44) for the unknown external and internal variables. This system of equations (16.43)(16.39)(16.44) can be solved numerically.

We will follow the same procedure to reduce (16.43)(16.39)(16.44) into a smaller system of (nonlinear) equations that involves only $\left(V_c, I_i^{\text{int}}, I_p^{\text{int}}, V_p^{\text{int}}\right)$. All other variables can then be derived from a solution $\left(V_c, I_i^{\text{int}}, I_p^{\text{int}}, V_p^{\text{int}}\right)$.

Substituting I_i, I_p in (16.39f) and (16.44b) respectively into (16.43) we have

$$\operatorname{diag}\left(\mathbb{I}_c, -\Gamma_i^{\mathsf{T}}, -\Gamma_p^{\mathsf{T}}\right) \begin{bmatrix} I_c \\ I_i^{\text{int}} \\ I_p^{\text{int}} \end{bmatrix} \;=\; \begin{bmatrix} Y_{cv} \\ Y_{iv} \\ Y_{pv} \end{bmatrix} V_v \;+\; \begin{bmatrix} Y_{cc} & Y_{ci} & Y_{cp} \\ Y_{ic} & Y_{ii} & Y_{ip} \\ Y_{pc} & Y_{pi} & Y_{pp} \end{bmatrix} \begin{bmatrix} V_c \\ V_i \\ V_p \end{bmatrix} \tag{16.45}$$

Suppose the inverse

$$\begin{bmatrix} Z_{cc} & Z_{ci} & Z_{cp} \\ Z_{ic} & Z_{ii} & Z_{ip} \\ Z_{pc} & Z_{pi} & Z_{pp} \end{bmatrix} := \begin{bmatrix} Y_{cc} & Y_{ci} & Y_{cp} \\ Y_{ic} & Y_{ii} & Y_{ip} \\ Y_{pc} & Y_{pi} & Y_{pp} \end{bmatrix}^{-1} \tag{16.46}$$

exists and multiplying both sides by this inverse and then by $\operatorname{diag}(\mathbb{I}_c, \Gamma_i, \Gamma_p)$ we have

$$\operatorname{diag}(\mathbb{I}_c, \Gamma_i, \Gamma_p) \begin{bmatrix} Z_{cc} & Z_{ci} & Z_{cp} \\ Z_{ic} & Z_{ii} & Z_{ip} \\ Z_{pc} & Z_{pi} & Z_{pp} \end{bmatrix} \operatorname{diag}\left(\mathbb{I}_c, -\Gamma_i^{\mathsf{T}}, -\Gamma_p^{\mathsf{T}}\right) \begin{bmatrix} I_c \\ I_i^{\text{int}} \\ I_p^{\text{int}} \end{bmatrix} \;=\; \begin{bmatrix} B_{cv} \\ B_{iv} \\ B_{pv} \end{bmatrix} V_v \;+\; \begin{bmatrix} V_c \\ \Gamma_i V_i \\ \Gamma_p V_p \end{bmatrix}$$

where

$$\begin{bmatrix} B_{cv} \\ B_{iv} \\ B_{pv} \end{bmatrix} \;:=\; \operatorname{diag}(\mathbb{I}_c, \Gamma_i, \Gamma_p) \begin{bmatrix} Z_{cc} & Z_{ci} & Z_{cp} \\ Z_{ic} & Z_{ii} & Z_{ip} \\ Z_{pc} & Z_{pi} & Z_{pp} \end{bmatrix} \begin{bmatrix} Y_{cv} \\ Y_{iv} \\ Y_{pv} \end{bmatrix}$$

Substituting $\Gamma_i V_i = V_i^{\text{int}} + \gamma_i \otimes \mathbf{1} = Z_i I_i^{\text{int}} + \gamma_i \otimes \mathbf{1}$ from (16.39e) and (16.39f), $\Gamma_p V_p = V_p^{\text{int}} + \gamma_p \otimes \mathbf{1}$ from (16.44b), and re-arranging, we have

$$\begin{bmatrix} \mathbb{I}_c & Z_{ci}\Gamma_i^{\mathsf{T}} & Z_{cp}\Gamma_p^{\mathsf{T}} & 0 \\ 0 & \Gamma_i Z_{ii}\Gamma_i^{\mathsf{T}} + Z_i & \Gamma_i Z_{ip}\Gamma_p^{\mathsf{T}} & 0 \\ 0 & \Gamma_p Z_{pi}\Gamma_i^{\mathsf{T}} & \Gamma_p Z_{pp}\Gamma_p^{\mathsf{T}} & \mathbb{I}_p \end{bmatrix} \begin{bmatrix} V_c \\ I_i^{\text{int}} \\ I_p^{\text{int}} \\ V_p^{\text{int}} \end{bmatrix} \;=\; \begin{bmatrix} Z_{cc} \\ \Gamma_i Z_{ic} \\ \Gamma_p Z_{pc} \end{bmatrix} I_c \;-\; \begin{bmatrix} B_{cv} \\ B_{iv} \\ B_{pv} \end{bmatrix} V_v \;-\; \begin{bmatrix} 0 \\ \gamma_i \\ \gamma_p \end{bmatrix} \otimes \mathbf{1} \tag{16.47a}$$

$$\operatorname{diag}\left(V_p^{\text{int}}\, I_p^{\text{intH}}\right) \;=\; \sigma_p \tag{16.47b}$$

The reduced system of (16.43)(16.39)(16.44) is (16.47) which must be solved numerically. The analysis problem therefore becomes: Solve (16.47) for $\left(V_c, I_i^{\text{int}}, I_p^{\text{int}}, V_p^{\text{int}}\right)$ and derive all other variables analytically (Exercise 16.13). As before, the analysis problem does not assume C16.1 and therefore each line (j,k) may model a transmission or distribution line, or a three-phase transformer where its series admittance matrices y_{jk}^s and y_{kj}^s may be different.

We make three remarks. First, compared with the reduced system (16.42) without power sources, the reduced system (16.47) involves two more variables $\left(V_p^{\text{int}}, V_p^{\text{int}}\right)$ with two additional sets of equations. While (16.42) is linear, (16.47) is quadratic because of the device model (16.47b) of power sources. Even if the inverse in (16.46) exists and the matrix on the left-hand side of (16.47a) is invertible, (16.47) may or may not have a solution which may or may not be unique because of the nonlinearity. Second, these inverses may not exist in which case more information is needed to determine a solution. For example, when there are no voltage sources as in Example 16.9 and the shunt admittances $y_{jk}^m = y_{kj}^m = 0$, the admittance matrix in (16.45) has zero row sums and is singular. In that case additional information $(\beta_j + \beta_k)$ is given, compared with the case in Example 16.8; see also Remark 16.8. Finally, the linearity of (16.47a) is the consequence of using the linear current balance equation $I = YV$ in (16.43), and this is always possible as discussed in Remark 16.7.

16.3 Balanced network

In this section we show that, if the voltage sources, current sources, and impedances are generalized balanced vectors and the lines are decoupled, then the analysis problem in Chapter 16.2 can be solved by analyzing certain simpler per-phase networks. The intuition is that the balanced voltage and current sources render all voltages and currents in the network to be balanced due to Corollary 1.3. To simplify exposition we only consider the case without power sources so that our problem remains linear.

With today's abundant computing power the smaller problem size may not be an important advantage of per-phase analysis. Rather, per-phase analysis clarifies the simple structure underlying a balanced network and enhances our conceptual understanding of three-phase networks in general, balanced or unbalanced.

We start in Chapter 16.3.1 by summarizing properties of Kronecker product which underlies the equivalence of three-phase analysis and per-phase analysis for a balanced network.

16.3.1 Kronecker product

The simple structure that underlies balanced networks depends critically on properties of the Kronecker product. For instance the admittance matrix Y of a balanced three-phase network can be written as the Kronecker product of a per-phase admittance matrix and the identity matrix $\mathbb{I}$ of size 3. This is explained in Chapter 16.3. In particular we will use the following properties in the proof of Theorem **??** there.

Lemma 16.6 (Kronecker product)**.** Let A, B, C, D be complex matrices of appropriate dimensions.

1 $(A+B)\otimes C = (A\otimes C) + (B\otimes C)$; $C\otimes(A+B) = (C\otimes A) + (C\otimes B)$.
2 $(A\otimes B)(C\otimes D) = (AC)\otimes(BD)$.
3 $(A\otimes B)^{\mathsf{T}} = A^{\mathsf{T}}\otimes B^{\mathsf{T}}$; $(A\otimes B)^{\mathsf{H}} = A^{\mathsf{H}}\otimes B^{\mathsf{H}}$.
4 $(A\otimes B)^{-1} = A^{-1}\otimes B^{-1}$; $(A\otimes B)^{\dagger} = A^{\dagger}\otimes B^{\dagger}$ where $A^{\dagger}$ denotes the pseudo-inverse of A.
5 rank $(A\otimes B)$ = rank $A\cdot$ rank B.
6 If $A\in\mathbb{C}^{m\times n}$ is invertible and $X, Y\in\mathbb{C}^{p\times q}$ then

$$A\otimes X = A\otimes Y, \qquad \Longleftrightarrow \qquad X = Y$$

The proof of the lemma is left as Exercise **??**

16.3.2 Three-phase analysis

We first explain how the device models and the admittance matrix simplify in a balanced system. We then use that to simplify the three-phase analysis problem in Chapter 16.2. Finally we show that the problem is equivalent to solving per-phase systems.

Balanced devices.

When the devices are balanced positive-sequence sets with parameters $\lambda_j, \mu_j, \zeta_j\in\mathbb{C}$:

$$E_j^{Y/\Delta} := \lambda_j\alpha_+, \quad j\in N_v, \quad J_j^{Y/\Delta} := \mu_j\alpha_+, \quad j\in N_c, \quad z_j^{Y/\Delta} := \zeta_j\mathbb{I}, \quad j\in N_i$$

their internal models in Table 16.3 reduce to those specified in Table 16.4. In vector form the voltage sources are

$$E_v^Y = \lambda_v^Y\otimes\alpha_+, \qquad E_v^\Delta = \lambda_v^\Delta\otimes\alpha_+, \qquad E_v := \begin{bmatrix} E_v^Y \\ E_v^\Delta \end{bmatrix} = \lambda_v\otimes\alpha_+$$

where $\lambda_v^Y := \left(\lambda_j, j\in N_v^Y\right)$, $\lambda_v^\Delta := \left(\lambda_j, j\in N_v^\Delta\right)$ and $\lambda_v := \left(\lambda_j, j\in N_v\right)$. Defining similar quantities for current sources and impedances, the specification (16.39a)(16.39b) in

vector form reduces to

$$E_v := \begin{bmatrix} \lambda_v^Y \\ \lambda_v^\Delta \end{bmatrix} \otimes \alpha_+ = \lambda_v \otimes \alpha_+, \qquad \gamma_v := \begin{bmatrix} \gamma_v^Y \\ \gamma_v^\Delta \end{bmatrix} \tag{16.48a}$$

$$J_c := \begin{bmatrix} \mu_c^Y \\ \mu_c^\Delta \end{bmatrix} \otimes \alpha_+ = \mu_c \otimes \alpha_+, \qquad \gamma_c^0 := \begin{bmatrix} \gamma_c^Y \\ 0 \end{bmatrix} \tag{16.48b}$$

$$Z_i := \operatorname{diag}\left(\zeta_i^Y, \zeta_i^\Delta\right) \otimes \mathbb{I} = \zeta_i \otimes \mathbb{I}, \qquad \gamma_i^0 := \begin{bmatrix} \gamma_i^Y \\ 0 \end{bmatrix} \tag{16.48c}$$

where $\zeta_i^Y := \operatorname{diag}\left(\zeta_j, j \in N_i^Y\right)$, $\zeta_i^\Delta := \operatorname{diag}\left(\zeta_j, j \in N_i^\Delta\right)$, $\zeta_i := \operatorname{diag}\left(\zeta_i^Y, \zeta_i^\Delta\right)$ are diagonal matrices of sizes $\left|V_i^Y\right|$, $\left|V_i^\Delta\right|$, $|V_i|$ respectively.

The external models in Table 16.4 are obtained by substituting these specifications into the external models in Table 16.3 and applying Corollary 1.3 and Theorem 14.2, specifically

$$\Gamma\alpha_+ = (1-\alpha)\alpha_+, \quad \Gamma^\mathsf{T}\alpha_+ = \left(1-\alpha^2\right)\alpha_+, \quad \Gamma^\dagger = \frac{1}{3}\Gamma^\mathsf{T}, \quad \Gamma^{\mathsf{T}\dagger} = \frac{1}{3}\Gamma$$

The derivation of the impedance model in Table 16.4 in Δ configuration is left as Exercise 16.14. These models are special cases of the three-phase devices in Chapters

Buses j	Specification	External model	Vars	Internal vars
N_v^Y	$E_j^Y = \lambda_j\alpha_+,\ \gamma_j$	$V_j = \lambda_j\alpha_+ + \gamma_j\mathbf{1}$	I_j	$I_j^Y = -I_j$
N_v^Δ	$E_j^\Delta = \lambda_j\alpha_+,\ \gamma_j, \beta_j$	$V_j = \frac{1}{3}(1-\alpha^2)\lambda_j\alpha_+ + \gamma_j\mathbf{1}$	I_j	$I_j^\Delta = -\Gamma^{\mathsf{T}\dagger}I_j + \beta_j\mathbf{1}$
N_c^Y	$J_j^Y = \mu_j\alpha_+, \gamma_j$	$I_j = -\mu_j\alpha_+$	V_j	$V_j^Y = V_j - \gamma_j\mathbf{1}$
N_c^Δ	$J_j^\Delta = \mu_j\alpha_+$	$I_j = -(1-\alpha^2)\mu_j\alpha_+$	V_j	$V_j^\Delta = \Gamma V_j,\ \gamma_j := \frac{1}{3}\mathbf{1}^\mathsf{T}V_j$ $\beta_j := \frac{1}{3}\mathbf{1}^\mathsf{T}I_j^\Delta$
N_i^Y	$z_j^Y = \zeta_j\mathbb{I},\ \gamma_j$	$I_j = -\eta_j\left(V_j - \gamma_j\mathbf{1}\right)$	(V_j, I_j)	$V_j^Y = V_j - \gamma_j\mathbf{1},\ I_j^Y = -I_j$
N_i^Δ	$z_j^\Delta = \zeta_j\mathbb{I},\ \beta_j$	$I_j = -3\eta_j\left(V_j - \gamma_j\mathbf{1}\right)$	(V_j, I_j)	$V_j^\Delta = \Gamma V_j,\ \gamma_j := \frac{1}{3}\mathbf{1}^\mathsf{T}V_j$ $I_j^\Delta = -\Gamma^{\mathsf{T}\dagger}I_j + \beta_j\mathbf{1}$

Table 16.4 Internal and external models of balanced positive-sequence sources and impedances with $\eta_j := \zeta_j^{-1}$. The impedance model for N_i^Δ in the table is equivalent to $I_j = -3\eta_j\left(V_j - \left(\frac{1}{3}\mathbf{1}^\mathsf{T}V_j\right)\mathbf{1}\right)$ which is the model $I_j = -Y_j^\Delta V_j$ in Table 16.3.

14.3.3 and 14.3.4. To simplify the notation for the external models of voltage and current sources, define

$$\hat{\alpha}_j := \begin{cases} 1 & \text{if} \quad j \in N_v^Y \cup N_c^Y \cup N_i^Y \\ (1-\alpha^2)/3 & \text{if} \quad j \in N_v^\Delta \quad \text{(voltage sources)} \\ (1-\alpha^2) & \text{if} \quad j \in N_c^\Delta \quad \text{(current sources)} \\ 3 & \text{if} \quad j \in N_i^\Delta \quad \text{(admittance)} \end{cases}$$

Then when the voltage and current sources are balanced, their external models

(16.39c)(16.39d) reduce to:

$$V_v = \left(\hat{\alpha}_j \lambda_j \alpha_+ + \gamma_j \mathbf{1}, j \in N_v\right) =: \hat{\lambda}_v \otimes \alpha_+ + \gamma_v \otimes \mathbf{1} \tag{16.48d}$$

$$I_c = \left(-\hat{\alpha}_j \mu_j \alpha_+, j \in N_c\right) =: -\hat{\mu}_c \otimes \alpha_+ \tag{16.48e}$$

where $\hat{\lambda}_v, \gamma_v \in \mathbb{C}^{|N_v|}$ and $\hat{\mu}_c \in \mathbb{C}^{|N_c|}$.

Remark 16.9 (Δ-Y transformation)**.** The specification (16.48d)(16.48e) corresponds to the first step of per-phase analysis in Chapter 1.2.5 that converts all Δ configured devices to their Y equivalents that have the same external behavior. It generalizes the standard practice of assuming $\gamma_j = 0$ to the case where γ_j may be nonzero, because some Y-configured devices on the network are not grounded, some are grounded through nonzero earthing impedances, and some Δ-configured devices have nonzero zero-sequence voltages. □

The internal models of impedances (16.39e) and the conversion rules (16.39f) become

$$V_i^{\text{int}} = Z_i I_i^{\text{int}} = (\zeta_i \otimes \mathbb{I}) I_i^{\text{int}} \tag{16.48f}$$

$$I_i = \begin{bmatrix} -I_i^Y \\ -\left(\mathbb{I}_i^\Delta \otimes \Gamma_i^{\Delta\mathsf{T}}\right) I_i^\Delta \end{bmatrix} = -\Gamma_i^{\mathsf{T}} I_i^{\text{int}} \tag{16.48g}$$

$$\Gamma_i V_i = \begin{bmatrix} V_i^Y + \gamma_i^Y \otimes \mathbf{1} \\ V_i^\Delta \end{bmatrix} = V_i^{\text{int}} + \gamma_i^0 \otimes \mathbf{1} \tag{16.48h}$$

where Z_i, ζ_i, γ_i^0 are defined in (16.48c), and $\mathbb{I}_i^Y, \mathbb{I}_i^\Delta$ are the identity matrices of sizes $\left|V_i^Y\right|, \left|V_i^\Delta\right|$ respectively.

Balanced admittance matrix Y.

We assume all lines are balanced, i.e.,

$$y_{jk}^s = \eta_{jk}^s \mathbb{I}, \qquad y_{jk}^m = \eta_{jk}^m \mathbb{I}, \qquad y_{kj}^m = \eta_{kj}^m \mathbb{I} \tag{16.49a}$$

for some constants $\eta_{jk}^s, \eta_{jk}^m, \eta_{kj}^m \in \mathbb{C}$. The terminal voltages and currents $V := (V_0, \ldots, V_N)$ and $I := (I_0, \ldots, I_N)$ are described by (16.5) which, with balanced lines, reduces to

$$I_j = \sum_{k:j\sim k} \left(y_{jk}^s + y_{jk}^m\right) V_j - \sum_{k:j\sim k} y_{jk}^s V_k = \sum_{k:j\sim k} \eta_{jk} V_j - \sum_{k:j\sim k} \eta_{jk}^s V_k, \qquad j \in \overline{N} \tag{16.49b}$$

where $\eta_{jk} := \eta_{jk}^s + \eta_{jk}^m$ and $V_j, I_j \in \mathbb{C}^3$. This in vector form is $I = YV$. The balanced lines in (16.49a) allow us to write the admittance matrix Y using the Kronecker product. This is the key mathematical structure, in addition to the conversion matrices $\Gamma, \Gamma^{\mathsf{T}}$ as described in Corollary 1.3, that underlies the balanced property of all voltages and currents in the network.

Specifically, define the $(N+1)\times(N+1)$ *per-phase admittance matrix* $Y^{1\phi}$ by

$$Y^{1\phi}_{jk} := \begin{cases} -\eta^s_{jk}, & (j,k)\in E,\ (j\neq k) \\ \sum_{k:j\sim k}\left(\eta^s_{jk}+\eta^m_{jk}\right), & j=k \\ 0 & \text{otherwise} \end{cases} \tag{16.50a}$$

As we will see, this is the bus admittance matrix studied in Chapter 4.2 for the per-phase circuit of a balanced three-phase network where each line is characterized by four complex scalars $\left(\eta^s_{jk},\eta^m_{jk}\right)$, $\left(\eta^s_{kj},\eta^m_{kj}\right)$. In particular Y does not assume C16.1 and hence $Y^{1\phi}$ may not satisfy C4.1. Therefore each line (j,k) may model a transmission or distribution line, or a three-phase transformer where its series admittance matrices y^s_{jk} and y^s_{kj} may be different.

Substituting (16.49a) into the admittance matrix Y in (16.6) for the three-phase network, we can write Y in terms of the per-phase admittance matrix $Y^{1\phi}$ using the Kronecker product:

$$Y = Y^{1\phi}\otimes\mathbb{I} \tag{16.50b}$$

The relation $I=YV$ for the three-phase network becomes

$$I = \left(Y^{1\phi}\otimes\mathbb{I}\right)V \tag{16.50c}$$

Three-phase analysis.

We are interested in determining the (column) vectors of terminal and internal variables

$$V_{-v} := (V_c,V_i) := \left(V_j,\, j\in N_c\cup N_i\right), \qquad I_{-c} := (I_v,I_i) := \left(I_j,\, j\in N_c\cup N_i\right) \tag{16.51a}$$

$$V^{\text{int}}_{-v} := \left(V^{\text{int}}_c,V^{\text{int}}_i\right) := \left(V^{Y/\Delta}_j,\, j\in N_c\cup N_i\right), \quad I^{\text{int}}_{-c} := \left(I^{\text{int}}_v,I^{\text{int}}_i\right) := \left(I^{Y/\Delta}_j,\, j\in N_c\cup N_i\right) \tag{16.51b}$$

$$\gamma^\Delta_{-v} := \left(\gamma^\Delta_c,\gamma^\Delta_i\right) := \left(\gamma_j,\, j\in N^\Delta_c\cup N^\Delta_i\right), \qquad \beta^\Delta_{-v} := \left(\beta^\Delta_j,\beta^\Delta_j\right) := \left(\beta_j,\, j\in N^\Delta_c\cup N^\Delta_i\right) \tag{16.51c}$$

Let $x := \left(V_{-v},I_{-c},V^{\text{int}}_{-v},I^{\text{int}}_{-c},\gamma^\Delta_{-v},\beta^\Delta_{-v}\right)$. When the network is balanced the three-phase analysis problem in Chapter 16.2 reduces to: solve for x given the device specification (16.48) and the network equation (16.50).

16.3.3 Balanced voltages and currents

In this subsection we prove a structural result that says that, when the internal voltages and currents of non-power sources are balanced, so are all other voltages and currents in the network.

Partition the per-phase admittance matrix $Y^{1\phi}$ defined in (16.50) into submatrices (A_{11}, A_{21}, A_{22}):

$$Y^{1\phi} =: \left[\begin{array}{c|cc} Y_{vv}^{1\phi} & Y_{vc}^{1\phi} & Y_{vi}^{1\phi} \\ \hline Y_{cv}^{1\phi} & Y_{cc}^{1\phi} & Y_{ci}^{1\phi} \\ Y_{iv}^{1\phi} & Y_{ic}^{1\phi} & Y_{ii}^{1\phi} \end{array}\right] =: \left[\begin{array}{c|c} A_{11} & A_{21}^{\mathsf{T}} \\ \hline A_{21} & A_{22} \end{array}\right] \tag{16.52}$$

The matrix A_{22} is complex symmetric and therefore a legitimate admittance matrix. We will make two assumptions on the per-phase admittance matrix $Y^{1\phi}$.

C16.7: The submatrix A_{22} is invertible.

Assuming C16.7 (see Chapter 4.2.3 for sufficient conditions for the invertibility of principal submatrices of an admittance matrix), denote the inverse of the submatrix A_{22} by

$$\begin{bmatrix} Z_{cc}^{1\phi} & Z_{ci}^{1\phi} \\ Z_{ic}^{1\phi} & Z_{ii}^{1\phi} \end{bmatrix} := \begin{bmatrix} Y_{cc}^{1\phi} & Y_{ci}^{1\phi} \\ Y_{ic}^{1\phi} & Y_{ii}^{1\phi} \end{bmatrix}^{-1} = A_{22}^{-1} \tag{16.53a}$$

Then the inverse in (16.41) exists and is:

$$\begin{bmatrix} Z_{cc} & Z_{ci} \\ Z_{ic} & Z_{ii} \end{bmatrix} := \begin{bmatrix} Y_{cc} & Y_{ci} \\ Y_{ic} & Y_{ii} \end{bmatrix}^{-1} = A_{22}^{-1} \otimes \mathbb{I} \tag{16.53b}$$

where we have used $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$ (Lemma 16.6). The second assumption is:

C16.8: The impedances $\zeta_j \in \mathbb{C}$ are nonzero for all $j \in N_i$, the submatrix $Z_{ii}^{1\phi}$ in (16.53a) and the matrix

$$\hat{C}_i = \left(\left(Z_{ii}^{1\phi}\right)^{-1} \otimes \mathbb{I}\right) + \Gamma_i^{\mathsf{T}} \left(\zeta_i^{-1} \otimes \mathbb{I}\right) \Gamma_i \tag{16.54}$$

are invertible.

Theorem 16.7 (Balanced voltages and currents)**.** Suppose C16.7 and C16.8 hold.

1. Any solution x of (16.48)(16.50) consists of generalized balanced vectors in positive sequence, i.e., any voltage or current x_j in (16.51) at bus j is of the form $x_j = a_j \alpha_+ + b_j \mathbf{1}$ for some $a_j, b_j \in \mathbb{C}$.
2. Moreover all x_j are balanced vectors, i.e., $b_j = 0$, if $\gamma_v = 0$ for all voltage sources and the neutral voltages $\gamma_i^Y = 0$ for all Y configured impedances.

In the rest of this subsection we prove the theorem following the solution strategy in Chapter 16.2.3 to show that any solution (V_c, I_i^{int}) of the reduced system (16.42) consists of generalized balanced vectors. All other variables can then be derived analytically in terms of the solution (V_c, I_i^{int}) and shown to be generalized balanced vectors (Exercise 16.15).

The variable (V_c, I_i^{int}) satisfies (16.42), reproduced here:

$$\begin{bmatrix} \mathbb{I}_c \otimes \mathbb{I} & Z_{ci}\Gamma_i^{\mathsf{T}} \\ 0 & \Gamma_i Z_{ii}\Gamma_i^{\mathsf{T}} + Z_i \end{bmatrix} \begin{bmatrix} V_c \\ I_i^{\text{int}} \end{bmatrix} = \begin{bmatrix} Z_{cc} \\ \Gamma_i Z_{ic} \end{bmatrix} I_c - \begin{bmatrix} A_{cv} \\ A_{iv} \end{bmatrix} V_v - \begin{bmatrix} 0 \\ \gamma_i^0 \otimes \mathbf{1} \end{bmatrix} \tag{16.55}$$

where

$$\begin{bmatrix} A_{cv} \\ A_{iv} \end{bmatrix} := \operatorname{diag}(\mathbb{I}_c \otimes \mathbb{I}, \Gamma_i) \begin{bmatrix} Z_{cc} & Z_{ci} \\ Z_{ic} & Z_{ii} \end{bmatrix} \begin{bmatrix} Y_{cv} \\ Y_{iv} \end{bmatrix}$$

We now prove Theorem 16.7 in the following three lemmas. The first lemma simplifies (16.55) using balanced devices (16.48) and balanced lines (16.50).

Lemma 16.8. Suppose C16.7 holds. Balanced devices and lines (16.48)(16.50) reduces (16.55) to

$$\underbrace{\begin{bmatrix} \mathbb{I}_c \otimes \mathbb{I} & \left(Z_{ci}^{1\phi} \otimes \mathbb{I}\right)\Gamma_i^{\mathsf{T}} \\ 0 & \Gamma_i\left(Z_{ii}^{1\phi} \otimes \mathbb{I}\right)\Gamma_i^{\mathsf{T}} + (\zeta_i \otimes \mathbb{I}) \end{bmatrix}}_{M} \begin{bmatrix} V_c \\ I_i^{\text{int}} \end{bmatrix} = a' \otimes \alpha_+ + b' \otimes \mathbf{1} \tag{16.56a}$$

where

$$a' := -\begin{bmatrix} Z_{cc}^{1\phi}\hat{\mu}_c + B_{cv}\hat{\lambda}_v \\ Z_{ic}^{1\phi,Y}\hat{\mu}_c + B_{iv}^{Y}\hat{\lambda}_v \\ (1-\alpha)\left(Z_{ic}^{1\phi,\Delta}\hat{\mu}_c + B_{iv}^{\Delta}\hat{\lambda}_v\right) \end{bmatrix}, \qquad b' := -\begin{bmatrix} B_{cv}\,\gamma_v \\ B_{iv}^{Y}\gamma_v^{Y} + \gamma_i^{Y} \\ 0 \end{bmatrix} \tag{16.56b}$$

for some matrices $B_{cv}, B_{iv}^{Y}, B_{iv}^{\Delta}$.

The second lemma shows that the inverse M^{-1} of the matrix in (16.53a) has a structure that preserve the balanced nature of voltages and currents.

Lemma 16.9. Suppose C16.7 and C16.8 hold.

1 The matrix M in (16.56) is invertible.
2 Each 3×3 block $\left[M^{-1}\right]_{jk}$ of M^{-1} corresponding to phases abc is of the form

$$\left[M^{-1}\right]_{jk} := v_{jk}\,\mathbb{I} + w_{jk}\,W_{jk} \tag{16.57}$$

where $v_{jk}, w_{jk} \in \mathbb{C}$ are scalars and $W_{jk} \in \mathbb{C}^{3\times 3}$ is one of $\mathbb{I}$, Γ, Γ^{T}, $\Gamma\Gamma^{\mathsf{T}}$ and $\Gamma^{\mathsf{T}}\Gamma$.

The structure (16.57) of M^{-1} in Lemma 16.9 is what allows (V_c, V_i^{int}) to remain generalized balanced vectors. It requires that $\hat{C}_i$ in C16.8 be invertible. The following lemma is the crucial fact in determining the inverse of $\hat{C}_i$ that appears in M^{-1}. The

lemma can be verified directly using $\left(\Gamma^{\mathsf{T}}\Gamma\right)\left(\Gamma^{\mathsf{T}}\Gamma\right) = 3\Gamma^{\mathsf{T}}\Gamma$ (Theorem 14.2). It says that taking the inverse of the sum of a Kronecker product with $\mathbb{I}$ and a Kronecker product with $\Gamma^{\mathsf{T}}\Gamma$ preserves the Kronecker structure.

Lemma 16.10. For any matrix A and B of appropriate sizes, if A and $A+3B$ are invertible then

$$\left(A\otimes\mathbb{I} + B\otimes\Gamma^{\mathsf{T}}\Gamma\right)^{-1} \;=\; \left(A^{-1}\otimes\mathbb{I} - \left((A+3B)^{-1}BA^{-1}\right)\otimes\Gamma^{\mathsf{T}}\Gamma\right)$$

We now prove Lemmas 16.8 and 16.9.

Proof of Lemma 16.8 From (16.53) and (16.48c), the matrix on the left-hand side of (16.55) reduces to

$$\begin{bmatrix} \mathbb{I}_c\otimes\mathbb{I} & \left(Z_{ci}^{1\phi}\otimes\mathbb{I}\right)\Gamma_i^{\mathsf{T}} \\ 0 & \Gamma_i\left(Z_{ii}^{1\phi}\otimes\mathbb{I}\right)\Gamma_i^{\mathsf{T}} + (\zeta_i\otimes\mathbb{I}) \end{bmatrix} \tag{16.58}$$

On the right-hand side partition $Z_{ic}^{1\phi}$ in (16.53) into submatrices corresponding to impedances in Y and Δ configurations:

$$Z_{ic}^{1\phi} \;=:\; \begin{bmatrix} Z_{ic}^{1\phi,YY} & Z_{ic}^{1\phi,Y\Delta} \\ Z_{ic}^{1\phi,\Delta Y} & Z_{ic}^{1\phi,\Delta\Delta} \end{bmatrix} \;=:\; \begin{bmatrix} Z_{ic}^{1\phi,Y} \\ Z_{ic}^{1\phi,\Delta} \end{bmatrix}$$

where $Z_{ic}^{1\phi,Y}$ denotes the first $|N_i^Y|$ rows of $Z_{ic}^{1\phi}$ corresponding to Y configured impedances and $Z_{ic}^{1\phi,\Delta}$ denotes the remaining $|N_i^\Delta|$ rows of $Z_{ic}^{1\phi}$ corresponding to Δ configured impedances. We then have, using $\Gamma_i = \mathrm{diag}\left(\mathbb{I}_i^Y\otimes\mathbb{I}, \mathbb{I}_i^\Delta\otimes\Gamma\right)$,

$$\Gamma_i Z_{ic} \;=\; \Gamma_i\left(Z_{ic}^{1\phi}\otimes\mathbb{I}\right) \;=\; \begin{bmatrix} Z_{ic}^{1\phi,Y}\otimes\mathbb{I} \\ Z_{ic}^{1\phi,\Delta}\otimes\Gamma \end{bmatrix}$$

The important structure is that the conversion matrix Γ appears on the right as "$\otimes\Gamma$" which allows the current I_c transformed by $\Gamma_i Z_{ic}$ to remain in $\mathrm{span}(\alpha_+)$ on the right-hand side (using (16.48e)):

$$\begin{bmatrix} Z_{cc} \\ \Gamma_i Z_{ic} \end{bmatrix} I_c \;=\; -\begin{bmatrix} Z_{cc}^{1\phi}\otimes\mathbb{I} \\ \Gamma_i\left(Z_{ic}^{1\phi}\otimes\mathbb{I}\right) \end{bmatrix}\hat{\mu}_c\otimes\alpha_+ \;=\; -\begin{bmatrix} Z_{cc}^{1\phi}\hat{\mu}_c \\ Z_{ic}^{1\phi,Y}\hat{\mu}_c \\ (1-\alpha)Z_{ic}^{1\phi,\Delta}\hat{\mu}_c \end{bmatrix}\otimes\alpha_+ \tag{16.59a}$$

where we have used $(A\otimes B)(C\otimes D) \;=\; (AC)\otimes(BD)$ (Lemma 16.6) and $\Gamma\alpha_+ = (1-\alpha)\alpha_+$ (Corollary 1.3).

The second term $\begin{bmatrix} A_{cv} \\ A_{iv} \end{bmatrix} V_v$ on the right-hand side of (16.55) can be simplified in a similar manner but with more steps. We have from (16.53)

$$\begin{bmatrix} A_{cv} \\ A_{iv} \end{bmatrix} \;=\; \begin{bmatrix} Z_{cc}^{1\phi}\otimes\mathbb{I} & Z_{ci}^{1\phi}\otimes\mathbb{I} \\ \Gamma_i\left(Z_{ic}^{1\phi}\otimes\mathbb{I}\right) & \Gamma_i\left(Z_{ii}^{1\phi}\otimes\mathbb{I}\right) \end{bmatrix}\begin{bmatrix} Y_{cv}^{1\phi}\otimes\mathbb{I} \\ Y_{iv}^{1\phi}\otimes\mathbb{I} \end{bmatrix}$$

Similarly partition $Z_{ii}^{1\phi}$ into its first $|N_i^Y|$ and the remaining $|N_i^\Delta|$ rows:

$$Z_{ii}^{1\phi} \;=:\; \begin{bmatrix} Z_{ii}^{1\phi,Y} \\ Z_{ii}^{1\phi,\Delta} \end{bmatrix}$$

Then, using (16.48d) and $\Gamma_i = \operatorname{diag}\left(\mathbb{I}_i^Y \otimes \mathbb{I}, \mathbb{I}_i^\Delta \otimes \Gamma\right)$, we have

$$\begin{aligned} \begin{bmatrix} A_{cv} \\ A_{iv} \end{bmatrix} V_v \;&=\; \begin{bmatrix} Z_{cc}^{1\phi} \otimes \mathbb{I} & Z_{ci}^{1\phi} \otimes \mathbb{I} \\ Z_{ic}^{1\phi,Y} \otimes \mathbb{I} & Z_{ii}^{1\phi,Y} \otimes \mathbb{I} \\ Z_{ic}^{1\phi,\Delta} \otimes \Gamma & Z_{ii}^{1\phi,\Delta} \otimes \Gamma \end{bmatrix} \begin{bmatrix} Y_{cv}^{1\phi} \otimes \mathbb{I} \\ Y_{iv}^{1\phi} \otimes \mathbb{I} \end{bmatrix} \left(\hat{\lambda}_v \otimes \alpha_+ + \gamma_v \otimes \mathbf{1}\right) \\ &=:\; \underbrace{\begin{bmatrix} B_{cv}\,\hat{\lambda}_v \\ B_{iv}^Y \hat{\lambda}_v \\ (1-\alpha) B_{iv}^\Delta \hat{\lambda}_v \end{bmatrix}}_{a'} \otimes \alpha_+ \;+\; \underbrace{\begin{bmatrix} B_{cv}\,\gamma_v \\ B_{iv}^Y \gamma_v^Y \\ 0 \end{bmatrix}}_{b'} \otimes \mathbf{1} \end{aligned} \tag{16.59b}$$

where

$$B_{cv} \;:=\; Z_{cc}^{1\phi} Y_{cv}^{1\phi} + Z_{ci}^{1\phi} Y_{iv}^{1\phi}, \quad B_{iv}^Y \;:=\; Z_{ic}^{1\phi,Y} Y_{cv}^{1\phi} + Z_{ii}^{1\phi,Y} Y_{iv}^{1\phi}, \quad B_{cv}^\Delta \;:=\; Z_{ic}^{1\phi,\Delta} Y_{cv}^{1\phi} + Z_{ii}^{1\phi,\Delta} Y_{iv}^{1}$$

The factor $1-\alpha$ in (16.59b) is due to $\Gamma\alpha_+ = (1-\alpha)\alpha_+$ and the 0 entry is due to $\Gamma\mathbf{1} = 0$ and originates from the fact that the internal voltages in a Δ configuration sum to zero, i.e., $\mathbf{1}^\mathsf{T} V^\Delta = 0$.

Substituting (16.58)(16.59) into (16.55) then yields (16.56) (recall from (16.48c) that $\gamma_i^0 := \left(\gamma_i^Y, 0\right)$). □

Proof of Lemma 16.9 The matrix in (16.56):

$$M \;:=\; \begin{bmatrix} \mathbb{I}_c \otimes \mathbb{I} & \left(Z_{ci}^{1\phi} \otimes \mathbb{I}\right) \Gamma_i^\mathsf{T} \\ 0 & \Gamma_i \left(Z_{ii}^{1\phi} \otimes \mathbb{I}\right) \Gamma_i^\mathsf{T} + (\zeta_i \otimes \mathbb{I}) \end{bmatrix}$$

is invertible if its submatrix

$$M_{22} \;:=\; (\zeta_i \otimes \mathbb{I}) \;+\; \Gamma_i \left(Z_{ii}^{1\phi} \otimes \mathbb{I}\right) \Gamma_i^\mathsf{T} \tag{16.60a}$$

is invertible in which case its inverse is

$$M^{-1} \;:=\; \begin{bmatrix} \mathbb{I}_c \otimes \mathbb{I} & -\left(Z_{ci}^{1\phi} \otimes \mathbb{I}\right) \Gamma_i^\mathsf{T} M_{22}^{-1} \\ 0 & M_{22}^{-1} \end{bmatrix} \tag{16.60b}$$

(see Appendix A.3 for discussions on Schur complement for the inverse of general block matrices). To study the invertibility of M_{22} we use the matrix inversion formula (A.6):

$$(A + BCD)^{-1} \;=\; A^{-1} \;-\; A^{-1} \left(B \tilde{C}^{-1} D\right) A^{-1}$$

where $\tilde{C} := C^{-1} + DA^{-1}B$ in Appendix A.3.2. The matrix $A + BCD$ is invertible if A, C and $\tilde{C} := C^{-1} + DA^{-1}B$ are invertible. Therefore M_{22} in (16.60a) is invertible if (i)

the impedances $\zeta_j \in \mathbb{C}$ are nonzero for all $j \in N_i$; (ii) $Z_{ii}^{1\phi}$ is invertible; and (iii) the matrix $\hat{C}_i$ in (16.54) is invertible, as claimed in Lemma 16.9.

We now prove (16.57). To apply Lemma 16.10 to determine the inverse of $\hat{C}_i$, use $\Gamma_i = \text{diag}\left(\mathbb{I}_i^Y \otimes \mathbb{I}, \mathbb{I}_i^\Delta \otimes \Gamma\right)$ to get

$$\begin{aligned}\Gamma_i^\mathsf{T}\left(\zeta_i^{-1}\otimes\mathbb{I}\right)\Gamma_i &= \text{diag}\left(\mathbb{I}_i^Y\otimes\mathbb{I}, \mathbb{I}_i^\Delta\otimes\Gamma^\mathsf{T}\right)\text{diag}\left(\left(\zeta_i^Y\right)^{-1}\otimes\mathbb{I}, \left(\zeta_i^\Delta\right)^{-1}\otimes\mathbb{I}\right)\text{diag}\left(\mathbb{I}_i^Y\otimes\mathbb{I}, \mathbb{I}_i^\Delta\otimes\Gamma\right)\\ &= \text{diag}\left(\eta_i^Y\otimes\mathbb{I}, \eta_i^\Delta\otimes\Gamma^\mathsf{T}\Gamma\right)\end{aligned}$$

where $\eta_i^{Y/\Delta} := \left(\zeta_i^{Y/\Delta}\right)^{-1}$. Partition $\left(Z_{ii}^{1\phi}\right)^{-1}$ into submatrices:

$$\left(Z_{ii}^{1\phi}\right)^{-1} =: \begin{bmatrix} A^{YY} & A^{Y\Delta} \\ A^{\Delta Y} & A^{\Delta\Delta}\end{bmatrix}$$

Then

$$\hat{C}_i := \left(\left(Z_{ii}^{1\phi}\right)^{-1}\otimes\mathbb{I}\right) + \Gamma_i^\mathsf{T}\left(\zeta_i^{-1}\otimes\mathbb{I}\right)\Gamma_i = \begin{bmatrix}\tilde{A}^{YY} & A^{Y\Delta}\\ A^{\Delta Y} & A^{\Delta\Delta}\end{bmatrix}\otimes\mathbb{I} + \text{diag}\left(0, \eta_i^\Delta\right)\otimes\Gamma^\mathsf{T}\Gamma$$

where $\tilde{A}^{YY} := A^{YY} + \eta_i^Y$. We can then apply Lemma 16.10 to get

$$\hat{C}_i^{-1} = \tilde{A}\otimes\mathbb{I} - \tilde{B}\otimes\Gamma^\mathsf{T}\Gamma \tag{16.61}$$

where

$$\tilde{A} := \begin{bmatrix}A^{YY}+\eta_i^Y & A^{Y\Delta}\\ A^{\Delta Y} & A^{\Delta\Delta}\end{bmatrix}^{-1}, \quad \tilde{B} := \begin{bmatrix}A^{YY}+\eta_i^Y & A^{Y\Delta}\\ A^{\Delta Y} & A^{\Delta\Delta}+3\eta_i^\Delta\end{bmatrix}^{-1}\text{diag}\left(0,\eta_i^\Delta\right)\tilde{A}$$

Applying the matrix inversion formula with $\hat{C}_i^{-1}$ given by (16.61) we obtain the inverse of M_{22} in (16.60a) as

$$\begin{aligned}M_{22}^{-1} &= (\eta_i\otimes\mathbb{I}) - (\eta_i\otimes\mathbb{I})\Gamma_i\left(\tilde{A}\otimes\mathbb{I} - \tilde{B}\otimes\Gamma^\mathsf{T}\Gamma\right)\Gamma_i^\mathsf{T}(\eta_i\otimes\mathbb{I})\\ &= (\eta_i\otimes\mathbb{I}) - \left(\begin{bmatrix}\hat{A}^{YY}\otimes\mathbb{I} & \hat{A}^{Y\Delta}\otimes\Gamma^\mathsf{T}\\ \hat{A}^{\Delta Y}\otimes\Gamma & \hat{A}^{\Delta\Delta}\otimes\Gamma\Gamma^\mathsf{T}\end{bmatrix} - \begin{bmatrix}\hat{B}^{YY}\otimes\Gamma^\mathsf{T}\Gamma & 3\hat{B}^{Y\Delta}\otimes\Gamma^\mathsf{T}\\ 3\hat{B}^{\Delta Y}\otimes\Gamma & 3\hat{B}^{\Delta\Delta}\otimes\Gamma\Gamma^\mathsf{T}\end{bmatrix}\right)\end{aligned} \tag{16.62}$$

where

$$\begin{aligned}[\hat{A}/\hat{B}]^{YY} &:= \eta_i^Y[\tilde{A}/\tilde{B}]^{YY}\eta_i^Y, & [\hat{A}/\hat{B}]^{Y\Delta} &:= \eta_i^Y[\tilde{A}/\tilde{B}]^{Y\Delta}\eta_i^\Delta\\ [\hat{A}/\hat{B}]^{\Delta Y} &:= \eta_i^\Delta[\tilde{A}/\tilde{B}]^{\Delta Y}\eta_i^Y, & [\hat{A}/\hat{B}]^{\Delta\Delta} &:= \eta_i^\Delta[\tilde{A}/\tilde{B}]^{\Delta\Delta}\eta_i^\Delta\end{aligned}$$

and $\eta_i^{Y/\Delta} := \left(\zeta_i^{Y/\Delta}\right)^{-1}$, $\eta_i := \text{diag}\left(\eta_i^Y, \eta_i^\Delta\right)$. Therefore each 3×3 block of M_{22}^{-1} is of the desired form of $v_{jk}\mathbb{I} + w_{jk}W_{jk}$ where $v_{jk}, w_{jk} \in \mathbb{C}$ are scalars and $W_{jk} \in \mathbb{C}^{3\times3}$ is one of $\mathbb{I}$, Γ, Γ^T, $\Gamma\Gamma^\mathsf{T}$ and $\Gamma^\mathsf{T}\Gamma$.

Finally substituting (16.61) into (16.60b) we see that each 3×3 block of the $3(|N_c|+|N_i|)\times3(|N_c|+|N_i|)$ matrix M^{-1} will also be of the desired form of $w_{jk}W_{jk}$

if this property holds for its off-diagonal submatrix $\left(Z_{ci}^{1\phi} \otimes \mathbb{I}\right) \Gamma_i^{\mathsf{T}} M_{22}^{-1}$. We now show that this is indeed the case. Partition $Z_{ci}^{1\phi}$ into submatrices corresponding to impedances in Y and Δ configurations:

$$Z_{ci}^{1\phi} \;=:\; \begin{bmatrix} Z_{ci}^{1\phi,YY} & Z_{ci}^{1\phi,Y\Delta} \\ Z_{ci}^{1\phi,\Delta Y} & Z_{ci}^{1\phi,\Delta\Delta} \end{bmatrix}$$

Using $Z_{ci}^{1\phi}$ and $\Gamma_i = \operatorname{diag}\left(\mathbb{I}_i^Y \otimes \mathbb{I}, \mathbb{I}_i^\Delta \otimes \Gamma\right)$ we have

$$\left(Z_{ci}^{1\phi} \otimes \mathbb{I}\right) \Gamma_i^{\mathsf{T}} M_{22}^{-1} \;=\; \begin{bmatrix} Z_{ci}^{1\phi,YY} \otimes \mathbb{I} & Z_{ci}^{1\phi,Y\Delta} \otimes \Gamma^{\mathsf{T}} \\ Z_{ci}^{1\phi,\Delta Y} \otimes \mathbb{I} & Z_{ci}^{1\phi,\Delta\Delta} \otimes \Gamma^{\mathsf{T}} \end{bmatrix} M_{22}^{-1}$$

Substituting M_{22}^{-1} in (16.62) and using

$$\Gamma^{\mathsf{T}} \Gamma \Gamma^{\mathsf{T}} \;=\; \left(3\mathbb{I} - \mathbf{1}\mathbf{1}^{\mathsf{T}}\right) \Gamma^{\mathsf{T}} \;=\; 3\Gamma^{\mathsf{T}}$$

we see that each 3×3 block of $\left(Z_{ci}^{1\phi} \otimes \mathbb{I}\right) \Gamma_i^{\mathsf{T}} M_{22}^{-1}$ is of the desired form of $v_{jk}\mathbb{I} + w_{jk} W_{jk}$ where $v_{jk}, w_{jk} \in \mathbb{C}$ are scalars and $W_{jk} \in \mathbb{C}^{3\times 3}$ is one of $\mathbb{I}$, Γ, Γ^{T}, $\Gamma\Gamma^{\mathsf{T}}$ and $\Gamma^{\mathsf{T}}\Gamma$.

This completes the proof of (16.57). □

Lemmas 16.8 and 16.9 imply Theorem 16.7.

Proof of Theorem 16.7 Multiplying both sides of (16.56) by M^{-1} in (16.57) we see that the jth 3×3 block of $\left(V_c, I_i^{\text{int}}\right)$ is of the form

$$\sum_k \left[M^{-1}\right]_{jk} \left(a'_k \alpha_+ + b'_k \mathbf{1}\right) \;=\; \sum_k a'_k \left(v_{jk}\mathbb{I} + w_{jk} W_{jk}\right) \alpha_+ \;+\; \sum_k b'_k \left(v_{jk}\mathbb{I} + w_{jk} W_{jk}\right) \mathbf{1}$$

Since

$$W_{jk}\alpha_+ \;=\; \begin{cases} \alpha_+ & \text{if} \quad W_{jk} = \mathbb{I} \\ (1-\alpha)\alpha_+ & \text{if} \quad W_{jk} = \Gamma \\ (1-\alpha^2)\alpha_+ & \text{if} \quad W_{jk} = \Gamma^{\mathsf{T}} \\ 3\alpha_+ & \text{if} \quad W_{jk} = \Gamma\Gamma^{\mathsf{T}} \text{ or } \Gamma^{\mathsf{T}}\Gamma \end{cases}$$

and $W_{jk}\mathbf{1} = \mathbf{1}$ if $W_{jk} = \mathbb{I}$ and 0 otherwise, $\left(V_c, I_i^{\text{int}}\right)$ consists of generalized balanced vectors of the form $a_j\alpha_+ + b_j\mathbf{1}$. When $\gamma_v = 0$ for all voltage sources and $\gamma_i^Y = 0$ for all Y configured impedances, then $b' = 0$ in (16.56) and hence $b = 0$. This completes the proof of Theorem 16.7. □

16.3.4 Phase decoupling and per-phase analysis

In this subsection we show that phases in a balanced network are decoupled so that the three-phase analysis problem can be solved by solving two per-phase networks.

Substitute the per-phase admittance matrix (16.52), and the external models of

voltage and current sources (16.48d)(16.48e) into the current balance equation $I = YV$ (16.50c) to get

$$\begin{bmatrix} I_v \\ -\hat{\mu}_c \otimes \alpha_+ \\ I_i \end{bmatrix} = \left(\begin{bmatrix} Y_{vv}^{1\phi} & Y_{vc}^{1\phi} & Y_{vi}^{1\phi} \\ Y_{cv}^{1\phi} & Y_{cc}^{1\phi} & Y_{ci}^{1\phi} \\ Y_{iv}^{1\phi} & Y_{ic}^{1\phi} & Y_{ii}^{1\phi} \end{bmatrix} \otimes \mathbb{I} \right) \begin{bmatrix} \hat{\lambda}_v \otimes \alpha_+ + \gamma_v \otimes \mathbf{1} \\ V_c \\ V_i \end{bmatrix} \tag{16.63}$$

Instead of following the solution strategy of Chapter 16.2.3 to compute the internal impedance current I_i^{int} from the reduced system (16.42) we will compute the terminal voltage V_i, as well as V_c, using (16.63). We can then compute (I_v, I_i) and all other variables such as internal voltages and currents and zero-sequence voltages and currents.

We know from Theorem 16.7 that all voltages and currents consist of generalized balanced vectors of the form $a_j \alpha_+ + b_j \mathbf{1}$. We now describe separately external models for devices in Δ and Y configurations.

Δ configuration.

Consider a Δ configured device $j \in N_v^\Delta \cup N_c^\Delta \cup N_i^\Delta$. Let

$$V_j =: v_j \alpha_+ + \gamma_j \mathbf{1}, \qquad j \in N_c^\Delta \cup N_i^\Delta \tag{16.64a}$$

$$I_j =: i_j \alpha_+, \qquad j \in N_v^\Delta \cup N_i^\Delta \tag{16.64b}$$

for some (v_j, γ_j) and i_j to be determined. Here $\gamma_j = \frac{1}{3}\mathbf{1}^\mathsf{T} V_j$ is the zero-sequence voltage of V_j. As expected, $\mathbf{1}^\mathsf{T} I_j = 0$ since $I_j = -\Gamma^\mathsf{T} I_j^\Delta$. For an impedance $j \in N_i^\Delta$, we can express its terminal current I_j in terms of its terminal voltage V_j using its external model (from Table 16.4)

$$I_j = -3\eta_j \left(V_j - \gamma_j \mathbf{1} \right) = -3\eta_j v_j \alpha_+, \qquad j \in N_i^\Delta \tag{16.64c}$$

Hence the variables (v_j, i_j) for an impedance $j \in N_i^\Delta$ satisfies $i_j = -3\eta_j v_j$, the negative sign due to the definition of I_j being injection from the device to the rest of the network.

Y configuration.

Consider a Y configured device $j \in N_v^Y \cup N_c^Y \cup N_i^Y$. Let its internal voltage and internal current be generalized balanced vectors:

$$V_j^Y =: v_j^{\text{int}} \alpha_+ + \gamma_j^{\text{int}} \mathbf{1}, \qquad j \in\in N_c^Y \cup N_i^Y$$

$$I_j^Y =: -\left(i_j^{\text{int}} \alpha_+ + \beta_j^{\text{int}} \mathbf{1} \right), \qquad j \in N_v^Y \cup N_i^Y$$

for some $(v_j^{\text{int}}, \gamma_j^{\text{int}})$ and $\left(i_j^{\text{int}}, \beta_j^{\text{int}}\right)$ to be determined. Here $\gamma_j^{\text{int}} := \frac{1}{3}\mathbf{1}^\mathsf{T} V_j^Y$ is the zero-sequence voltage of the *internal* voltage V_j^Y, not the neutral voltage $\gamma_j := V_j^n$, and $\beta_j^{\text{int}} :=$

$\frac{1}{3}\mathbf{1}^\mathsf{T} I_j^Y$ is the zero-sequence current of the *internal* current I_j^Y. Since $V_j = V_j^Y + V_j^n \mathbf{1}$ and $I_j = -I_j^Y$, the terminal voltage and current are:

$$V_j =: v_j^{\text{int}} \alpha_+ + \left(\gamma_j^{\text{int}} + \gamma_j\right)\mathbf{1}, \qquad j \in N_c^Y \cup N_i^Y \tag{16.65a}$$

$$I_j =: i_j^{\text{int}} \alpha_+ + \beta_j^{\text{int}} \mathbf{1}, \qquad j \in N_v^Y \cup N_i^Y \tag{16.65b}$$

Recall that the neutral voltages $\gamma_j := V_j^n$ are given for all Y configured devices. The zero-sequence voltage of the terminal voltage V_j is the sum of the zero-sequence voltage γ_j^{int} of the internal voltage V_j^Y and the neutral voltage γ_j. Hence the terminal voltage V_j is balanced if and only if the neutral voltage γ_j is offset by γ_j^{int} so that $\gamma_j^{\text{int}} + \gamma_j = 0$ (see below for a sufficient condition). Moreover $\mathbf{1}^\mathsf{T} I_j = -\mathbf{1}^\mathsf{T} I_j^Y = -I_j^n$ is the negative of the neutral current. Hence $\beta_j^{\text{int}} = \frac{1}{3}\mathbf{1}^\mathsf{T} I_j = 0$ if device j has no neutral line. For an impedance $j \in N_i^Y$, we can express its terminal current I_j in terms of its terminal voltage V_j using the external model (from Table 16.4 and (16.65a))

$$I_j = -\eta_j \left(V_j - \gamma_j \mathbf{1}\right) = -\eta_j \left(v_j^{\text{int}} \alpha_+ + \gamma_j^{\text{int}} \mathbf{1}\right), \qquad j \in N_i^Y \tag{16.65c}$$

Hence $i_j^{\text{int}} = -\eta_j v_j^{\text{int}}$ and $\beta_j^{\text{int}} = -\eta_j \gamma_j^{\text{int}}$.

Before substituting (16.64)(16.65) into the network equation (16.63) we unify notations by defining

$$\hat{v}_j := \begin{cases} v_j^{\text{int}}, \\ v_j, \end{cases} \qquad \hat{\gamma}_j := \begin{cases} \gamma_j^{\text{int}} + \gamma_j, & j \in N_c^Y \cup N_i^Y \\ \gamma_j, & j \in N_c^\Delta \cup N_i^\Delta \end{cases} \tag{16.66a}$$

$$\hat{i}_j := \begin{cases} i_j^{\text{int}}, \\ i_j, \end{cases} \qquad \hat{\beta}_j := \begin{cases} \beta_j^{\text{int}}, & j \in N_v^Y \cup N_i^Y \\ 0, & j \in N_v^\Delta \cup N_i^\Delta \end{cases} \tag{16.66b}$$

Even though $\gamma_j = V_j^n$ are given for $j \in N_c^Y \cup N_i^Y$, γ_j^{int} (as well as $\gamma_j := \frac{1}{3}\mathbf{1}^\mathsf{T} V_j$ for $j \in N_c^\Delta \cup N_i^\Delta$) are unknown, and hence $\hat{\gamma}_j$ is unknown for $j \in N_c \cup N_i$. Therefore all the quantities in (16.66a) (16.66b) are to be determined. Collect currents and voltages associated with voltage and current sources respectively into

$$\hat{i}_v := \left(\hat{i}_j, j \in N_v\right), \quad \hat{\beta}_v := \left(\hat{\beta}_j, j \in N_v\right), \quad \hat{v}_c := \left(\hat{v}_j, j \in N_c\right), \quad \hat{\gamma}_c := \left(\hat{\gamma}_j, j \in N_c\right) \tag{16.66c}$$

Collect currents and voltages associated with impedances into

$$\hat{i}_i := \left(\hat{i}_j, j \in N_i\right), \quad \hat{\beta}_i := \left(\hat{\beta}_j, j \in N_i\right), \quad \hat{v}_i := \left(\hat{v}_j, j \in N_i\right), \quad \hat{\gamma}_i := \left(\hat{\gamma}_j, j \in N_i\right) \tag{16.66d}$$

Using the same notation for $\hat{\alpha}_j$ as in (16.48d)(16.48e), we can apply (16.66) to the external impedance models (16.65c) and (16.64c) to relate $\hat{v}_i$ and $\hat{i}_i$:

$$\hat{i}_i \otimes \alpha_+ + \hat{\beta}_i \otimes \mathbf{1} = -\left(\hat{\eta}_i \otimes \mathbb{I}\right)\left(\hat{v}_i \otimes \alpha_+ + (\hat{\gamma}_i - \gamma_i) \otimes \mathbf{1}\right) \tag{16.67a}$$

where the diagonal matrix $\hat{\eta}_i \in \mathbb{C}^{|N_i|\times|N_i|}$ and the vector $\gamma_i \in \mathbb{C}^{|N_i|}$ are defined as

$$\hat{\eta}_i \;:=\; \mathrm{diag}\left(\hat{\alpha}_j \eta_j, j \in N_i\right), \quad \gamma_i \;:=\; \begin{bmatrix}\gamma_i^Y \\ \gamma_i^\Delta\end{bmatrix} \;:=\; \begin{bmatrix}\left(\gamma_j := V_j^n, j \in N_i^Y\right) \\ \left(\gamma_j := \frac{1}{3}\mathbf{1}^\mathsf{T} V_j, j \in N_i^\Delta\right)\end{bmatrix} \tag{16.67b}$$

Hence $\hat{\gamma}_i - \gamma_i = \begin{bmatrix}\gamma_i^{\mathrm{int}} \\ 0\end{bmatrix}$ with $\gamma_i^{\mathrm{int}} := \left(\gamma_j^{\mathrm{int}}, j \in N_i^Y\right)$. Note the difference between γ_i defined here and the specification $\gamma_i^0 := \begin{bmatrix}\gamma_i^Y \\ 0\end{bmatrix}$ defined in (16.48c). Recall that γ_i^Y is given, but γ_i^{int} and hence $\hat{\gamma}_i$ are to be determined.

Substituting (16.66) into (16.63) we have

$$\begin{bmatrix}\hat{\imath}_v \\ -\hat{\mu}_c \\ \hat{\imath}_i\end{bmatrix} \otimes \alpha_+ + \begin{bmatrix}\hat{\beta}_v \\ 0 \\ \hat{\beta}_i\end{bmatrix} \otimes \mathbf{1} \;=\; \left(\begin{bmatrix}Y_{vv}^{1\phi} & Y_{vc}^{1\phi} & Y_{vi}^{1\phi} \\ Y_{cv}^{1\phi} & Y_{cc}^{1\phi} & Y_{ci}^{1\phi} \\ Y_{iv}^{1\phi} & Y_{ic}^{1\phi} & Y_{ii}^{1\phi}\end{bmatrix} \otimes \mathbb{I}\right)\left(\begin{bmatrix}\hat{\lambda}_v \\ \hat{v}_c \\ \hat{v}_i\end{bmatrix} \otimes \alpha_+ + \begin{bmatrix}\gamma_v \\ \hat{\gamma}_c \\ \hat{\gamma}_i\end{bmatrix} \otimes \mathbf{1}\right) \tag{16.68}$$

where the voltage sources $\hat{\lambda}_v$, current sources $-\hat{\mu}_c$, as well as $\left(\gamma_v, \gamma_c^0, \gamma_i^0\right)$ are given, and $\left(\hat{v}_{-v}, \hat{\gamma}_{-v}, \hat{\imath}_{-c}, \hat{\beta}_{-c}\right)$ are variables to be determined. Since α_+ and $\mathbf{1}$ are orthogonal this induces two sets of equations that can be interpreted as two per-phase networks.

Positive-sequence per-phase network.

Equating the α_+ coordinates on both sides of (16.68) the per-phase variables must satisfy

$$\begin{bmatrix}\hat{\imath}_v \\ -\hat{\mu}_c \\ \hat{\imath}_i\end{bmatrix} \;=\; \begin{bmatrix}Y_{vv}^{1\phi} & Y_{vc}^{1\phi} & Y_{vi}^{1\phi} \\ Y_{cv}^{1\phi} & Y_{cc}^{1\phi} & Y_{ci}^{1\phi} \\ Y_{iv}^{1\phi} & Y_{ic}^{1\phi} & Y_{ii}^{1\phi}\end{bmatrix}\begin{bmatrix}\hat{\lambda}_v \\ \hat{v}_c \\ \hat{v}_i\end{bmatrix} \tag{16.69a}$$

This defines the following per-phase network:

- The admittance matrix is $Y^{1\phi}$.
- The voltage sources have given voltages $\hat{\lambda}_v$.
- The current sources have given currents $-\hat{\mu}_c$.
- The impedances are $\hat{\eta}_i$ so that (from (16.67a))

$$\hat{\imath}_i \;=\; -\hat{\eta}_i \hat{v}_i \tag{16.69b}$$

This is a system of 4 sets of equations in 4 sets of variables $\left(\hat{v}_c, \hat{v}_i, \hat{\imath}_v, \hat{\imath}_i\right)$. Substituting (16.69b) into (16.69a) we obtain

$$\begin{bmatrix}Y_{cc}^{1\phi} & Y_{ci}^{1\phi} \\ Y_{ic}^{1\phi} & Y_{ii}^{1\phi} + \hat{\eta}_i\end{bmatrix}\begin{bmatrix}\hat{v}_c \\ \hat{v}_i\end{bmatrix} \;=\; -\left(\begin{bmatrix}\hat{\mu}_c \\ 0\end{bmatrix} + \begin{bmatrix}Y_{cv}^{1\phi} \\ Y_{iv}^{1\phi}\end{bmatrix}\hat{\lambda}_v\right) \tag{16.70}$$

If the matrix on the left-hand side is invertible then $(\hat{v}_c, \hat{v}_i)$ can be uniquely determined. The other variables $\left(\hat{\imath}_v, \hat{\imath}_i\right)$ can then be derived in terms of a solution $(\hat{v}_c, \hat{v}_i)$.

Zero-sequence per-phase network.

Equating the **1** coordinates in (16.68) the per-phase variables must satisfy

$$\begin{bmatrix} \hat{\beta}_v \\ 0 \\ \hat{\beta}_i \end{bmatrix} = \begin{bmatrix} Y^{1\phi}_{vv} & Y^{1\phi}_{vc} & Y^{1\phi}_{vi} \\ Y^{1\phi}_{cv} & Y^{1\phi}_{cc} & Y^{1\phi}_{ci} \\ Y^{1\phi}_{iv} & Y^{1\phi}_{ic} & Y^{1\phi}_{ii} \end{bmatrix} \begin{bmatrix} \gamma_v \\ \hat{\gamma}_c \\ \hat{\gamma}_i \end{bmatrix} \tag{16.71a}$$

This defines the following per-phase network:

- The network is described by the admittance matrix is $Y^{1\phi}$.
- The voltage sources have given voltages γ_v.
- The current sources inject 0 currents, i.e., no device is connected at buses j of the zero-sequence per-phase network where three-phase current sources are connected in the original network.
- The impedances are $\hat{\eta}_i$ so that (from (16.67a))

$$\hat{\beta}_i = -\hat{\eta}_i (\hat{\gamma}_i - \gamma_i) = -\mathrm{diag}\left(\hat{\eta}^Y_i, 0\right) \begin{bmatrix} (\hat{\gamma}^Y_i - \gamma^Y_i) \\ 0 \end{bmatrix} \tag{16.71b}$$

  where $\hat{\eta}^Y_i := \mathrm{diag}\left(\eta_j, j \in N^Y_i\right)$, $\hat{\gamma}^Y_i := \left(\hat{\gamma}_j, j \in N^Y_i\right)$ and $\gamma^Y_i := \left(V^n_j, j \in N^Y_i\right)$. Note that γ^Y_i is given and $\hat{\gamma}^Y_i$ is unknown.

This is a system of 4 sets of equations in 4 sets of variables $\left(\hat{\gamma}_c, \hat{\gamma}_i, \hat{\beta}_v, \hat{\beta}_i\right)$. Substituting (16.71b) into (16.71a) we obtain

$$\begin{bmatrix} Y^{1\phi}_{cc} & Y^{1\phi}_{ci} \\ Y^{1\phi}_{ic} & Y^{1\phi}_{ii} + \mathrm{diag}\left(\hat{\eta}^Y_i, 0\right) \end{bmatrix} \begin{bmatrix} \hat{\gamma}_c \\ \hat{\gamma}_i \end{bmatrix} = -\begin{bmatrix} Y^{1\phi}_{cv} \\ Y^{1\phi}_{iv} \end{bmatrix} \gamma_v + \begin{bmatrix} 0 \\ \hat{\eta}_i \gamma^0_i \end{bmatrix} \tag{16.72}$$

where we recall $\hat{\eta}_i$ in (16.67) and the given neutral voltages $\gamma^0_i := \begin{bmatrix} \gamma^Y_i \\ 0 \end{bmatrix}$. If the matrix on the left-hand side is invertible then $(\hat{\gamma}_c, \hat{\gamma}_i)$ can be uniquely determined. The other variables $\left(\hat{\beta}_v, \hat{\beta}_i\right)$ can then be derived in terms of a solution $(\hat{\gamma}_c, \hat{\gamma}_i)$.

Assume the matrix in (16.72) is invertible. If $\gamma_v = 0$ and $\gamma^Y_i = 0$ as in Theorem 16.7.2, then $\hat{\gamma}_c = 0$ and $\hat{\gamma}_i = 0$ and all voltages consist of balanced vectors. In this case we do not have to compute the zero-sequence network but simply set $\hat{\gamma}_{-v} := 0$ and $\hat{\beta}_{-c} := 0$. Recall from (16.66a)(16.66b) that this means $\gamma^{\mathrm{int}}_j + V^n_j = 0$ and $\beta^{\mathrm{int}}_j = 0$ for Y configured devices and $\gamma_j = 0$ for Δ configured devices.

Note that, even though $\hat{\beta}_{-c}$ is determined from (16.72) (16.71), its components $\hat{\beta}_j = 0$ for $j \in N^\Delta_v \cup N^\Delta_i$ from (16.66b). This is consistent because, for $j \in N^\Delta_v \cup N^\Delta_i$, multiplying both sides of (16.49b) by $\mathbf{1}^\mathsf{T}$ gives, using $\gamma_j := \frac{1}{3}\mathbf{1}^\mathsf{T} V_j$,

$$\sum_{k:j\sim k} \left(y^s_{jk} + y^m_{jk}\right)\gamma_j - \sum_{k:j\sim k} y^s_{jk}\gamma_k = 0$$

which is (16.71) for rows corresponding to $j \in N^\Delta_v \cup N^\Delta_i$.

Per-phase analysis.

Per-phase analysis for solving (16.63) is as follows:

1 Solve the positive-sequence per-phase network (16.70) for $(\hat{v}_c, \hat{v}_i)$ and then derive $(\hat{i}_v, \hat{i}_i)$.
2 If $\gamma_v = 0$ and $\gamma_i^Y = 0$, set $\hat{\gamma}_{-v} := 0$, $\hat{\beta}_{-c} := 0$, and goto the next step. Otherwise, solve the zero-sequence per-phase network (16.72) for $(\hat{\gamma}_c, \hat{\gamma}_i)$ and then derive $(\hat{\beta}_v, \hat{\beta}_i)$.
3 Substitute into (16.64)(16.65) to obtain (V_{-v}, I_{-c}).

Example 16.12 ($\gamma^Y = 0$)**.** Explain per-phase analysis in the special case where all neutrals are grounded with zero neutral impedances and voltages are defined with respect to the ground, i.e., $\gamma_j = 0$ for $j \in N_v^Y \cup N_c^Y \cup N_i^Y$. □

16.4 Symmetric network

We have formulated a general three-phase analysis problem in Chapter 16.2.2 and described a solution strategy in Chapter 16.2.3. When the network is balanced, the phases are decoupled and the network decomposes into two independent per-phase networks and the problem can be solved using per-phase analysis as explained in Chapter 16.3.

When the network is not balanced, e.g., the sources are unbalanced or the transmission lines are not phase-decoupled, then we can apply the similarity transformation F defined in Chapter 14.2.2 to transform *terminal* phase voltage and current (V, I) into sequence voltage and current $(\tilde{V}, \tilde{I})$. Even though the phases are coupled, we show in Chapters 16.4.1–16.4.4 that if three-phase lines are symmetric and loads are identical, then their external models are decoupled in the sequence coordinate. They define sequence networks that can be analyzed separately, similar to the per-phase networks of a balanced network studied in Chapter 16.3. The results from analyzing the sequence networks can then be transformed back to the original phase coordinate. We describe in Chapter 16.4.5 how to compose the sequence networks from the sequence models of individual devices and how to solve the three-phase analysis problem using these decoupled sequence networks when the original network is symmetric.

Symmetric components and sequence networks are most useful for fault analysis in a system that is more or less balanced, e.g., a three-phase network that remains balanced until the fault location. Without any symmetry, symmetrical components may not offer much advantage because they do not lead to decoupled sequence networks. Even though we do not study fault analysis in this book, the discussion in this section illustrates the application of various three-phase models developed in this chapter.

16.4.1 Sequence impedances

Y configuration (z^Y, z^n).

Consider the four-wire three-phase impedance (z^Y, z^n) in Y configuration shown in Figure 14.7 of Chapter 14.3.3. Under assumption C14.1 (all neutrals are grounded and all voltages are defined with respect to the ground), recall the external model (14.19b) relating the terminal voltage and current (V, I):

$$V = -Z^Y I \quad \text{with} \quad Z^Y := z^Y + z^n \mathbf{1}\mathbf{1}^\mathsf{T} = \begin{bmatrix} z^{an}+z^n & z^n & z^n \\ z^n & z^{an}+z^n & z^n \\ z^n & z^n & z^{cn}+z^n \end{bmatrix}$$

Substitute $V = F\tilde{V}$ and $I = F\tilde{I}$ to obtain the external model in the sequence coordinate:

$$\tilde{V} = -\underbrace{\overline{F} Z^Y F}_{\tilde{Z}^Y} \tilde{I} = -\tilde{Z}^Y \tilde{I}$$

where F from (14.6b) and its inverse $F^{-1} = \overline{F}$ from (14.7) are

$$F = \frac{1}{\sqrt{3}} \begin{bmatrix} \mathbf{1} & \alpha_+ & \alpha_- \end{bmatrix} = \frac{1}{\sqrt{3}} \begin{bmatrix} \mathbf{1}^\mathsf{T} \\ \alpha_+^\mathsf{T} \\ \alpha_-^\mathsf{T} \end{bmatrix} := \frac{1}{\sqrt{3}} \begin{bmatrix} 1 & 1 & 1 \\ 1 & \alpha & \alpha^2 \\ 1 & \alpha^2 & \alpha \end{bmatrix} \tag{16.73a}$$

$$\overline{F} = \frac{1}{\sqrt{3}} \begin{bmatrix} \mathbf{1} & \alpha_- & \alpha_+ \end{bmatrix} = \frac{1}{\sqrt{3}} \begin{bmatrix} \mathbf{1}^\mathsf{T} \\ \alpha_-^\mathsf{T} \\ \alpha_+^\mathsf{T} \end{bmatrix} := \frac{1}{\sqrt{3}} \begin{bmatrix} 1 & 1 & 1 \\ 1 & \alpha^2 & \alpha \\ 1 & \alpha & \alpha^2 \end{bmatrix} \tag{16.73b}$$

We call $\tilde{Z}^Y$ a *sequence impedance matrix* to differentiate it from the (phase) impedance matrix Z^Y. Substituting $Z^Y = z^Y + z^n \mathbf{1}\mathbf{1}^\mathsf{T}$, F and $\overline{F}$, we have (Exercise 16.18)

$$\tilde{Z}^Y = \frac{1}{3} \begin{bmatrix} \mathbf{1}^T z & \alpha_+^\mathsf{T} z & \alpha_-^\mathsf{T} z \\ \alpha_-^\mathsf{T} z & \mathbf{1}^T z & \alpha_+^\mathsf{T} z \\ \alpha_+^\mathsf{T} z & \alpha_-^\mathsf{T} z & \mathbf{1}^T z \end{bmatrix} + \begin{bmatrix} 3z^n & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

where $z := (z^{an}, z^{bn}, z^{cn})$ is the column vector of phase impedances. Hence the neutral impedance z^n appears only in the zero-sequence impedance.

If the impedance is balanced $z^{an} = z^{bn} = z^{cn}$, then $\mathbf{1}^T z = 3z^{an}$ and $\alpha_+^\mathsf{T} z = \alpha_-^\mathsf{T} z = 0$ and

$$\tilde{Z}^Y = \begin{bmatrix} z^{an}+3z^n & 0 & 0 \\ 0 & z^{an} & 0 \\ 0 & 0 & z^{an} \end{bmatrix} \tag{16.74a}$$

Hence the sequence impedance matrix $\tilde{Z}^Y$ is diagonal even though the phase impedance Z^Y is not. This implies that the external model $\tilde{V} = -\tilde{Z}^Y \tilde{I}$ relating the sequence voltage

and current in the sequence coordinate is decoupled:

$$\begin{bmatrix}\tilde{V}_0\\ \tilde{V}_+\\ \tilde{V}_-\end{bmatrix} = -\begin{bmatrix}z^{an}+3z^n & 0 & 0\\ 0 & z^{an} & 0\\ 0 & 0 & z^{an}\end{bmatrix}\begin{bmatrix}\tilde{I}_0\\ \tilde{I}_+\\ \tilde{I}_-\end{bmatrix} \tag{16.74b}$$

i.e., the external model consists of three separate impedances:

$$\begin{aligned}\text{zero-seq impedance:} \quad & \tilde{V}_0 = -(z^{an}+3z^n)\,\tilde{I}_0\\ \text{positive-seq impedance:} \quad & \tilde{V}_+ = -z^{an}\tilde{I}_+\\ \text{negative-seq impedance:} \quad & \tilde{V}_- = -z^{an}\tilde{I}_-\end{aligned}$$

The interpretation is as follows. When the similarity transformation defined by the unitary matrix F transforms a power network from the abc phase coordinate to 0+− sequence coordinate (see Chapter 14.2.2), a balanced impedance with $z^{an} = z^{bn} = z^{cn}$ becomes decoupled in the sequence coordinate. If all devices are decoupled in the sequence coordinate, the entire *sequence networks* are decoupled and the sequence impedances are impedances on these decoupled sequence networks. Each sequence network can be analyzed separately like a single-phase network. We will explain in Chapter 16.4.5 on how to compose the sequence networks from sequence models of individual devices.

Note that if the impedance is not balanced then the relation $\tilde{V} = \tilde{Z}^Y \tilde{I}$ is generally coupled and power flow analysis using the sequence variables may not offer any advantage over using the phase variables.

Δ configuration z^Δ.

Consider the three-wire three-phase impedance z^Δ in Δ configuration shown in Figure 14.8 of Chapter 14.3.4. Recall the external model (14.27b) relating the terminal voltage and current (V, I):

$$V = -Z^\Delta I + \gamma\mathbf{1}, \qquad \mathbf{1}^{\mathsf{T}} I = 0 \tag{16.75}$$

where the zero-sequence voltage $\gamma := \frac{1}{3}\mathbf{1}^{\mathsf{T}} V$ is also a variable to be determined in an analysis problem and

$$Z^\Delta := \frac{1}{9}\Gamma^{\mathsf{T}} \underbrace{z^\Delta\left(\mathbb{I} - \frac{1}{\zeta}\mathbf{1}\tilde{z}^{\Delta\mathsf{T}}\right)}_{\hat{z}^\Delta}\Gamma$$

Substitute $V = F\tilde{V}$ and $I = F\tilde{I}$ to obtain the external model in the sequence coordinate:

$$\tilde{V} = -\underbrace{\left(\overline{F}Z^\Delta F\right)}_{\tilde{Z}^\Delta}\tilde{I} + \gamma\overline{F}\mathbf{1}, \qquad \mathbf{1}^{\mathsf{T}} F\tilde{I} = 0 \tag{16.76}$$

where F and its inverse $\overline{F}$ is given in (16.73). It can be shown (Exercise 16.19) that

$$\tilde{Z}^{\Delta} := \frac{1}{9}(F\Lambda)^{\mathsf{H}}\hat{z}^{\Delta}(F\Lambda) \qquad \text{with} \qquad \Lambda := \begin{bmatrix} 0 & & \\ & 1-\alpha & \\ & & 1-\alpha^2 \end{bmatrix}$$

Moreover $\gamma\overline{F}\mathbf{1} = \tilde{V}_0 e_1$ and $\mathbf{1}^{\mathsf{T}}F\tilde{I} = \sqrt{3}\tilde{I}_0 = 0$.

If the impedance is balanced, i.e., $z^{ab} = z^{bc} = z^{ca}$ then (Exercise 16.19)

$$Z^{\Delta} = \frac{z^{ab}}{3}\left(\mathbb{I} - \frac{1}{3}\mathbf{1}\mathbf{1}^{\mathsf{T}}\right), \qquad \tilde{Z}^{\Delta} = \frac{z^{ab}}{3}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{16.77a}$$

and the external model (16.76) of a Δ-configured impedance in the sequence coordinate becomes decoupled:

$$\begin{bmatrix} 0 \\ \tilde{V}_+ \\ \tilde{V}_- \end{bmatrix} = -\frac{z^{ab}}{3}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} \tilde{I}_0 \\ \tilde{I}_+ \\ \tilde{I}_- \end{bmatrix}, \qquad \tilde{I}_0 = \frac{1}{\sqrt{3}}(I_a + I_b + I_c) = 0 \tag{16.77b}$$

For a Δ-configured load, $\tilde{I}_0 = 0$ because there is no neutral wire and therefore KCL dictates that the line currents sum to zero. The model (16.77) defines three separate impedances in the sequence coordinate:

$$\begin{aligned} &\text{zero-seq impedance:} && \text{null} \quad (\tilde{I}_0 = 0,\ \tilde{Z}_0 = \infty,\ \text{open circuit}) \\ &\text{positive-seq impedance:} && \tilde{V}_+ = -\frac{z^{ab}}{3}\tilde{I}_+ \\ &\text{negative-seq impedance:} && \tilde{V}_- = -\frac{z^{ab}}{3}\tilde{I}_- \end{aligned}$$

The interpretation is that a balanced Δ-configured impedance with $z^{ab} = z^{bc} = z^{ca}$ connected to a bus in a power network is transformed into an impedance of $z^{ab}/3$ at that bus (as we have seen in Chapter 1.2.4) in the positive and the negative-sequence networks and no impedance at that bus in the zero-sequence network (i.e., in the circuit model for the zero-sequence network, the connection between this bus and the ground is open; see (**??**) and discussions therein). This does not mean that the voltage $V_{j,0} = 0$ at bus j in the zero-sequence network where the impedance is connected. Rather, it means that there is zero injection at bus j ($\tilde{I}_{j,0} = 0$) and $\tilde{V}_{j,0}$ will be determined by the network equation; see Chapter 16.4.5.

Remark 16.10 (Terminal variables)**.** It is important to remember that the external models derived in this section relate the sequence variables $(\tilde{V}, \tilde{I})$ of the terminal voltage and current (V, I), not the internal voltage and current $(V^{Y/\Delta}, I^{Y/\Delta})$. See Example 16.13 on how to use sequence networks to calculate internal currents and powers. □

16.4.2 Sequence voltage sources

Y configuration $\left(E^Y, z^Y, z^n\right)$.

Consider the four-wire three-phase voltage source $\left(E^Y, z^Y, z^n\right)$ in Y configuration shown in Figure 14.7 of Chapter 14.3.3. Under assumption C14.1 (all neutrals are grounded and all voltages are defined with respect to the ground), recall the external model (14.13b) relating the terminal voltage and current (V, I):

$$V \;=\; E^Y \;-\; Z^Y I \qquad\qquad \text{with} \qquad\qquad Z^Y \;:=\; z^Y \;+\; z^n \mathbf{1}\mathbf{1}^{\mathsf{T}}$$

where Z^Y is the same matrix as that for Y-configured impedance. Substitute $V = F\tilde{V}$ and $I = F\tilde{I}$ to obtain the external model in the sequence coordinate:

$$\tilde{V} = \underbrace{\overline{F}E^Y}_{\tilde{E}^Y} \;-\; \underbrace{\overline{F}Z^Y F}_{\tilde{Z}^Y}\tilde{I} \;=:\; \tilde{E}^Y \;-\; \tilde{Z}^Y\tilde{I}$$

The sequence impedance matrix $\tilde{Z}^Y := \overline{F}Z^Y F$ is the same matrix as that for Y-configured impedance and the sequence internal voltage is:

$$\tilde{E}^Y \;:=\; \overline{F}E^Y \;=\; \frac{1}{\sqrt{3}}\begin{bmatrix}\mathbf{1}^{\mathsf{H}}E^Y\\ \alpha_+^{\mathsf{H}}E^Y\\ \alpha_-^{\mathsf{H}}E^Y\end{bmatrix}$$

When the impedance z^Y is balanced, i.e., $z^{an} = z^{bn} = z^{cn}$, even if the internal voltage E^Y is unbalanced, its external model in the sequence coordinate becomes decoupled (using (16.74b)):

$$\begin{bmatrix}\tilde{V}_0\\ \tilde{V}_+\\ \tilde{V}_-\end{bmatrix} \;=\; \begin{bmatrix}\tilde{E}_0^Y\\ \tilde{E}_+^Y\\ \tilde{E}_-^Y\end{bmatrix} \;-\; \begin{bmatrix}z^{an}+3z^n & 0 & 0\\ 0 & z^{an} & 0\\ 0 & 0 & z^{an}\end{bmatrix}\begin{bmatrix}\tilde{I}_0\\ \tilde{I}_+\\ \tilde{I}_-\end{bmatrix} \tag{16.78a}$$

This defines three separate non-ideal voltage sources:

$$\begin{aligned}\text{zero-seq voltage source:} \qquad & \tilde{V}_0 \;=\; \tilde{E}_0^Y \;-\; (z^{an}+3z^n)\,\tilde{I}_0\\ \text{positive-seq voltage source:} \qquad & \tilde{V}_+ \;=\; \tilde{E}_+^Y \;-\; z^{an}\tilde{I}_+\\ \text{negative-seq voltage source:} \qquad & \tilde{V}_- \;=\; \tilde{E}_-^Y \;-\; z^{an}\tilde{I}_-\end{aligned}$$

As for a balanced impedance, the voltage source becomes decoupled in the sequence coordinate even if they remain unbalanced.

Furthermore, if $E^Y = E^{an}\alpha_+$ is a balanced positive-sequence set then only the positive-sequence voltage is nonzero:

$$\overline{F}E^Y \;=\; \tilde{E}^Y \;=\; \frac{1}{\sqrt{3}}\begin{bmatrix}\mathbf{1}^{\mathsf{T}}\\ \alpha_-^{\mathsf{T}}\\ \alpha_+^{\mathsf{T}}\end{bmatrix}(E^{an}\alpha_+) \;=\; \frac{E^{an}}{\sqrt{3}}\begin{bmatrix}\mathbf{1}^{\mathsf{H}}\alpha_+\\ \alpha_+^{\mathsf{H}}\alpha_+\\ \alpha_-^{\mathsf{H}}\alpha_+\end{bmatrix} \;=\; \begin{bmatrix}0\\ \sqrt{3}E^{an}\\ 0\end{bmatrix}$$

The external model of a balanced Y-configured voltage source in the sequence coordinate becomes (from (16.78a)):

$$\begin{bmatrix} \tilde{V}_0 \\ \tilde{V}_+ \\ \tilde{V}_- \end{bmatrix} = \begin{bmatrix} 0 \\ \sqrt{3}E^{an} \\ 0 \end{bmatrix} - \begin{bmatrix} z^{an}+3z^n & 0 & 0 \\ 0 & z^{an} & 0 \\ 0 & 0 & z^{an} \end{bmatrix} \begin{bmatrix} \tilde{I}_0 \\ \tilde{I}_+ \\ \tilde{I}_- \end{bmatrix} \tag{16.78b}$$

This defines a voltage source $\left(\sqrt{3}E^{an}, z^{an}\right)$ on the positive-sequence network and impedances on the other sequence networks:

$$\begin{aligned} \text{zero-seq impedance:} \qquad & \tilde{V}_0 = -\left(z^{an}+3z^n\right)\tilde{I}_0 \\ \text{positive-seq voltage source:} \qquad & \tilde{V}_+ = \sqrt{3}E^{an} - z^{an}\tilde{I}_+ \\ \text{negative-seq impedance:} \qquad & \tilde{V}_- = -z^{an}\tilde{I}_- \end{aligned}$$

They are illustrated in Figure 16.8. [6]

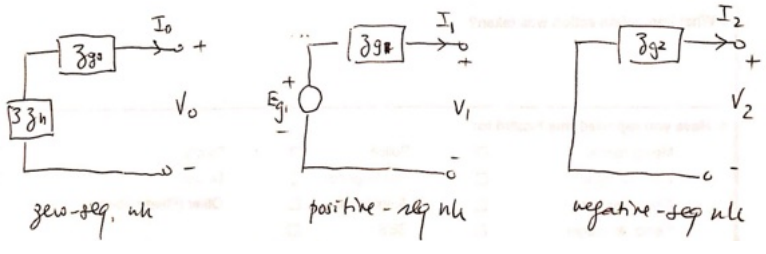


Figure 16.8 The sequence networks of a balanced voltage source $\left(E^Y, z^Y, z^n\right)$ in Y configuration.

Δ configuration $\left(E^\Delta, z^\Delta\right)$.

Consider the three-phase voltage source $\left(E^\Delta, z^\Delta\right)$ in Δ configuration shown in Figure 14.8 of Chapter 14.3.4. One of its external models is (14.21b), reproduced here [7]

$$V = \hat{\Gamma}E^\Delta - Z^\Delta I + \gamma\mathbf{1}, \qquad \mathbf{1}^\mathsf{T} I = 0$$

where

$$\hat{\Gamma} := \frac{1}{3}\Gamma^\mathsf{T}\left(\mathbb{I} - \frac{1}{\zeta}\tilde{z}^\Delta\mathbf{1}^\mathsf{T}\right), \qquad Z^\Delta := \frac{1}{9}\Gamma^\mathsf{T} z^\Delta\left(\mathbb{I} - \frac{1}{\zeta}\mathbf{1}\tilde{z}^{\Delta\mathsf{T}}\right)\Gamma$$

where $\tilde{z}^\Delta := z^\Delta\mathbf{1}$ is a column vector and $\zeta := \mathbf{1}^\mathsf{T}\tilde{z}^\Delta$ is a scalar. This is similar to the model (16.75) of Δ-configured impedance with the extra term $\hat{\Gamma}E^\Delta$. Substitute $V = F\tilde{V}$

[6] The sequence networks of synchronous generators are generally more complicated and their sequence impedances (mostly reactances) are generally unequal unlike the model in (16.78b); see e.g. [160, Section 2.3].

[7]

and $I = F\tilde{I}$ to obtain the external model in the sequence coordinate:

$$\tilde{V} = \underbrace{\overline{F}\hat{\Gamma}E^{\Delta}}_{\tilde{E}^{\Delta}} - \underbrace{\overline{F}Z^{\Delta}F}_{\tilde{Z}^{\Delta}}\tilde{I} + \gamma\overline{F}\mathbf{1} =: \tilde{E}^{\Delta} - \tilde{Z}^{\Delta}\tilde{I} + \tilde{V}_0 e_1, \qquad \mathbf{1}^{\mathsf{T}}F\tilde{I} = 0$$

where $\mathbf{1}^{\mathsf{T}}F\tilde{I} = \sqrt{3}\tilde{I}_0 = 0$. This is similar to (16.76) with the extra term (Exercise 16.20)

$$\tilde{E}^{\Delta} := \overline{F}\hat{\Gamma}E^{\Delta} = \Lambda^{\dagger}\overline{F}\left(\mathbb{I} - \frac{1}{\zeta}\tilde{z}^{\Delta}\mathbf{1}^{\mathsf{T}}\right)E^{\Delta} \quad \text{with} \quad \Lambda^{\dagger} := \begin{bmatrix} 0 & & \\ & (1-\alpha)^{-1} & \\ & & (1-\alpha^2)^{-1} \end{bmatrix}$$

If the impedance is balanced, i.e., $z^{ab} = z^{bc} = z^{ca}$ then $\tilde{z}^{\Delta} := z^{ab}\mathbf{1}$, $\zeta := 3z^{ab}$, and (Exercise 16.20 and from (16.77a))

$$\tilde{E}^{\Delta} = \begin{bmatrix} 0 \\ (1-\alpha)^{-1}\tilde{E}^{\Delta}_{+} \\ (1-\alpha^2)^{-1}\tilde{E}^{\Delta}_{-} \end{bmatrix}, \quad Z^{\Delta} = \frac{z^{ab}}{3}\left(\mathbb{I} - \frac{1}{3}\mathbf{1}\mathbf{1}^{\mathsf{T}}\right), \quad \tilde{Z}^{\Delta} = \frac{z^{ab}}{3}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

where the sequence voltages are $\tilde{E}^{\Delta}_{+} := \frac{1}{3}\alpha_{+}^{\mathsf{H}}E^{\Delta}$ and $\tilde{E}^{\Delta}_{-} := \frac{1}{3}\alpha_{-}^{\mathsf{H}}E^{\Delta}$. The zero-sequence voltage $\tilde{E}^{\Delta}_{0} = 0$ because there is no neutral line in Δ configuration. Hence the external model in the sequence coordinate is

$$\begin{bmatrix} 0 \\ \tilde{V}_{+} \\ \tilde{V}_{-} \end{bmatrix} = \begin{bmatrix} 0 \\ (1-\alpha)^{-1}\tilde{E}^{\Delta}_{+} \\ (1-\alpha^2)^{-1}\tilde{E}^{\Delta}_{-} \end{bmatrix} - \frac{z^{ab}}{3}\begin{bmatrix} 0 \\ \tilde{I}_{+} \\ \tilde{I}_{-} \end{bmatrix}, \quad \tilde{I}_0 = \frac{1}{\sqrt{3}}\left(I^a + I^b + I^c\right) = 0 \qquad (16.79a)$$

Hence the voltage sources in the sequence coordinate are unbalanced but decoupled:

$$\begin{aligned} &\text{zero-seq voltage source:} && \text{null} \quad (\tilde{I}_0 = 0,\ \tilde{Z}_0 = \infty, \text{ open circuit}) \\ &\text{positive-seq voltage source:} && \tilde{V}_{+} = \frac{E^{\Delta}_{+}}{1-\alpha} - \frac{z^{ab}}{3}\tilde{I}_{+} \\ &\text{negative-seq voltage source:} && \tilde{V}_{-} = \frac{E^{\Delta}_{-}}{1-\alpha^2} - \frac{z^{ab}}{3}\tilde{I}_{-} \end{aligned}$$

As for a Δ-configured impedance, a symmetric voltage source in a power network is transformed into voltage sources in the positive and negative-sequence networks. The equivalent series impedance of the sequence voltage sources is $z^{ab}/3$ as we have seen in Chapter 1.2.4. There is no device (open circuit) in the zero-sequence network, which means that, when the voltage source is connected to bus j, there is zero injection at bus j in the zero-sequence network ($\tilde{I}_{j,0} = 0$) and $\tilde{V}_{j,0}$ will be determined by the network equation; see Chapter 16.4.5.

Furthermore, if $E^{\Delta} := E^{ab}\alpha_{+}$ is a balanced positive-sequence set then

$$\tilde{E}^{\Delta}_{+} = \sqrt{3}E^{ab}, \qquad \tilde{E}^{\Delta}_{-} = 0$$

and

$$\begin{bmatrix} 0 \\ \tilde{V}_{+} \\ \tilde{V}_{-} \end{bmatrix} = \begin{bmatrix} 0 \\ e^{-\mathbf{i}\pi/6}E^{ab} \\ 0 \end{bmatrix} - \frac{z^{ab}}{3}\begin{bmatrix} 0 \\ \tilde{I}_{+} \\ \tilde{I}_{-} \end{bmatrix} \qquad (16.79b)$$

since $\sqrt{3}/(1-\alpha) = e^{-\mathbf{i}\pi/6}$. This defines a voltage source $\left(e^{-\mathbf{i}\pi/6}E^{ab}, z^{ab}/3\right)$ in the positive-sequence network and an impedance $z^{ab}/3$ in the negative-sequence network:

$$\begin{aligned}
\text{zero-seq voltage source:} \quad & \text{null} \quad (\tilde{I}_0 = 0,\ \tilde{Z}_0 = \infty,\ \text{open circuit}) \\
\text{positive-seq voltage source:} \quad & \tilde{V}_+ = e^{-\mathbf{i}\pi/6}E^{ab} - \frac{z^{ab}}{3}\tilde{I}_+ \\
\text{negative-seq voltage source:} \quad & \tilde{V}_- = -\frac{z^{ab}}{3}\tilde{I}_-
\end{aligned}$$

There is no device (open circuit) in the zero-sequence network.

16.4.3 Sequence current sources

Y configuration $\left(J^Y, y^Y, z^n\right)$.

An external model of a *Y*-configured current source $\left(J^Y, y^Y, z^n\right)$ is (from (14.15a)):

$$I = -J^Y - y^Y\left(V - V^n\mathbf{1}\right)$$

Substitute $V = F\tilde{V}$ and $I = F\tilde{I}$ to obtain the external model in the sequence coordinate:

$$\tilde{I} = -\underbrace{\overline{F}J^Y}_{\tilde{J}^Y} - \underbrace{\overline{F}y^YF}_{\tilde{Y}^Y}\tilde{V} + V^n\overline{F}y^Y\mathbf{1}$$

where $\tilde{J}^Y := \overline{F}J^Y$ and

$$\tilde{Y}^Y := \overline{F}y^YF = \frac{1}{3}\left(y^{an}\mathbf{1}\mathbf{1}^{\mathsf{H}} + y^{bn}\alpha_-\alpha_-^{\mathsf{H}} + y^{cn}\alpha_+\alpha_+^{\mathsf{H}}\right) \tag{16.80}$$

If the phase admittance $y^Y := y^{an}\mathbb{I}$ is balanced then the sequence admittance is also balanced:

$$\tilde{Y}^Y := \overline{F}y^YF = y^{an}\mathbb{I}, \qquad \overline{F}y^Y\mathbf{1} = y^{an}\overline{F}\mathbf{1} = y^{an}\begin{bmatrix}\sqrt{3}\\0\\0\end{bmatrix}$$

The current source becomes decoupled in the sequence coordinate even though it is unbalanced:

$$\begin{bmatrix}\tilde{I}_0\\ \tilde{I}_+\\ \tilde{I}_-\end{bmatrix} = -\begin{bmatrix}\tilde{J}^Y_0\\ \tilde{J}^Y_+\\ \tilde{J}^Y_-\end{bmatrix} - y^{an}\left(\begin{bmatrix}\tilde{V}_0\\ \tilde{V}_+\\ \tilde{V}_-\end{bmatrix} - \begin{bmatrix}\sqrt{3}V^n\\0\\0\end{bmatrix}\right)$$

In particular the neutral voltage V^n appears only in the zero-sequence network. If, furthermore, the current source $J^Y := J^{an}\alpha_+$ is in a balanced positive sequence then

$$\tilde{J}^Y = \overline{F}J^Y = \frac{J^{an}}{\sqrt{3}}\begin{bmatrix}\mathbf{1}^{\mathsf{H}}\\ \alpha_+^{\mathsf{H}}\\ \alpha_-^{\mathsf{H}}\end{bmatrix}\alpha_+ = \begin{bmatrix}0\\ \sqrt{3}J^{an}\\0\end{bmatrix}$$

The current source in the sequence coordinate becomes a current source $\left(\sqrt{3}J^{an}, y^{an}\right)$ in the positive-sequence network and the impedance $(y^{an})^{-1}$ in each of the other two sequence networks:

$$\begin{aligned} \text{zero-seq impedance:} \qquad & \tilde{I}_0 = -y^{an}\left(\tilde{V}_0 - \sqrt{3}V^n\right) \\ \text{positive-seq current source:} \qquad & \tilde{I}_+ = -\sqrt{3}J^{an} - y^{an}\tilde{V}_+ \\ \text{negative-seq impedance:} \qquad & \tilde{I}_- = -y^{an}\tilde{V}_- \end{aligned}$$

The interpretation of the zero-sequence impedance is that the voltage drop across the impedance $(y^{an})^{-1}$ is $\tilde{V}_0 - \sqrt{3}V^n$ with one end of the impedance at a potential $\sqrt{3}V^n$ with respect to the common voltage reference point.

When assumption C14.1 holds (the neutral is grounded and voltages are defined with respect to the ground) so that $V^n = -z^n\left(\mathbf{1}^{\mathsf{T}} I\right)$, we have

$$V^n = -z^n\left(\mathbf{1}^{\mathsf{T}} F\tilde{I}\right) = -\frac{z^n}{\sqrt{3}}\left(\mathbf{1}^{\mathsf{T}}\begin{bmatrix}\mathbf{1} & \alpha_+ & \alpha_-\end{bmatrix}\tilde{I}\right) = -\sqrt{3}z^n\tilde{I}_0$$

i.e., the neutral voltage depends only on the zero-sequence current $\tilde{I}_0$ (of the terminal current I). Substitute this into expressions above, the sequence voltage and current $\left(\tilde{V}, \tilde{I}\right)$ satisfies, when $y^Y := y^{an}\mathbb{I}$,

$$\begin{bmatrix}(1+3\,y^{an}z^n)\,\tilde{I}_0 \\ \tilde{I}_+ \\ \tilde{I}_-\end{bmatrix} = -\begin{bmatrix}\tilde{J}_0^Y \\ \tilde{J}_+^Y \\ \tilde{J}_-^Y\end{bmatrix} - y^{an}\begin{bmatrix}\tilde{V}_0 \\ \tilde{V}_+ \\ \tilde{V}_-\end{bmatrix} \tag{16.81a}$$

and the current source becomes decoupled in the sequence coordinate even if they remain unbalanced:

$$\begin{aligned} \text{zero-seq current source:} \qquad & \tilde{I}_0 = -\frac{\tilde{J}_0^Y}{1+3\,y^{an}z^n} - \frac{y^{an}}{1+3\,y^{an}z^n}\tilde{V}_0 \\ \text{positive-seq current source:} \qquad & \tilde{I}_+ = -\tilde{J}_+^Y - y^{an}\tilde{V}_+ \\ \text{negative-seq current source:} \qquad & \tilde{I}_- = -\tilde{J}_-^Y - y^{an}\tilde{V}_- \end{aligned}$$

If, furthermore, the current source $J^Y := J^{an}\alpha_+$ they become:

$$\text{zero-seq admittance:} \qquad \tilde{I}_0 = -\frac{y^{an}}{1+3\,y^{an}z^n}\tilde{V}_0 \tag{16.81b}$$

$$\text{positive-seq current source:} \qquad \tilde{I}_+ = -\sqrt{3}J^{an} - y^{an}\tilde{V}_+ \tag{16.81c}$$

$$\text{negative-seq admittance:} \qquad \tilde{I}_- = -y^{an}\tilde{V}_- \tag{16.81d}$$

Instead of sequence current sources in (16.81), equivalent voltage sources in the sequence domain can also be derived starting from the external model of a current source (from (14.15b)): $V = -\left(z^Y J^Y + Z^Y I\right)$ where $z^Y := \left(y^Y\right)^{-1}$ and $Z^Y := z^Y + z^n\,\mathbf{1}\mathbf{1}^{\mathsf{T}}$; see Exercise 16.22.

Δ configuration $\left(J^\Delta, y^\Delta\right)$.

The external model of a Δ-configured current source is (from (14.23a)):

$$I \;=\; -\left(\Gamma^\mathsf{T} J^\Delta \;+\; Y^\Delta V\right)$$

where $Y^\Delta := \Gamma^\mathsf{T} y^\Delta \Gamma$ is the matrix in (14.21a). Substitute $V = F\tilde{V}$ and $I = F\tilde{I}$ to obtain the external model in the sequence coordinate:

$$\tilde{I} \;=\; -\left(\underbrace{\overline{F}\Gamma^\mathsf{T} J^\Delta}_{\tilde{J}^\Delta} + \underbrace{\overline{F} Y^\Delta F}_{\tilde{Y}^\Delta}\tilde{V}\right) \;=:\; -\left(\tilde{J}^\Delta + \tilde{Y}^\Delta \tilde{V}\right)$$

where

$$\begin{aligned}
\tilde{J}^\Delta \;&:=\; \overline{F}\Gamma^\mathsf{T} J^\Delta \;=\; 3\Lambda^\dagger \overline{F} J^\Delta \\
\tilde{Y}^\Delta \;&:=\; \overline{F}\left(\Gamma^\mathsf{T} y^\Delta \Gamma\right) F \;=\; \overline{F}\left(3F\Lambda^\dagger \overline{F}\right) y^\Delta \left(F\Lambda\overline{F}\right) F \;=\; 3\Lambda^\dagger \left(\overline{F} y^\Delta F\right)\Lambda
\end{aligned}$$

where we have used $\Gamma = F\Lambda\overline{F}$ and $\Gamma^\mathsf{T} = 3\Gamma^\dagger = 3F\Lambda^\dagger\overline{F}$ from (14.6).

If the phase admittance $y^Y := y^{ab}\mathbb{I}$ is balanced, then the effective phase admittance Y^Δ is not diagonal but its sequence admittance $\tilde{Y}^\Delta$ is unbalanced but diagonal:

$$\begin{aligned}
Y^\Delta \;&:=\; y^{ab}\Gamma^\mathsf{T}\Gamma \;=\; 3y^{ab}\left(\mathbb{I} - \frac{1}{3}\mathbf{1}\mathbf{1}^\mathsf{T}\right) \\
\tilde{Y}^\Delta \;&:=\; \overline{F} Y^\Delta F \;=\; 3y^{ab}\left(\mathbb{I} \;-\; e_1 e_1^\mathsf{T}\right)
\end{aligned}$$

where we have used $\Gamma^\mathsf{T}\Gamma = 3\left(\mathbb{I} - \frac{1}{3}\mathbf{1}\mathbf{1}^\mathsf{T}\right)$ from Theorem 14.2 and $\overline{F}\mathbf{1} = \sqrt{3}e_1$. Hence the current source is unbalanced but decoupled in the sequence coordinate:

$$\begin{bmatrix}\tilde{I}_0\\ \tilde{I}_+\\ \tilde{I}_-\end{bmatrix} \;=\; -\begin{bmatrix}\tilde{J}_0^\Delta\\ \tilde{J}_+^\Delta\\ \tilde{J}_-^\Delta\end{bmatrix} \;-\; 3y^{ab}\begin{bmatrix}0&0&0\\0&1&0\\0&0&1\end{bmatrix}\begin{bmatrix}\tilde{V}_0\\ \tilde{V}_+\\ \tilde{V}_-\end{bmatrix} \;=\; -\begin{bmatrix}\tilde{J}_0^\Delta\\ \tilde{J}_+^\Delta\\ \tilde{J}_-^\Delta\end{bmatrix} \;-\; 3y^{ab}\begin{bmatrix}0\\ \tilde{V}_+\\ \tilde{V}_-\end{bmatrix} \tag{16.82a}$$

The zero-sequence network has an ideal current source $\tilde{J}_0^\Delta$ and the other two sequence networks each has a non-ideal current source:

$$\begin{aligned}
&\text{zero-seq current source:} && \tilde{I}_0 \;=\; -\tilde{J}_0^\Delta \\
&\text{positive-seq current source:} && \tilde{I}_+ \;=\; -\tilde{J}_+^\Delta \;-\; 3y^{ab}\tilde{V}_+ \\
&\text{negative-seq current source:} && \tilde{I}_- \;=\; -\tilde{J}_-^\Delta \;-\; 3y^{ab}\tilde{V}_-
\end{aligned}$$

If, furthermore, the current source $J^\Delta := J^{ab}\alpha_+$ is a balanced positive sequence then

$$\tilde{J}^\Delta \;:=\; 3J^{ab}\Lambda^\dagger\overline{F}\alpha_+ \;=\; 3J^{ab}\begin{bmatrix}0&&\\&(1-\alpha)^{-1}&\\&&\left(1-\alpha^2\right)^{-1}\end{bmatrix}\begin{bmatrix}0\\ \sqrt{3}\\ 0\end{bmatrix} \;=\; \begin{bmatrix}0\\ 3e^{-\mathbf{i}\pi/6}J^{ab}\\ 0\end{bmatrix}$$

where we have used $\overline{F}\alpha_+ = \sqrt{3}e_2$ and $\sqrt{3}/(1-\alpha) = e^{-\mathbf{i}\pi/6}$. A balanced

positive-sequence current source is therefore transformed into a current source $\left(3e^{-\mathbf{i}\pi/6}J^{ab}, 3y^{ab}\right)$ in the positive-sequence network and an admittance $3y^{ab}$ in the negative-sequence network:

$$\text{zero-seq current source:} \quad \text{null} \quad (\tilde{I}_0 = 0) \tag{16.82b}$$

$$\text{positive-seq current source:} \quad \tilde{I}_+ = -3e^{-\mathbf{i}\pi/6}J^{ab} - 3y^{ab}\tilde{V}_+ \tag{16.82c}$$

$$\text{negative-seq admittance:} \quad \tilde{I}_- = -3y^{ab}\tilde{V}_- \tag{16.82d}$$

There is no device in the zero-sequence network because Δ configuration has no neutral line.

16.4.4 Sequence line model

Consider a three-phase line connecting bus j and bus k that is modeled by only a series phase impedance matrix z^s_{jk}. We omit shunt admittances for simplicity.[8] The terminal voltages and the line current is related by Ohm's law:

$$V_j - V_k = z^s_{jk} I_{jk}$$

Convert to the sequence coordinate by substituting $V_j = F\tilde{V}_j$, $V_k = F\tilde{V}_k$ and $I_{jk} = F\tilde{I}_{jk}$ to get

$$\tilde{V}_j - \tilde{V}_k = \underbrace{\left(\overline{F} z^s_{jk} F\right)}_{\tilde{z}^s_{jk}} \tilde{I}_{jk} =: \tilde{z}^s_{jk}\tilde{I}_{jk} \tag{16.83a}$$

where $\tilde{z}^s_{jk} := \overline{F} z^s_{jk} F$ is called the *sequence impedance matrix* of line (j,k). This does not assume C16.1, i.e., z^s_{jk} and z^s_{kj} may be different.

If the phase impedance matrix z^s_{jk} is symmetric of the form in (15.9) then (omitting the subscript jk for simplicity)

$$\tilde{z}^s_{jk} = \frac{1}{3}\begin{bmatrix}1 & 1 & 1\\ 1 & \alpha^2 & \alpha\\ 1 & \alpha & \alpha^2\end{bmatrix}\begin{bmatrix}z^1 & z^2 & z^2\\ z^2 & z^1 & z^2\\ z^2 & z^2 & z^1\end{bmatrix}\begin{bmatrix}1 & 1 & 1\\ 1 & \alpha & \alpha^2\\ 1 & \alpha^2 & \alpha\end{bmatrix} = \begin{bmatrix}z^1+2z^2 & 0 & 0\\ 0 & z^1-z^2 & 0\\ 0 & 0 & z^1-z^2\end{bmatrix} \tag{16.83b}$$

[8] Shunt admittances can be included using (15.8a): $I_{jk} = y^s_{jk}(V_j - V_k) + y^m_{jk}V_j$ in which case the sequence admittance matrices $\left(\tilde{y}^s_{jk}, \tilde{y}^m_{jk}, \tilde{y}^m_{kj}\right)$ are given by:

$$\tilde{I}_{jk} = \underbrace{\left(\overline{F} y^s_{jk} F\right)}_{\tilde{y}^s_{jk}}(\tilde{V}_j - \tilde{V}_k) + \underbrace{\left(\overline{F} y^m_{jk} F\right)}_{\tilde{y}^m_{jk}}\tilde{V}_j$$

i.e., the sequence impedance matrix of line (j,k) is diagonal. This defines three separate sequence networks:

$$\begin{aligned}
\text{zero-seq impedance:} \qquad & \tilde{V}_{j,0} - \tilde{V}_{k,0} &&= \left(z^1 + 2z^2\right) \tilde{I}_{jk,0} \\
\text{positive-seq impedance:} \qquad & \tilde{V}_{j,+} - \tilde{V}_{k,+} &&= \left(z^1 - z^2\right) \tilde{I}_{jk,+} \\
\text{negative-seq impedance:} \qquad & \tilde{V}_{j,-} - \tilde{V}_{k,-} &&= \left(z^1 - z^2\right) \tilde{I}_{jk,-}
\end{aligned}$$

The phase impedance matrix z^s_{jk} in (15.9) is complex symmetric but not Hermitian. In general a complex symmetric matrix may not be diagonalizable (see Exercise 16.23 for an example). The matrix z^s_{jk} however is normal and hence unitarily diagonalizable through the unitary matrix F (Exercise 16.24).

16.4.5 Three-phase analysis

We now explain how to compose sequence networks from individual device models in the sequence coordinate derived in Chapters 16.4.1–16.4.4. We will show that if a network is unbalanced but symmetric, its sequence networks are decoupled and can be analyzed separately.

Definition 16.2 (Symmetric network)**.** A network $G := \left(\overline{N}, E\right)$ that connects a set of three-phase devices by three-phase lines is called *symmetric* if the following assumptions hold:

C16.9: All impedances are symmetric $z^{Y/\Delta}_j = z^{an/ab}_j \mathbb{I}$.
C16.10: All voltage sources have symmetric series impedances $z^{Y/\Delta}_j = z^{an/ab}_j \mathbb{I}$.
C16.11: All current sources have symmetric shunt admittances $y^{Y/\Delta}_j = y^{an/ab}_j \mathbb{I}$.
C16.12: All three-phase lines (j,k) have series impedances $z^s_{jk} = z^s_{kj}$ that satisfy (15.9) and zero shunt admittances. In particular we assume for simplicity that assumption C16.1 holds.

Suppose we are given a symmetric network with a single three-phase device at each bus. As before, partition the set $\overline{N}$ of buses into 6 disjoint subsets:

- $N^{Y/\Delta}_v$: buses with non-ideal voltage sources in Y or Δ configurations: $\left(E^Y, z^Y, z^n\right)$, $\left(E^\Delta, z^\Delta\right)$.
- $N^{Y/\Delta}_c$: buses with non-ideal current sources in Y or Δ configurations: $\left(J^Y, y^Y, z^n\right)$, $\left(J^\Delta, y^\Delta\right)$.
- $N^{Y/\Delta}_i$: buses with impedances in Y or Δ configurations: $\left(z^Y, z^n\right)$, z^Δ.

Suppose assumption C14.1 holds (i.e., all neutrals are grounded and voltages are defined with respect to the ground). C14.1 and the assumption of a single three-phase

device at each bus are made without loss of generality only to simplify presentation (see Example 16.13 for a network where there are two devices connected to a single bus). We will follow the solution strategy of Chapter 16.3.4 that solves

$$\begin{bmatrix} I_v \\ I_c \\ I_i \end{bmatrix} = \underbrace{\begin{bmatrix} Y_{vv} & Y_{vc} & Y_{vi} \\ Y_{cv} & Y_{cc} & Y_{ci} \\ Y_{iv} & Y_{ic} & Y_{ii} \end{bmatrix}}_{Y} \begin{bmatrix} V_v \\ V_c \\ V_i \end{bmatrix} \tag{16.84}$$

for the terminal voltage $V_{-v} := (V_c, V_i)$ and current $I_{-c} := (I_v, I_i)$. All other variables such as internal voltages and currents $\left(V^{Y/\Delta}, I^{Y/\Delta}\right)$ can then be derived in terms of the terminal voltages and currents (V, I).

We now show that (16.84) decomposes into three separate sequence networks so that it can be solved by analyzing three simpler networks. Furthermore, if not only is the network symmetric but all voltage and current sources are also balanced positive-sequence sets, then it is sufficient to analyze only the positive-sequence network. This is because in that case there are only impedances and admittances, but no voltage or current sources, in the zero-sequence and the negative-sequence networks.

Let $\mathbb{I}_{N+1}$ be the identity matrix of size $N+1$ so that $\mathbb{I}_{N+1} \otimes F$ is a matrix of size $3(N+1) \times 3(N+1)$. Convert both sides of (16.84) into the sequence coordinate by substituting

$$I \;=:\; (\mathbb{I}_{N+1} \otimes F)\,\tilde{I}, \qquad\qquad V \;=:\; (\mathbb{I}_{N+1} \otimes F)\,\tilde{V}$$

to obtain

$$\begin{bmatrix} \tilde{I}_v \\ \tilde{I}_c \\ \tilde{I}_i \end{bmatrix} = \underbrace{\begin{bmatrix} \tilde{Y}_{vv} & \tilde{Y}_{vc} & \tilde{Y}_{vi} \\ \tilde{Y}_{cv} & \tilde{Y}_{cc} & \tilde{Y}_{ci} \\ \tilde{Y}_{iv} & \tilde{Y}_{ic} & \tilde{Y}_{ii} \end{bmatrix}}_{\tilde{Y}} \begin{bmatrix} \tilde{V}_v \\ \tilde{V}_c \\ \tilde{V}_i \end{bmatrix} \quad \text{where} \quad \tilde{Y} := \left(\mathbb{I}_{N+1} \otimes \overline{F}\right) Y \,(\mathbb{I}_{N+1} \otimes F) \tag{16.85a}$$

and we have used $(\mathbb{I}_{N+1} \otimes F)^{-1} = \mathbb{I}_{N+1} \otimes \overline{F}$ from Lemma 16.6. The three rows $(3j+1, 3j+2, 3j+3)$ of (16.85a) corresponding to the sequence current $\tilde{I}_j \in \mathbb{C}^3$ of device $j = 0, \ldots, N$, are:

$$\tilde{I}_j = \sum_{\substack{j:j\sim k \\ k\in N_v}} \tilde{y}_{jk}\left(\tilde{V}_j - \tilde{V}_k\right) + \sum_{\substack{j:j\sim k \\ k\in N_c}} \tilde{y}_{jk}\left(\tilde{V}_j - \tilde{V}_k\right) + \sum_{\substack{j:j\sim k \\ k\in N_i}} \tilde{y}_{jk}\left(\tilde{V}_j - \tilde{V}_k\right), \quad j \in \overline{N} \tag{16.85b}$$

where $\tilde{y}_{jk} := \left(\tilde{z}_{jk}\right)^{-1} := \left(\overline{F} z^s_{jk} F\right)^{-1}$ are the series admittance matrices of lines (j,k) in the sequence coordinate from (16.83). The network equation (16.85) relates terminal variables. To show that the three-phase network decomposes into decoupled sequence networks we have to show both of the following:

1 The three rows of (16.85b) are decoupled, i.e., the zero-sequence current $\tilde{I}_{j,0}$ depends only on voltages $\tilde{V}_{k,0}$ of its adjacent buses $k \neq j$ in the zero-sequence network but not on voltages $\tilde{V}_{k,s}$ in the other sequence networks $s \in \{+,-\}$. Similarly for the positive and negative-sequence currents $\left(\tilde{I}_{j,+}, \tilde{I}_{j,-}\right)$.
2 At each bus j, the terminal voltage and current $\left(\tilde{V}_j, \tilde{I}_j\right)$ are decoupled, i.e., the zero-sequence voltage $\tilde{V}_{j,0}$ does not depend on the positive or negative-sequence currents $\left(\tilde{I}_{j,+}, \tilde{I}_{j,-}\right)$ at bus j. Similarly for $\tilde{V}_{j,+}$ and $\tilde{V}_{j,-}$.

The first claim follows from C16.12 in Definition 16.2 which implies that $\tilde{y}_{jk}$ is diagonal (from (16.83)). This means that the three rows of (16.85b) are decoupled at all buses $j \in \overline{N}$. We hence only need to prove the second claim that locally at each bus j the sequence voltage $\tilde{V}_{j,s}$, $s \in \{0,+,-\}$, does not couple the sequence currents $\tilde{I}_{j,s'}$, $s' \neq s$. This can be shown using the models derived in Chapters 16.4.1–16.4.3.

Specifically the external models of the three-phase devices are as follows.

1 Voltage source $j \in N_v$ from (16.78a) and (16.79a):

$$\begin{bmatrix} \tilde{V}_{j,0} \\ \tilde{V}_{j,+} \\ \tilde{V}_{j,-} \end{bmatrix} = \begin{bmatrix} \tilde{E}^Y_{j,0} \\ \tilde{E}^Y_{j,+} \\ \tilde{E}^Y_{j,-} \end{bmatrix} - \begin{bmatrix} z^{an}_j + 3z^n_j & & \\ & z^{an}_j & \\ & & z^{an}_j \end{bmatrix} \begin{bmatrix} \tilde{I}_{j,0} \\ \tilde{I}_{j,+} \\ \tilde{I}_{j,-} \end{bmatrix}, \quad j \in N^Y_v \tag{16.86a}$$

$$\begin{bmatrix} 0 \\ \tilde{V}_{j,+} \\ \tilde{V}_{j,-} \end{bmatrix} = \begin{bmatrix} 0 & & \\ & \frac{1}{1-\alpha} & \\ & & \frac{1}{1-\alpha^2} \end{bmatrix} \begin{bmatrix} \tilde{E}^\Delta_{j,0} \\ \tilde{E}^\Delta_{j,+} \\ \tilde{E}^\Delta_{j,-} \end{bmatrix} - \frac{z^{ab}_j}{3} \begin{bmatrix} \tilde{I}_{j,0} \\ \tilde{I}_{j,+} \\ \tilde{I}_{j,-} \end{bmatrix}, \quad j \in N^\Delta_v \tag{16.86b}$$

2 Current sources $j \in N_c$ from (16.81a) and (16.82a):

$$\begin{bmatrix} \tilde{I}_{j,0} \\ \tilde{I}_{0,+} \\ \tilde{I}_{j,-} \end{bmatrix} = -\frac{1}{1+3\,y^{an}\,z^n} \begin{bmatrix} \tilde{J}^Y_{j,0} \\ \tilde{J}^Y_{0,+} \\ \tilde{J}^Y_{0,-} \end{bmatrix} - \frac{y^{an}}{1+3\,y^{an}\,z^n} \begin{bmatrix} \tilde{V}_{j,0} \\ \tilde{V}_{j,+} \\ \tilde{V}_{0,-} \end{bmatrix}, \quad j \in N^Y_c \tag{16.86c}$$

$$\begin{bmatrix} \tilde{I}_{j,0} \\ \tilde{I}_{0,+} \\ \tilde{I}_{j,-} \end{bmatrix} = -\begin{bmatrix} \tilde{J}^\Delta_{j,0} \\ \tilde{J}^\Delta_{j,+} \\ \tilde{J}^\Delta_{j,-} \end{bmatrix} - 3y^{ab} \begin{bmatrix} 0 \\ \tilde{V}_{j,+} \\ \tilde{V}_{j,-} \end{bmatrix}, \quad j \in N^\Delta_c \tag{16.86d}$$

3 Impedances $j \in N_i$ from (16.74b) and (16.77b):

$$\begin{bmatrix} \tilde{V}_{j,0} \\ \tilde{V}_{j,+} \\ \tilde{V}_{j,-} \end{bmatrix} = -\begin{bmatrix} z^{an} + 3z^n & 0 & 0 \\ 0 & z^{an} & 0 \\ 0 & 0 & z^{an} \end{bmatrix} \begin{bmatrix} \tilde{I}_{j,0} \\ \tilde{I}_{j,+} \\ \tilde{I}_{j,-} \end{bmatrix}, \quad j \in N^Y_i \tag{16.86e}$$

$$\begin{bmatrix} 0 \\ \tilde{V}_{j,+} \\ \tilde{V}_{j,-} \end{bmatrix} = -\frac{z^{ab}}{3} \begin{bmatrix} \tilde{I}_{j,0} \\ \tilde{I}_{j,+} \\ \tilde{I}_{j,-} \end{bmatrix}, \quad j \in N^\Delta_i \tag{16.86f}$$

Therefore the terminal voltage and current $\left(\tilde{V}_j, \tilde{I}_j\right)$ at each bus j are decoupled, even if they are unbalanced. The network equation (16.85) and the device models (16.86) thus

decompose into separate 0/+/− sequence networks that can be analyzed separately, similar to per-phase analysis for balanced networks.

We illustrate the analysis of sequence networks with an example.

Example 16.13 (Sequence network analysis)**.** Consider the network shown in Figure 16.9 where a voltage source and a current source supply power through two lines to two loads in parallel. Suppose the network is symmetric (Definition 16.2) and C14.1 holds

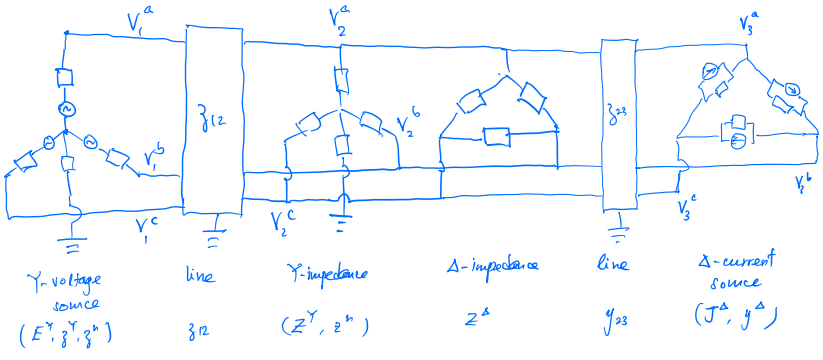


Figure 16.9 Example 16.13: Three-phase unbalanced sources supplies power two balance loads in parallel through symmetric lines.

(i.e., all neutrals are grounded and voltages are defined with respect to the ground). Given the Y-configured voltage source (E^Y, z^Y, z^n), the Δ-configured current source (J^Δ, y^Δ), the balanced impedances (z^Y, z^n), z^Δ, and the symmetric lines with series impedance matrices (z_{12}, z_{23}), calculate:

1 the terminal load voltages $V_2 := (V_2^a, V_2^b, V_2^c)$;
2 the internal current $I_2^Y := (I_2^{an}, I_2^{bn}, I_2^{cn})$ and the total complex power $\mathbf{1}^\mathsf{T} s_2^Y$ delivered to the Y-configured load;
3 the internal current $I_2^\Delta := (I_2^{ab}, I_2^{bc}, I_2^{ca})$ and the total complex power $\mathbf{1}^\mathsf{T} s_2^\Delta$ delivered to the Δ-configured load;

Solution. The network equation (16.85) and the device models (16.86) decompose into separate 0/+/− sequence networks as shown in Figure 16.10. We will first determine the terminal sequence voltage $\tilde{V}_2$ and then the *terminal* sequence currents $\tilde{I}_2^1$ and $\tilde{I}_2^2$ coming out of the Y-configured and Δ-configured impedances respectively. The *terminal* phase variables are then $V_2 = F\tilde{V}_2$, $I_2^1 = F\tilde{I}_2^1$, and $I_2^2 = F\tilde{I}_2^2$. Given these terminal variables we can determine internal currents (I_2^Y, I_2^Δ) and powers (s_2^Y, s_2^Δ) using the conversion rules.

To determine $\tilde{V}_2$, apply KCL at bus 2 of the zero-sequence networks to get

$$\frac{\tilde{E}_{1,0}^Y - \tilde{V}_{2,0}}{\left(z_1^{an} + 3z_1^n\right) + \left(z_{12}^s + 2z_{12}^m\right)} = \frac{\tilde{V}_{2,0}}{z_2^{an} + 3z_2^n} + \tilde{J}_{3,0}^\Delta \tag{16.87a}$$

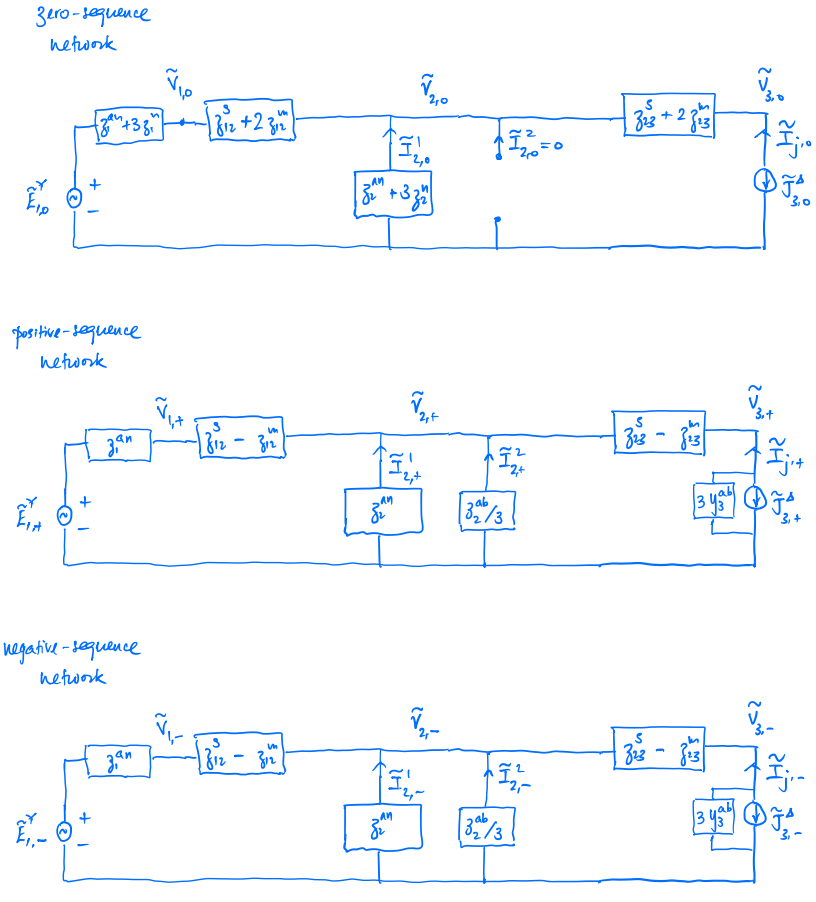


Figure 16.10 Example 16.13: Circuit models of sequence networks.

To analyze the positive and negative-sequence networks let the Thévenin equivalent load admittance be

$$\tilde{Y}_2 \;=\; y_2^{an} \,+\, 3y_2^{ab}$$

where $y_2^{an} := \left(z_2^{an}\right)^{-1}$ and $y_2^{ab} := \left(z_2^{ab}\right)^{-1}$. KCL at bus 2 of the positive-sequence network gives

$$\frac{\tilde{E}_{1,+}^{Y} - \tilde{V}_{2,+}}{z_1^{an} + \left(z_{12}^{s} - z_{12}^{m}\right)} \;=\; \tilde{Y}_2 \tilde{V}_{2,+} \,+\, 3y_3^{ab}\tilde{V}_{3,+} \,+\, \tilde{J}_{3,+}^{\Delta}$$

Hence we have, after eliminating $\tilde{V}_{3,+}$

$$\frac{\tilde{E}_{1,+}^{Y} - \tilde{V}_{2,+}}{z_1^{an} + z_{12}^{s} - z_{12}^{m}} \;=\; \left(\tilde{Y}_2 + 3\tilde{\rho}_3 y_3^{ab}\right)\tilde{V}_{2,+} \,+\, \left(1 \,-\, 3\tilde{\rho}_3 y_3^{ab}\left(z_{23}^{s} - z_{23}^{m}\right)\right)\tilde{J}_{3,+}^{\Delta} \qquad (16.87b)$$

Similarly, from the negative-sequence network, we get

$$\frac{\tilde{E}^Y_{1,-} - \tilde{V}_{2,-}}{z^{an}_1 + z^s_{12} - z^m_{12}} = \left(\tilde{Y}_2 + 3\tilde{\rho}_3 y^{ab}_3\right)\tilde{V}_{2,-} - \left(1 - 3\tilde{\rho}_3 y^{ab}_3 \left(z^s_{23} - z^m_{23}\right)\right)\tilde{J}^\Delta_{3,-} \quad (16.87c)$$

The terminal sequence voltage $\tilde{V}_2 := \left(\tilde{V}_{2,0}, \tilde{V}_{2,+}, \tilde{V}_{2,-}\right)$ can be obtained from (16.87). From the 0/+/− sequence networks, the terminal sequence load currents are

$$\tilde{I}^1_{2,0} = -\frac{\tilde{V}_{2,0}}{z^{an}_2 + 3z^n_2}, \qquad \tilde{I}^1_{2,+} = -\frac{\tilde{V}_{2,+}}{z^{an}_2}, \qquad \tilde{I}^1_{2,-} = -\frac{\tilde{V}_{2,-}}{z^{an}_2}$$
$$\tilde{I}^2_{2,0} = 0, \qquad \tilde{I}^2_{2,+} = -\frac{3\tilde{V}_{2,+}}{z^{ab}_2}, \qquad \tilde{I}^2_{2,-} = -\frac{3\tilde{V}_{2,-}}{z^{ab}_2}$$

From the terminal sequence variables $\left(\tilde{V}_2, \tilde{I}^1_2, \tilde{I}^2_2\right)$ we can obtain the terminal phase variables

$$V_2 = F\tilde{V}_2, \qquad I^1_2 = F\tilde{I}^1_2, \qquad I^2_2 = F\tilde{I}^2_2$$

To obtain the internal currents I^Y_2 and I^Δ_2, apply the conversion rules to get

$$I^Y_2 = -I^1_2, \qquad I^\Delta_2 = -\Gamma^{\mathsf{T}\dagger} I^2_2 + \beta_2 \mathbf{1} = -\frac{1}{3}\Gamma I^2_2 + \beta_2 \mathbf{1}$$

for an arbitrary $\beta \in \mathbb{C}$, where I^Δ_2 exists because $\tilde{I}^2_{2,0} = 0$ means $\mathbf{1}^{\mathsf{T}} I^2_2 = 0$.

Finally to calculate the internal powers s^Y_2 and s^Δ_2 we first obtain the internal voltages:

$$V^Y_2 = V_2 - V^n_2 \mathbf{1} = V_2 + z^n_2 \left(\mathbf{1}\mathbf{1}^{\mathsf{T}}\right) I^1_2, \qquad V^\Delta_2 = \Gamma V_2$$

where the second equality follows from $V^n_2 = -z^n_2 \left(\mathbf{1}^{\mathsf{T}} I^1_2\right)$ under C14.1. Hence

$$s^Y_2 := \operatorname{diag}\left(V^Y_2 I^{Y\mathsf{H}}_2\right) = -\operatorname{diag}\left(V_2 I^{1\mathsf{H}}_2 + z^n_2 \left(\mathbf{1}\mathbf{1}^{\mathsf{T}}\right) I^1_2 I^{1\mathsf{H}}_2\right)$$
$$s^\Delta_2 := \operatorname{diag}\left(V^\Delta_2 I^{\Delta\mathsf{H}}_2\right) = -\operatorname{diag}\left(\Gamma V_2 I^{2\mathsf{H}}_2 \Gamma^\dagger\right) + \overline{\beta}_2 \Gamma V_2$$

The total internal powers are $\mathbf{1}^{\mathsf{T}} s^Y_2$ and $\mathbf{1}^{\mathsf{T}} s^\Delta_2$ which is independent of β_2. □

16.5 Bibliographical notes

Three-phase load flow solvers have been developed since at least the 1960s, e.g., see [170] for solution in the sequence coordinate and [41, 160] in the phase coordinate. A three-phase network is equivalent to a single-phase circuit where each node in the equivalent circuit is indexed by a (bus, phase) pair [160]. The main difference with a single-phase network is the models of three-phase devices in the equivalent circuit, such as models for generators and loads studied in Chapter 14, and lines and transformers studied in Chapter 15. Single-phase power flow algorithms such as Newton Raphson [171] or Fast Decoupled methods [172] can be directly applied to the equivalent circuit.

See also [5, Chapter 11] for recent algorithms for solving three-phase power flows. A sufficient condition is derived in [173] to ensure a fixed-point iteration of an AC power flow equation converges to a unique power flow solution. Sufficient conditions are also proved in [16] for the invertibility of three-phase admittance matrix which then ensures the validity of Z-bus method for computing power flow solutions. Finally recent studies on three-phase AC optimal power flow problems and their semidefinite relaxations include e.g. [103, 104, 174].

16.6 Problems

Chapter 16.1.

Exercise 16.1 (Symmetry and block symmetry)**.** Consider a $3n \times 3n$ matrix A partitioned as in Definition 16.1.

1 Suppose A is symmetric. Show that it is block symmetric if all its off-diagonal blocks are symmetric, i.e., $A_{jk}^{\mathsf{T}} = A_{jk}$, for all $j \neq k$.
2 Suppose A is block symmetric. Show that it is symmetric if all blocks A_{jk}, including the diagonal blocks, are symmetric.

Exercise 16.2 (Invertibility of Y)**.** Prove Theorem 16.2.

Exercise 16.3 (Invertibility of Y)**.** This exercise shows that the set of conditions in Theorem 16.1 and that in Theorem 16.2 each ensures $\alpha^{\mathsf{H}} Y \alpha \neq 0$ for any nonzero $\alpha \in \mathbb{C}^{3(N+1)}$. Suppose C16.2 is satisfied, i.e., $y^s_{jk} = y^s_{kj}$, y^m_{jk} and y^m_{kj} are complex symmetric, so that the admittance matrix Y is both symmetric and block symmetric. Consider $\alpha^{\mathsf{H}} Y \alpha$ for any $\alpha \in \mathbb{C}^{3(N+1)}$, and write $y^s_{jk}, y^m_{jj} := \sum_{k:j\sim k} y^m_{jk}$ and α_j in terms of their real and imaginary parts:

$$y^s_{jk} =: g^s_{jk} + \mathbf{i} b^s_{jk} \in \mathbb{C}^{3\times 3}, \quad y^m_{jj} =: g^m_{jj} + \mathbf{i} b^m_{jj} \in \mathbb{C}^{3\times 3}, \quad \alpha_j =: \rho_j + \mathbf{i}\epsilon_j \in \mathbb{C}^3$$

1 Show that the real and imaginary parts of $\alpha^{\mathsf{H}} Y \alpha$ are:

$$\mathrm{Re}\left(\alpha^{\mathsf{H}} Y \alpha\right) = \sum_{(j,k)\in E} \left(\begin{bmatrix}\rho_j \\ \epsilon_j\end{bmatrix} - \begin{bmatrix}\rho_k \\ \epsilon_k\end{bmatrix}\right)^{\mathsf{T}} \begin{bmatrix} g^s_{jk} & 0 \\ 0 & g^s_{jk}\end{bmatrix} \left(\begin{bmatrix}\rho_j \\ \epsilon_j\end{bmatrix} - \begin{bmatrix}\rho_k \\ \epsilon_k\end{bmatrix}\right) + \sum_{j\in\overline{N}} \begin{bmatrix}\rho_j^{\mathsf{T}} & \epsilon_j^{\mathsf{T}}\end{bmatrix} \begin{bmatrix} g^m_{jj} & 0 \\ 0 & g^m_{jj}\end{bmatrix} \begin{bmatrix}$$

$$\mathrm{Im}\left(\alpha^{\mathsf{H}} Y \alpha\right) = \sum_{(j,k)\in E} \left(\begin{bmatrix}\rho_j \\ \epsilon_j\end{bmatrix} - \begin{bmatrix}\rho_k \\ \epsilon_k\end{bmatrix}\right)^{\mathsf{T}} \begin{bmatrix} b^s_{jk} & 0 \\ 0 & b^s_{jk}\end{bmatrix} \left(\begin{bmatrix}\rho_j \\ \epsilon_j\end{bmatrix} - \begin{bmatrix}\rho_k \\ \epsilon_k\end{bmatrix}\right) + \sum_{j\in\overline{N}} \begin{bmatrix}\rho_j^{\mathsf{T}} & \epsilon_j^{\mathsf{T}}\end{bmatrix} \begin{bmatrix} b^m_{jj} & 0 \\ 0 & b^m_{jj}\end{bmatrix} \begin{bmatrix}$$

2 Show that the conditions in Theorem 16.1 ensure $\alpha^{\mathsf{H}} Y \alpha \neq 0$ for any nonzero $\alpha \in \mathbb{C}^{3(N+1)}$.

3 Show that the conditions in Theorem 16.2 ensure $\alpha^{\mathsf{H}} Y \alpha \neq 0$ for any nonzero $\alpha \in \mathbb{C}^{3(N+1)}$.

Exercise 16.4 (Invertibility of Y_{22})**.** Prove Theorem 16.3.

Exercise 16.5 (Power flow equation)**.** Express the three-phase power injection $s_j \in \mathbb{C}^3$ in terms of the voltage vector $V \in \mathbb{C}^{3(N+1)}$:

$$s_j = \sum_{k:j\sim k} \operatorname{diag}\left(\left(e_j^{\mathsf{T}} \otimes \mathbb{I}\right) V V^{\mathsf{H}} \left((e_j - e_k) \otimes y_{jk}^{s\mathsf{H}}\right) + \left(e_j^{\mathsf{T}} \otimes \mathbb{I}\right) V V^{\mathsf{H}} \left(e_j \otimes y_{jk}^{m\mathsf{H}}\right)\right)$$

Chapter 16.2.

Exercise 16.6 (Four-wire model in Y-configured)**.** For Example 16.3 express the neutral voltages (γ_j, γ_k) in terms of the phase voltages and currents (V_j, V_k, I_j, I_k).

Exercise 16.7 (Four-wire model in Y-configured)**.** Repeat Example 16.5 but for the case where the neutrals n of the voltage source and the impedance are connected through impedances $\left(z_j^{n'n}, z_k^{n'n}\right)$ to their respective external neutral terminals n' which are then connected to the four-wire line. See Figure 16.6.

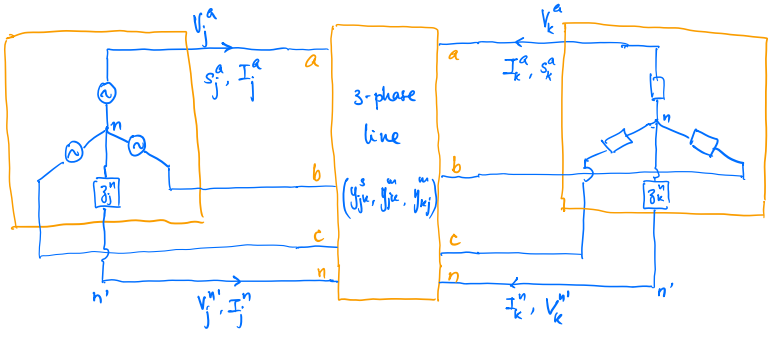


Figure 16.11 Exercise 16.7: A Y-configured generator connected through a four-wire line to a Y-configured impedance load.

Note that V_j^n is the voltage (with respect to a common reference point) at the neutral internal of the device, and $V_j^{n'}$ is the voltage at the terminal of the neutral line of the device, and that $\left(V_j^{n'}, V_k^{n'}\right)$ do not need to be given or grounded.

Exercise 16.8 (Current Source in Δ configuration)**.** Consider Example 16.6 but with an ideal current source instead of the ideal voltage source. Specifically suppose the following are specified:

- Current source $\left(J_j^{\Delta}, \gamma_j\right)$.

- Impedance z_k^{Δ}. (Note that β_k need not be specified but can be derived.)
- Line admittances $\left(\left(z_{jk}^s\right)^{-1}, y_{jk}^m = y_{kj}^m := 0\right)$. We have assumed for simplicity that shunt admittances are zero.

1 Compute all the other quantities in Table 16.2.
2 Show that if z_{jk}^s is symmetric of the form in (15.9) with $z_{jk}^1 + 2z_{jk}^2 \neq 0$, then $\gamma_k = \gamma_j$.
3 Show the following relation between the loop flows β_j and β_k:
   - $\beta_k = -\beta_j$ if and only if $z_k^{ab} J_j^{ab} + z_k^{bc} J_j^{bc} + z_k^{ca} J_j^{ca} = 0$.
   - $\beta_k = 0$ if and only if $z_k^{ab} J_j^{ab} + z_k^{bc} J_j^{bc} + z_k^{ca} J_j^{ca} = \zeta_k \beta_j$ where $\zeta_k := \mathbf{1}^{\mathsf{T}} \zeta_k \mathbf{1}$.
   - $\beta_k = 0$ if the impedance $z_k^{\Delta} = \frac{\zeta_k}{3}\mathbb{I}$ is balanced, regardless of whether J_j^{Δ} is balanced or whether β_j is zero. The converse does not necessarily hold.

Note that if the shunt admittances $\left(y_{jk}^m, y_{kj}^m\right)$ are nonzero, then γ_j need not be specified and can be derived; see Remark 16.8.

Exercise 16.9 (Y and Δ devices)**.** Consider a Y-configured current source connected to a Δ-configured impedance as shown in Figure 16.12. Suppose the following are

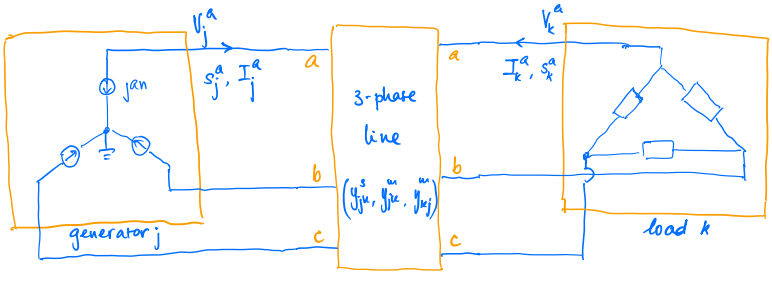


Figure 16.12 Three-phase Y-configured current source connected through a three-phase line to a Δ-configured impedance load.

specified:

- Current source J_j^Y.
- Impedance z_k^{Δ}.
- Line admittances $\left(\left(z_{jk}^s\right)^{-1}, y_{jk}^m, y_{kj}^m\right)$ with at least one of $\left(y_{jk}^m, y_{kj}^m\right)$ being nonzero.

Follow the solution strategy outlined in Chapter 16.2.3 to solve the network. State any invertibility assumptions in your derivation. An alternative approach is that used in Exercise 16.8.

Exercise 16.10 (Balanced power source)**.** Solve Example 16.9 when the system is balanced, i.e.,

- Power source $\left(\sigma_j^\Delta, \gamma_j\right)$ with $\sigma_j^\Delta = a_j\alpha_+ + b_j\mathbf{1}$ for given (a_j, b_j). i.e., a balanced power source must be a generalized balanced vector. Moreover its voltage and current $\left(V_j^\Delta, I_j^\Delta\right)$ are generalized balanced vectors.
- Impedance $z_k^\Delta := \zeta_k^\Delta \mathbb{I}$.
- Line admittances $\left(y_{jk}^s, y_{jk}^m, y_{kj}^m\right) := \left(\eta_{jk}^s \mathbb{I}, 0, 0\right)$.
- $\beta_j + \beta_k := \frac{1}{3}\mathbf{1}^\mathsf{T}\left(I_j^\Delta + I_k^\Delta\right) = \beta'$.

Use the external model (14.27b) of impedance.

Exercise 16.11 (Power sources)**.** Repeat Example 16.11 when the shunt admittances are zero, i.e., the three-phase line is specified as $\left(y_{jk}^s, y_{jk}^m = y_{kj}^m = 0\right)$ with nonsingular y_{jk}^s, as in Example 16.9. Since the admittance matrix is no longer invertible, suppose $\beta_j + \beta_k := \frac{1}{3}\mathbf{1}^\mathsf{T}\left(I_j^\Delta + I_k^\Delta\right) = \beta'$ is also given.

Exercise 16.12 (Balanced power sources)**.** Consider the system in Figure 16.7 where both the generator and load are power sources and the lines have zero shunt admittances, as in Example 16.9. Suppose the system is balanced and the following are specified:

- Power source $\left(\sigma_j^\Delta, \gamma_j\right)$ with $\sigma_j^\Delta = a_j\alpha_+ + b_j\mathbf{1}$ for given (a_j, b_j), with its voltage and current $\left(V_j^\Delta, I_j^\Delta\right)$ being generalized balanced vectors.
- Power source $\sigma_k^\Delta = a_k\alpha_+ + b_k\mathbf{1}$ for given (a_k, b_k), with its voltage and current $\left(V_j^\Delta, I_j^\Delta\right)$ being generalized balanced vectors. Note that γ_k is not specified.
- Line admittances $\left(y_{jk}^s, y_{jk}^m, y_{kj}^m\right) := \left(\eta_{jk}^s \mathbb{I}, 0, 0\right)$.
- Suppose a reference voltage $\angle V_j^a := \theta_j^a$ is given.

Show how to derive all variables $\left(V_i^\Delta, I_i^\Delta, \beta_i\right)$ and $\left(V_i, I_i, \gamma_j\right)$, $i = j, k$, *analytically*. In particular show that $\gamma_j = \gamma_k$.

Exercise 16.13 (Power sources)**.** Given a solution $\left(V_c, I_i^{\text{int}}, I_p^{\text{int}}, V_p^{\text{int}}\right)$ to the reduced system (16.47), derive all the unknown internal variables $\left(V_j^{Y/\Delta}, I_j^{Y/\Delta}, s_j^{Y/\Delta}, \beta_j\right)$ and external variables $\left(V_j, I_j, s_j, \gamma_j\right)$ over the network.

Chapter 16.3

Exercise 16.14 (Balanced network)**.** The two equivalent external models of an impedance z_j^Δ in Tables 14.3 and 14.4 are

$$\begin{aligned} V_j &= -Z^\Delta I_j + \gamma_j \mathbf{1}, & \mathbf{1}^\mathsf{T} I_j &= 0 \\ I_j &= -Y^\Delta V_j \end{aligned}$$

where the effective impedance and admittance matrices are $Z_j^\Delta := \frac{1}{9}\Gamma^\mathsf{T} z_j^\Delta \Gamma$ and and $Y_j^\Delta := \Gamma^\mathsf{T} y_j^\Delta \Gamma$. For balanced networks where the impedance $z_j^\Delta = \epsilon_j^{-1}\mathbb{I}$, show that these models reduce to:

$$\begin{aligned} V_j &= -\frac{1}{3\epsilon_j} I_j + \gamma_j \mathbf{1}, & \mathbf{1}^\mathsf{T} I_j &= 0 \\ I_j &= -3\epsilon_j \left(V_j - \gamma_j \mathbf{1}\right) \end{aligned}$$

Exercise 16.15 (Balanced voltages & currents)**.** Consider the reduced system (16.42) of (16.48)(16.50). We have shown that any solution $\left(V_c, I_i^{\text{int}}\right)$ of (16.42) consists of generalized balanced vectors. Derive all other variables analytically in terms of the solution $\left(V_c, I_i^{\text{int}}\right)$ and show that they are generalized balanced positive-sequence sets.

Exercise 16.16 (Balanced network)**.** Suppose $(A \times \mathbb{I})V = b \otimes \alpha_+ + c \otimes \mathbf{1}$ where $A \in \mathbb{C}^{n\times n}$, $b, c \in \mathbb{C}^n$, $\mathbb{I}$ is the identity matrix of size 3 and $\mathbf{1}$ is the vector of all 1s of size 3. Let $\gamma_j := \frac{1}{3}\mathbf{1}^\mathsf{T} V_j$ be the zero-sequence component of $V_j \in \mathbb{C}^3$. Show that $A\gamma = c$.

Chapter 16.4.

Exercise 16.17. Prove that if a vector V of three-phase voltages is a balanced negative sequence then the negative-sequence voltage $\tilde{V}_- = \sqrt{3}V_a$ and the zero-sequence and the positive-sequence voltages are both zero, $\tilde{V}_0 = \tilde{V}_+ = 0$.

Exercise 16.18 (Sequence impedance $\tilde{Z}^Y$)**.** Consider the phase impedance matrix $Z^Y := z^Y + z^n \mathbf{1}\mathbf{1}^\mathsf{T}$ of a Y-configured impedance z^Y. Show that its sequence impedance matrix is

$$\tilde{Z}^Y = \frac{1}{3}\begin{bmatrix} \mathbf{1}^T z & \alpha_+^\mathsf{T} z & \alpha_-^\mathsf{T} z \\ \alpha_-^\mathsf{T} z & \mathbf{1}^T z & \alpha_+^\mathsf{T} z \\ \alpha_+^\mathsf{T} z & \alpha_-^\mathsf{T} z & \mathbf{1}^T z \end{bmatrix} + \begin{bmatrix} 3z^n & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

If $z^{an} = z^{bn} = z^{cn}$ then

$$\tilde{Z}^Y \;=\; \begin{bmatrix} z^{an}+3z^n & 0 & 0 \\ 0 & z^{an} & 0 \\ 0 & 0 & z^{an} \end{bmatrix}$$

Exercise 16.19 (Sequence impedance $\tilde{Z}^\Delta$)**.** Consider a Δ-configured impedance z^Δ whose external model is (from (16.75)):

$$V \;=\; -Z^\Delta I + \gamma \mathbf{1}, \qquad\qquad \mathbf{1}^{\mathsf{T}} I \;=\; 0 \tag{16.88}$$

where the zero-sequence voltage $\gamma := \frac{1}{3}\mathbf{1}^{\mathsf{T}} V$ is also a variable to be determined and

$$Z^\Delta \;:=\; \frac{1}{9}\Gamma^{\mathsf{T}} \underbrace{z^\Delta \left(\mathbb{I} - \frac{1}{\zeta}\mathbf{1}\tilde{z}^{\Delta\mathsf{T}}\right)}_{\hat{z}^\Delta} \Gamma$$

Show that its sequence impedance matrix is

$$\tilde{Z}^\Delta \;:=\; \frac{1}{9}(F\Lambda)^{\mathsf{H}} \hat{z}^\Delta (F\Lambda)$$

where F is given in (**??**) and

$$\Lambda \;:=\; \begin{bmatrix} 0 & & \\ & 1-\alpha & \\ & & 1-\alpha^2 \end{bmatrix}$$

If $z^{ab} = z^{bc} = z^{ca}$ then

$$\tilde{Z}^\Delta \;=\; \frac{z^{ab}}{3}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and the external model of the Δ-configured impedance in the sequence coordinate is:

$$\begin{bmatrix} 0 \\ \tilde{V}_+ \\ \tilde{V}_- \end{bmatrix} \;=\; -\frac{z^{ab}}{3}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} \tilde{I}_0 \\ \tilde{I}_+ \\ \tilde{I}_- \end{bmatrix}, \qquad\qquad \tilde{I}_0 \;=\; 0$$

Exercise 16.20 (Sequence network: Δ-configured voltage source)**.** One of the external models of a Δ-configured voltage source is (from (14.21b)):

$$V \;=\; \hat{\Gamma} E^\Delta \,-\, Z^\Delta I + \gamma \mathbf{1}, \qquad\qquad \mathbf{1}^{\mathsf{T}} I \;=\; 0$$

where

$$\hat{\Gamma} \;:=\; \frac{1}{3}\Gamma^{\mathsf{T}}\left(\mathbb{I} - \frac{1}{\zeta}\tilde{z}^\Delta \mathbf{1}^{\mathsf{T}}\right), \qquad\qquad Z^\Delta \;:=\; \frac{1}{9}\Gamma^{\mathsf{T}} z^\Delta \left(\mathbb{I} - \frac{1}{\zeta}\mathbf{1}\tilde{z}^{\Delta\mathsf{T}}\right)\Gamma$$

where $\tilde{z}^\Delta := \mathrm{diag}\left(z^\Delta\right)\mathbf{1}$ and $\zeta := \mathbf{1}^{\mathsf{T}}\tilde{z}^\Delta$.

1 Show that

Exercise 16.21 (Sequence network: Δ-configured voltage source)**.** Repeat Exercise 16.20 starting with the alternative external models of a Δ-configured voltage source is (from (14.21a)).

Exercise 16.22 (Sequence network: Y-configured current source)**.** Suppose assumption C16.1 holds (all neutrals are grounded and voltages are defined with respect to the ground) so that $V^n = -z^n \left(\mathbf{1}^\mathsf{T} I\right)$. Derive the sequence networks for a Y-configured current source (as those in Chapter 16.4.3) starting from the external model in the phase domain (from (14.15b)):

$$V \;=\; -\left(z^Y J^Y \;+\; Z^Y I\right)$$

where $z^Y := \left(y^Y\right)^{-1}$ and $Z^Y := z^Y \;+\; z^n \mathbf{1}\mathbf{1}^\mathsf{T}$.

Exercise 16.23. Consider the complex symmetric matrix

$$M := \begin{bmatrix} 1 & i \\ i & -1 \end{bmatrix}$$

Show that M is not diagonalizable by computing its Jordan form and that:

1 Its eigenvalue $\lambda = 0$ has algebraic multiplicity of 2 and geometric multiplicity of 1.
2 Its eigenvector is $v_1 = (-i, 1)$ and generalized eigenvector is $v_2 = (-2i, 1)$.

Exercise 16.24. Consider the complex symmetric phase impedance matrix

$$z := \begin{bmatrix} s & m & m \\ m & s & m \\ m & m & s \end{bmatrix}$$

where $s, m \in \mathbb{C}$.

1 Check directly that $zz^\mathsf{H} = z^\mathsf{H} z$. Hence, even though z is symmetric but not Hermitian, it is normal.
2 Since z is normal, it is unitarily similar to a diagonal matrix $\tilde{z}$, i.e., there exists a unitary matrix F such that $\tilde{z} = F^\mathsf{H} z F$. Find F and $\tilde{z}$.

Exercise 16.25 (Unbalanced currents)**.** Consider a balanced load in (a) Y configuration, or (b) Δ configuration, with one of the loads open-circuited, as shown in Figure 16.13. Find the sequence currents $\tilde{I} := (\tilde{I}_1, \tilde{I}_2, \tilde{I}_3)$ and the neutral current I_n (for Y configuration) when the terminal phase currents are

$$I = \begin{bmatrix} i_a \\ i_a e^{\mathbf{i}2\pi/3} \\ I_c \end{bmatrix}$$

Why is only the negative-sequence component nonzero even though the loads are unbalanced because of the open circuit?

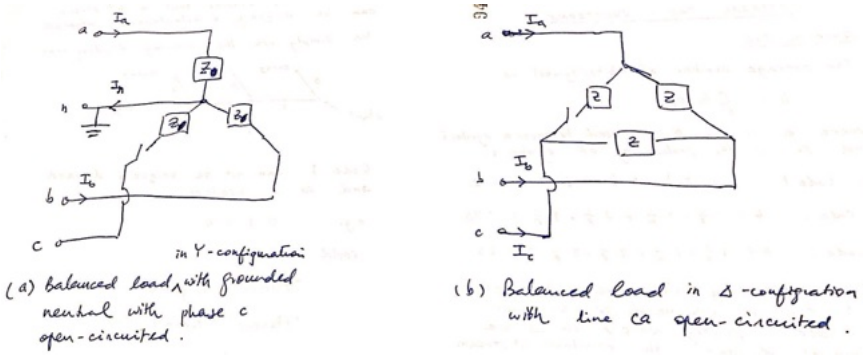


Figure 16.13 Sequence components of unbalanced phase currents.

Exercise 16.26. Repeat Example 16.13 without using symmetrical components and sequence networks.

Exercise 16.27. Repeat Example 16.13 but with the Y and Δ-impedances in series (instead of in parallel) connected by a line with the same series-phase impedance matrix z_{line}, as shown in Figure 16.14.

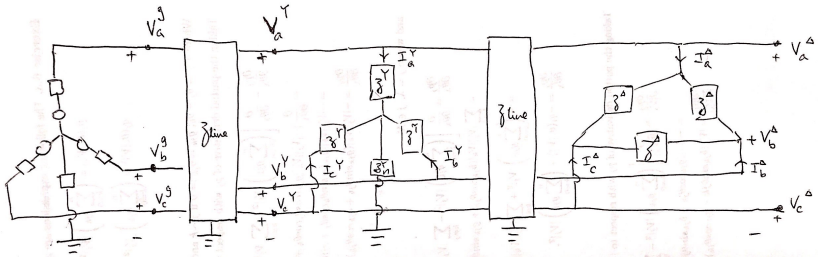


Figure 16.14 Exercise 16.27: A three-phase unbalanced voltage source supplies power two balance loads in series through symmetric lines.

Exercise 16.28. Repeat Exercise 16.27 without using symmetrical components and sequence networks.

17 Branch flow models: radial networks

In this chapter we extend the single-phase branch flow models of Chapter 5 to unbalanced three-phase networks. We will build on materials in Chapter 16 on unbalanced bus injection models.

17.1 Three-phase BFM for radial networks

17.1.1 Line model

We use the three-phase line model of Chapter 16.1.1 where each line $(j,k) \in E$ characterized by four 3×3 series and shunt admittance matrices, $\left(y^s_{jk}, y^m_{jk}\right)$ from j to k and $\left(y^s_{kj}, y^m_{kj}\right)$ from k to j, that define the relation between (V_j, V_k) and (I_{jk}, I_{kj}):

$$\begin{bmatrix} I_{jk} \\ I_{kj} \end{bmatrix} = \underbrace{\begin{bmatrix} y^s_{jk} + y^m_{jk} & -y^s_{jk} \\ -y^s_{kj} & y^s_{kj} + y^m_{kj} \end{bmatrix}}_{Y_{jk}} \begin{bmatrix} V_j \\ V_k \end{bmatrix}$$

We emphasize that y^s_{jk} and y^s_{kj} may be different (i.e., Y_{jk} may not be block symmetric) and y^m_{jk} and y^m_{kj} may be different. Moreover, when (j,k) models a three-phase transformer, any of these 3×3 admittance matrices may be singular and the shunt admittances (y^m_{jk}, y^m_{kj}) of the line model are generally nonzero even when the shunt admittances of the constituent single-phase transformers are assumed zero; see Remark 16.1 and (16.2)(16.3) for line parameters when (j,k) models a three-phase transformer. Therefore we assume $\left(y^s_{jk}, y^m_{jk}\right)$ and $\left(y^s_{kj}, y^m_{kj}\right)$ are given for each line $(j,k) \in E$, but series impedance matrices $z^s_{jk} := \left(y^s_{jk}\right)^{-1}$ and $z^s_{kj} := \left(y^s_{kj}\right)^{-1}$ may not exist. Generally we will write power flow equations in terms of the series admittance matrices instead of the series impedance matrices (unless the series admittance matrices are assumed nonsingular).

17.1.2 With shunt admittances

To extend the branch flow model (5.1) for single-phase networks to unbalanced three-phase networks define the following variables:

$$\begin{array}{rrr} s_j \in \mathbb{C}^3, & v_j \in \mathbb{S}^3_+, & j \in \overline{N} \\ \ell_{jk}, \ell_{kj} \in \mathbb{S}^3_+, & S_{jk}, S_{kj} \in \mathbb{C}^{3\times 3}, & (j,k) \in E \end{array}$$

where $\mathbb{S}^n_+ \subseteq \mathbb{C}^{n\times n}$ is the set of of $n \times n$ complex (Hermitian and) positive semidefinte matrices. It will become clear later that v_j, ℓ_{jk}, S_{jk} are rank-1 matrices. The diagonal entries of v_j are the squared magnitudes of the nodal voltages (V_j^a, V_j^b, V_j^c), the diagonal entries of ℓ_{jk} are the squared magnitudes of the sending-end line currents $(I_{jk}^a, I_{jk}^b, I_{jk}^c)$, the diagonal entries of S_{jk} are the sending-end line power flows $(S_{jk}^a, S_{jk}^b, S_{jk}^c)$, and similarly in the opposite direction. Let $s := (s_j, j \in \overline{N}), v := (v_j, j \in \overline{N})$, $\ell := (\ell_{jk}, \ell_{kj}, (j,k) \in E), S := (S_{jk}, S_{kj}, (j,k) \in E)$, and let $x := (s, v, \ell, S) \in \mathbb{C}^{12(N+1)+36M}$. Define for each $(j,k) \in E$ the total admittance matrix

$$\tilde{y}_{jk} := y_{jk}^s + y_{jk}^m, \qquad \tilde{y}_{kj} := y_{kj}^s + y_{kj}^m$$

Hence $\tilde{y}_{jk} = y_{jk}^s$ and $\tilde{y}_{kj} = y_{kj}^s$ if and only if $y_{jk}^m = y_{kj}^m = 0 \in \mathbb{C}^{3\times 3}$. The extension of (5.1) to an unbalanced three-phase network is the following model:

$$s_j = \sum_{k:j\sim k} \operatorname{diag}(S_{jk}), \qquad j \in \overline{N} \tag{17.1a}$$

$$\tilde{y}_{jk} v_j \tilde{y}_{jk}^{\mathsf{H}} - y_{jk}^s v_k \left(y_{jk}^s\right)^{\mathsf{H}} = 2\operatorname{Re}\left(\tilde{y}_{jk} S_{jk}\right) - \ell_{jk}, \qquad (j,k) \in E \tag{17.1b}$$

$$\tilde{y}_{kj} v_k \tilde{y}_{kj}^{\mathsf{H}} - y_{kj}^s v_j \left(y_{kj}^s\right)^{\mathsf{H}} = 2\operatorname{Re}\left(\tilde{y}_{kj} S_{kj}\right) - \ell_{kj}, \qquad (j,k) \in E \tag{17.1c}$$

$$\begin{bmatrix} v_j & S_{jk} \\ S_{jk}^{\mathsf{H}} & \ell_{jk} \end{bmatrix} \succeq 0, \quad \begin{bmatrix} v_k & S_{kj} \\ S_{kj}^{\mathsf{H}} & \ell_{kj} \end{bmatrix} \succeq 0, \qquad (j,k) \in E \tag{17.1d}$$

$$\operatorname{rank}\begin{bmatrix} v_j & S_{jk} \\ S_{jk}^{\mathsf{H}} & \ell_{jk} \end{bmatrix} = 1, \quad \operatorname{rank}\begin{bmatrix} v_k & S_{kj} \\ S_{kj}^{\mathsf{H}} & \ell_{kj} \end{bmatrix} = 1, \qquad (j,k) \in E \tag{17.1e}$$

$$y_{kj}^s \left(v_j \tilde{y}_{jk}^{\mathsf{H}} - S_{jk}\right) = \left(v_k \tilde{y}_{kj}^{\mathsf{H}} - S_{kj}\right)^{\mathsf{H}} \left(y_{jk}^s\right)^{\mathsf{H}}, \qquad (j,k) \in E \tag{17.1f}$$

$$\operatorname{col}\left(v_j \tilde{y}_{jk}^{\mathsf{H}} - S_{jk}\right)^{\mathsf{H}} \subseteq \operatorname{range}\left(y_{jk}^s\right), \quad \operatorname{col}\left(v_k \tilde{y}_{kj}^{\mathsf{H}} - S_{kj}\right)^{\mathsf{H}} \subseteq \operatorname{range}\left(y_{kj}^s\right), \quad (j,k) \in E \tag{17.1g}$$

where $V_0 \in \mathbb{C}^3$ is given and $v_0 := V_0 V_0^{\mathsf{H}}$ and $\operatorname{col} A$ denotes the columns of A. These equations extend (5.1) from single-phase to three-phase networks and express the same four properties that a power flow solution $x := (s, v, \ell, S)$ satisfies:

1 *Power balance*: Unlike the power balance equation (5.1a), (17.1a) constrains only the diagonal terms of 3×3 matrices S_{jk}. Their off-diagonal terms are determined jointly with the other equations.
2 *Ohm's law*: (17.1b)(17.1c) originate from the Ohm's law $I_{jk} = y^s_{jk}(V_j - V_k) + y^m_{jk} V_j$, but unlike (5.1b)(5.1c), (17.1b)(17.1c) use only admittance matrices (y^s_{jk}, y^m_{jk}) and (y^s_{kj}, y^m_{kj}), but not impedance matrices because these admittances may be singular (e.g., when they model transformers in Δ configuration).
3 *Apparent power*: The explicit definition (5.1d) of apparent power for single-phase networks becomes the implicit psd rank-1 conditions (17.1d)(17.1e). They ensure the existence of (V_j, I_{jk}) so that

$$v_j = V_j V_j^{\mathsf{H}}, \qquad \ell_{jk} = I_{jk} I_{jk}^{\mathsf{H}}, \qquad S_{jk} = V_j I_{jk}^{\mathsf{H}} \tag{17.2a}$$

or equivalently

$$\begin{bmatrix} v_j & S_{jk} \\ S_{jk}^{\mathsf{H}} & \ell_{jk} \end{bmatrix} = \begin{bmatrix} V_j \\ I_{jk} \end{bmatrix} \cdot \begin{bmatrix} V_j^{\mathsf{H}} & I_{jk}^{\mathsf{H}} \end{bmatrix}, \qquad j \to k \in E \tag{17.2b}$$

as well as the quantities in the opposite direction. The vectors $\left(V_j, I_{jk}\right)$ are unique up to a reference angle $\varphi_{jk} \in (-\pi, \pi]$, one for each $(j,k) \in E$. When the network graph is a tree and the linear cycle condition (17.1f) is satisfied, φ_{jk} as well as φ_{kj} in the opposite direction are the same for all lines $(j,k) \in E$. Moreover a given V_0 at the reference bus 0 will fix the angles of all variables in x, as discussed in the proof of Theorem 17.1 and in Chapter 17.2.3. [1] See also Example 17.1 in Chapter 17.3.
4 *Cycle condition*: The linear cycle condition (5.1e) becomes (17.1f)(17.1g). The condition (17.1g) is linear and equivalent to: $\left(v_j \tilde{y}_{jk}^{\mathsf{H}} - S_{jk}\right)^{\mathsf{H}} = y^s_{jk} w$ for some matrix $w \in \mathbb{C}^{3\times 3}$. It is necessary because any of admittance matrices (y^s_{jk}, y^m_{jk}) and (y^s_{kj}, y^m_{kj}) may be singular, the main challenge in extending (5.1) to three-phase networks. If (y^s_{jk}, y^s_{kj}) are nonsingular then multiplying both sides of (17.1f) by $z^s_{kj} := \left(y^s_{jk}\right)^{-1}$ and taking Hermitian transpose we obtain

$$\left(v_j \tilde{y}_{jk}^{\mathsf{H}} - S_{jk}\right)^{\mathsf{H}} = y^s_{jk} \left(v_k \tilde{y}_{kj}^{\mathsf{H}} - S_{kj}\right) z_{jk}^{s\mathsf{H}}$$

which implies (17.1g); similarly in the opposite direction k to j. Therefore if (y^s_{jk}, y^s_{kj}) are nonsingular then the condition (17.1g) is vacuous. We will discuss

[1] A fixed V_0 is needed in the equivalence Theorem 17.1. A given V_0 also enables Algorithm 5 in Chapter 17.2.3 that explicitly constructs voltage and current phasors (V, I) from a power flow solution $x := (s, v, \ell, S)$ of (17.1), and enables a backward forward sweep method in Chapter 17.4.2. Note however that fixing V_0 may not guarantee the uniqueness of power flow solutions x since (17.1) is nonlinear.

in Chapter 17.2.3 the role of tree topology, cycle conditions, and angle recovery after we have extended (17.1) to general networks that may contain cycles.

Like the single-phase model (5.1) for radial networks, (17.1) does not require $y^s_{jk} = y^s_{kj}$ (assumption assumption C17.1 below) and allows nonzero shunt admittances (y^m_{jk}, y^m_{kj}). It is therefore suitable for modeling three-phase transformers in standard configurations in addition to distribution and short transmission lines (line parameters when (j,k) models a three-phase transformer are given in (16.2)(16.3)). If the admittances $\left(y^s_{jk}, y^m_{jk}\right)$ and $\left(y^s_{kj}, y^m_{kj}\right)$ are nonzero scalars then (17.1) reduces to (5.1) for single-phase networks.

17.1.3 Without shunt admittances

Suppose the following condition holds:

C17.1: For every line $(j,k) \in E$, the series admittance matrices satisfy $y^s_{jk} = y^s_{kj}$.

This means that the $3(N+1) \times 3(N+1)$ admittance matrix Y is block symmetric and has a three-phase Π circuit representation. We also assume that the shunt admittances $y^m_{jk} = y^m_{kj} = 0$ as well in which case the admittance matrix Y has zero block row sums. In this case (j,k) can model a distribution or short transmission line, but is not suitable for modeling a transformer since their shunt admittances are generally nonzero; see Remark 16.1. Hence in this case we assume series impedance matrices $z^s_{jk} := \left(y^s_{jk}\right)^{-1}$ exist for all $(j,k) \in E$. This allows us to adopt a directed graph for network model since in this case

$$S_{jk} + S_{kj} \;=\; z^s_{jk} \ell_{jk}, \qquad \ell_{jk} \;=\; \ell_{kj} \tag{17.3}$$

and use line variables (ℓ_{jk}, S_{jk}) (only) on each directed line $j \to k$.

Substituting (17.3) into (17.1) leads to the following model proposed in [104] that generalizes DistFlow equations from the single-phase to the three-phase setting (Exercise 17.2):

$$\sum_{k:j\to k} \operatorname{diag}(S_{jk}) \;=\; \sum_{i:i\to j} \operatorname{diag}\left(S_{ij} - z^s_{ij}\ell_{ij}\right) + s_j, \qquad j \in \overline{N} \tag{17.4a}$$

$$v_j - v_k \;=\; \left(z^s_{jk} S^{\mathsf{H}}_{jk} + S_{jk} z^{s\mathsf{H}}_{jk}\right) - z^s_{jk}\, \ell_{jk}\, z^{s\mathsf{H}}_{jk}, \qquad j \to k \in E \tag{17.4b}$$

$$\begin{bmatrix} v_j & S_{jk} \\ S^{\mathsf{H}}_{jk} & \ell_{jk} \end{bmatrix} \succeq 0, \qquad \operatorname{rank} \begin{bmatrix} v_j & S_{jk} \\ S^{\mathsf{H}}_{jk} & \ell_{jk} \end{bmatrix} \;=\; 1, \qquad j \to k \in E \tag{17.4c}$$

where $V_0 \in \mathbb{C}^3$ is given. In particular the cycle condition (17.1f) becomes vacuous when assumption C17.1 holds and shunt admittances are zero.

Angle recovery.

We now explain how to recover the phase angles for voltage and current phasors (V, I) for a radial network with zero shunt admittance matrices $y^m_{jk} = y^m_{kj} = 0$ and under assumption C5.1, i.e., given a power solution $x = (s, v, \ell, S)$ that satisfies (17.4) we will construct the phasors (V, I).

The BFM (17.4) does not contain the vectors V_j or I_{jk}, but the psd rank-1 constraints (17.4c) ensure that there exist V_j and I_{jk} such that

$$v_j \;=\; V_j V_j^{\mathsf{H}}, \qquad \ell_{jk} \;=\; I_{jk} I_{jk}^{\mathsf{H}}, \qquad S_{jk} \;=\; V_j I_{jk}^{\mathsf{H}} \tag{17.5a}$$

or equivalently

$$\begin{bmatrix} v_j & S_{jk} \\ S_{jk}^{\mathsf{H}} & \ell_{jk} \end{bmatrix} \;=\; \begin{bmatrix} V_j \\ I_{jk} \end{bmatrix} \cdot \begin{bmatrix} V_j^{\mathsf{H}} & I_{jk}^{\mathsf{H}} \end{bmatrix}, \qquad j \to k \in E \tag{17.5b}$$

Given matrices $\left(v_j, \ell_{jk}, S_{jk}\right)$, the vectors $\left(V_j, I_{jk}\right)$ are determined uniquely up to a reference angle. If a reference angle is given, e.g., $\angle V_0^a = 0°$, the power flow equation (17.4) will fix the angles of all variables. See Example 17.1 in Chapter 17.3.

If V_0 is given, not just V_0^ϕ, $\phi \in \{a, b, c\}$, then given a power solution $x := (s, v, \ell, S)$ that satisfies (17.4), an $\tilde{x} := (s, V, I, \ell, S) \in \tilde{\mathbb{X}}$ can be explicitly constructed using the iterative Algorithm 5 from [104] that makes use of the tree topology. The basic idea in Step 5 of the algorithm is to compute the phasors V_i and I_{ij} recursively, starting from bus 0 when V_0 is given: since $S_{ij} = V_i I_{ij}^{\mathsf{H}}$, taking the Hermitian transpose and multiplying both sides by V_i, we have

$$V_i I_{ij}^{\mathsf{H}} \;=\; S_{ij} \;\;\Rightarrow\;\; I_{ij}\left(V_i^{\mathsf{H}} V_i\right) \;=\; S_{ij}^{\mathsf{H}} V_i \;\;\Rightarrow\;\; I_{ij} \;=\; \frac{1}{\mathrm{tr}(v_i)}\, S_{ij}^{\mathsf{H}} V_i \tag{17.6}$$

Tree topology and cycle condition.

An x satisfying (17.4) is a legitimate power flow solution, i.e., from which a unique (up to an arbitrary reference angle) phasor (V, I) can be constructed as described above, only if the network is radial. To see this, substituting $I_{jk} = y_{jk}\left(V_j - V_k\right)$ into $S_{jk} = V_j I_{jk}^{\mathsf{H}}$ we get

$$V_j V_k^{\mathsf{H}} \;=\; v_j - S_{jk} z_{jk}^{\mathsf{H}}, \qquad j \to k \in E$$

Taking the diagonal vectors on both sides, we conclude that given a solution x of (17.4), voltage phasors V_j exist if and only if there exist $\theta_j := \left(\theta_j^a, \theta_j^b, \theta_j^c\right)$, for all $j \in \overline{N}$, such

Algorithm 4: Recover $\tilde{x} = (s, V, I, \ell, S)$ from $x = (s, v, \ell, S)$.
Down orientation where all lines point away from root bus 0.

Input: $x = (s, v, \ell, S) \in \mathbb{X}$; $V_0 \in \mathbb{C}^3$.
Output: $\tilde{x} = (\tilde{s}, \tilde{V}, \tilde{I}, \tilde{\ell}, \tilde{S}) \in \tilde{\mathbb{X}}$

1: $\tilde{s} \leftarrow s$; $\tilde{\ell} \leftarrow \ell$; $\tilde{S} \leftarrow S$;
2: $N_{\text{visit}} \leftarrow \{0\}$;
3: **while** $N_{\text{visit}} \neq \overline{N}$ **do**
4: find $i \to j$ such that $i \in N_{\text{visit}}$ and $j \notin N_{\text{visit}}$;
5: compute

$$\begin{aligned} \tilde{I}_{ij} &\leftarrow \frac{1}{\operatorname{tr}(v_i)} S_{ij}^{\mathsf{H}} \tilde{V}_i \\ \tilde{V}_j &\leftarrow \tilde{V}_i - z_{ij} \tilde{I}_{ij} \\ N_{\text{visit}} &\leftarrow N_{\text{visit}} \cup \{j\} \end{aligned}$$

6: **end while**

that

$$\begin{bmatrix} |V_j^a V_k^a| e^{\mathbf{i}\left(\theta_j^a - \theta_k^a\right)} \\ |V_j^b V_k^b| e^{\mathbf{i}\left(\theta_j^b - \theta_k^b\right)} \\ |V_j^c V_k^c| e^{\mathbf{i}\left(\theta_j^c - \theta_k^c\right)} \end{bmatrix} = \begin{bmatrix} |U_{jk}^a| e^{\mathbf{i}\beta_{jk}^a} \\ |U_{jk}^b| e^{\mathbf{i}\beta_{jk}^b} \\ |U_{jk}^c| e^{\mathbf{i}\beta_{jk}^c} \end{bmatrix}, \qquad j \to k \in E$$

where the vectors $\beta_{jk} := \beta_{jk}(x) \in \mathbb{R}^3$ of angles depend on x and are defined by $\beta_{jk}(x) := \angle \operatorname{diag}\left(v_j - S_{jk} z_{jk}^{\mathsf{H}}\right)$. In particular there must exist $\theta := \left(\theta_j \in \mathbb{R}^3, j \in \overline{N}\right) \in \mathbb{R}^{3(N+1)}$ such that

$$\beta(x) = \left(C^{\mathsf{T}} \otimes \mathbb{I}\right) \theta \tag{17.7a}$$

where $\beta(x) := \left(\beta_{jk}(x), j \to k \in E\right) \in \mathbb{C}^{3M}$ and C is the $(N+1) \times M$ bus-by-line incidence matrix whose rank is N. See Chapter A.11 for more properties of C. The condition (17.17) is the cycle condition that generalizes (**??**) from single-phase to three-phase networks. We now show that the cycle condition is vacuous for radial networks, i.e., any x satisfying (17.4) also satisfies (17.17) when the network is radial.

Partition C into its first row c_0^{T} and an $N \times M$ matrix $\hat{C}$ of the remaining rows so that

$$C^{\mathsf{T}} =: \begin{bmatrix} c_0 & \hat{C}^{\mathsf{T}} \end{bmatrix}$$

Similarly partition $\theta =: \left(\theta_0, \hat{\theta}\right) \in \mathbb{R}^{3(N+1)}$. Suppose G is a (connected) tree with $M = N$. Then $\hat{C}^{\mathsf{T}}$ is $N \times N$ and of full rank. Therefore $c_0 = \hat{C}^{\mathsf{T}} \eta$ for some $\eta \in \mathbb{C}^N$. It is proved in Exercise 17.1 that $\left(\hat{C}^{\mathsf{T}} \eta\right) \otimes \mathbb{I} = \left(\hat{C}^{\mathsf{T}} \otimes \mathbb{I}\right)(\eta \otimes \mathbb{I})$. Hence (17.17) becomes

$$\beta(x) = \left(c_0^{\mathsf{T}} \otimes \mathbb{I}\right) \theta_0 + \left(\hat{C}^{\mathsf{T}} \otimes \mathbb{I}\right) \hat{\theta} = \left(\hat{C}^{\mathsf{T}} \otimes \mathbb{I}\right)\left(\hat{\theta} + (\eta \otimes \theta_0)\right) \tag{17.7b}$$

where we have used $(\eta \otimes \mathbb{I})\,\theta_0 = \eta \otimes \theta_0$. Since $\hat{C}^\mathsf{T}$ and hence $\left(\hat{C}^\mathsf{T} \otimes \mathbb{I}\right)$ are invertible, for any x satisfying (17.4), there always exists an $\theta = (\theta_0, \hat{\theta}) \in \mathbb{R}^{3(N+1)}$ that satisfies (17.21a). Indeed the solution θ of (17.21a) is not unique. Given any $\theta_0 \in \mathbb{C}^3$, there is always a (unique) $\hat{\theta} := \left(\left(\hat{C}^\mathsf{T}\right)^{-1} \otimes \mathbb{I}\right)\beta(x) - \eta \otimes \theta_0$ that satisfies (17.21a).[2]

If G contains cycles, on the other hand, then $M > N$ and the $3M \times 3(N+1)$ matrix $\left(C^\mathsf{T} \otimes \mathbb{I}\right)$ in (17.17) has a column rank of $3N < 3M$ since rank $(A \otimes B) =$ rank $A \cdot$ rank B from Lemma 16.6. This means that the column space of $\left(C^\mathsf{T} \otimes \mathbb{I}\right)$ does not span $\mathbb{R}^{3M}$ and hence there may be $\beta(x)$ for which no θ exists that satisfies (17.17), regardless of whether θ_0 is given. A power flow model for a meshed network consists of (17.4) augmented with the cycle condition (17.17).

17.2 Equivalence, cycle condition and angle recovery

The branch flow models for an unbalanced three-phase radial networks are (17.1) with shunt admittances and without assumption C17.1 and the generalized DistFlow equations (17.4) when shunt admittances are zero and assumption C17.1 holds. We will show that they are equivalent to the bus injection model (16.12b) studied in Chapter 16.1.4, reproduced here:

$$s_j = \sum_{k:j\sim k} \operatorname{diag}\left(V_j (V_j - V_k)^\mathsf{H} \left(y^s_{jk}\right)^\mathsf{H} + V_j V_j^\mathsf{H} \left(y^m_{jk}\right)^\mathsf{H}\right), \qquad j \in \overline{N} \tag{17.8}$$

To this end we first extend the branch flow models (17.1) and (17.4) to general networks possibly with cycles. We then use these generalized branch flow models as a bridge to relate BFM (17.1) and (17.4) for radial networks to BIM (17.8) for general networks.

17.2.1 Extension to general networks

To extend the branch flow model (5.20) for a general network possibly with cycles from the single-phase setting to the unbalanced three-phase setting, define the following variables:

$$\begin{aligned} s_j &\in \mathbb{C}^3, & V_j &\in \mathbb{C}^3, & & & j &\in \overline{N} \\ I_{jk}, I_{kj} &\in \mathbb{C}^3, & \ell_{jk}, \ell_{kj} &\in \mathbb{S}^3_+, & S_{jk}, S_{kj} &\in \mathbb{C}^{3\times 3}, & (j,k) &\in E \end{aligned}$$

where $\mathbb{S}^n_+ \subseteq \mathbb{C}^{n\times n}$ is the set of of $n \times n$ complex (Hermitian and) positive semidefinte matrices. Let $s := (s_j, j \in \overline{N})$, $V := (V_j, j \in \overline{N})$, $I := (I_{jk}, I_{kj}, (j,k) \in E)$, $\ell := (\ell_{jk}, \ell_{kj}, (j,k) \in E)$ and $S := (S_{jk}, S_{kj}, (j,k) \in E)$. Let $\tilde{x} := (s, V, I, \ell, S) \in$

[2] Here the vector θ_0 can be arbitrary to satisfy (17.21a) whereas a single angle e.g. θ^a_0 fixes all other angles in (17.2). This is because (17.2) uses the matrix v_j whereas (17.21) uses only the diagonal entries of v_j.

$\mathbb{C}^{6(N+1)+42M}$. The branch flow model for a general three-phase network is the following power flow equations in $\tilde{x}$:

$$s_j = \sum_{k:j\sim k} \operatorname{diag}\left(S_{jk}\right), \qquad j \in \overline{N} \tag{17.9a}$$

$$I_{jk} = \tilde{y}_{jk} V_j - y^s_{jk} V_k, \qquad I_{kj} = \tilde{y}_{kj} V_k - y^s_{kj} V_j, \qquad (j,k) \in E \tag{17.9b}$$

$$\ell_{jk} = I_{jk} I^{\mathsf{H}}_{jk}, \qquad \ell_{kj} = I_{kj} I^{\mathsf{H}}_{kj}, \qquad (j,k) \in E \tag{17.9c}$$

$$S_{jk} = V_j I^{\mathsf{H}}_{jk}, \qquad S_{kj} = V_k I^{\mathsf{H}}_{kj}, \qquad (j,k) \in E \tag{17.9d}$$

where $\tilde{y}_{jk} := y^s_{jk} + y^m_{jk}$ and $\tilde{y}_{kj} = y^s_{kj} + y^m_{kj}$. The equation (17.9a) imposes power balance at each bus, (17.9b) describes the Kirchhoff's and Ohm's laws, (17.9c) defines the squared current magnitude matrices, and (17.9d) defines branch power in terms of the associated voltage and current. A key to generalizing single-phase BFM to the 3-phase setting is the generalization in (17.9c)(17.9d) of the quadratic relation between between $\left(S_{jk}, \ell_{jk}\right)$ and $\left(V_j, I_{jk}\right)$ using outer products. This relation is explicit in BFM (17.9) for general networks that include voltage and current angles, but is implicit in BFMs for radial networks that do not include voltage and current angles (see (17.1d)(17.1e) and (17.4c)). For convenience we assume here the vector V_0, not just V_0^ϕ, $\phi \in \{a,b,c\}$, is given (see angle recovery in Chapter 17.1.3 and a backward forward sweep method in Chapter 17.4.2). Since this model does not require assumption C17.1 and allows nonzero shunt admittance matrices (y^m_{jk}, y^m_{kj}), it is suitable for modeling three-phase transformers in YY, $\Delta\Delta$, ΔY and $Y\Delta$ configurations.

When assumption C17.1 holds and shunt admittance matrices $y^m_{jk} = y^m_{kj} = 0 \in \mathbb{C}^{3\times 3}$, we may assume series impedance matrices $z^s_{jk} := \left(y^s_{jk}\right)^{-1}$ exist for all $(j,k) \in E$. This allows us to adopt a directed graph and obtain the following simpler BFM by substituting (17.3) into (17.9):

$$s_j + \sum_{i:i\to j} \operatorname{diag}\left(S_{ij} - z_{ij}\ell_{ij}\right) = \sum_{k:j\to k} \operatorname{diag}(S_{jk}), \qquad j \in \overline{N} \tag{17.10a}$$

$$V_j - V_k = z_{jk} I_{jk}, \qquad j \to k \in E \tag{17.10b}$$

$$\ell_{jk} = I_{jk} I^{\mathsf{H}}_{jk}, \qquad j \to k \in E \tag{17.10c}$$

$$S_{jk} = V_j I^{\mathsf{H}}_{jk}, \qquad j \to k \in E \tag{17.10d}$$

with a given $V_0 \in \mathbb{C}^3$. In this case the line variables are directed with $s := (s_j, j \in \overline{N})$, $V := (V_j, j \in \overline{N})$, $I := (I_{jk}, j \to \in E)$, $\ell := (\ell_{jk}, j \to k \in E)$ $S := (S_{jk}, j \to k \in E)$, and $\tilde{x} := (s, V, I, \ell, S) \in \mathbb{C}^{6(N+1)+21M}$.

17.2.2 Equivalence of BFM and BIM

We now show that BFMs (17.1) and (17.4) for radial networks and (17.9) and (17.10) for general networks are all equivalent to the BIM (17.8), in the following sense. Define

the solution sets:

$$\mathbb{V} := \mathbb{V}(V_0) := \left\{ (s,V) \in \mathbb{C}^{6(N+1)} \mid (s,V) \text{ satisfies (17.8) with a given } V_0 \right\}$$

$$\tilde{\mathbb{X}} := \tilde{\mathbb{X}}(V_0) := \left\{ \tilde{x} := (s,V,I,\ell,S) \in \mathbb{C}^{6(N+1)+42M} \mid \tilde{x} \text{ satisfies (17.9) with a given } V_0 \right\}$$

$$\mathbb{X}_{\text{tree}} := \mathbb{X}_{\text{tree}}(V_0) := \left\{ x := (s,v,\ell,S) \in \mathbb{C}^{12(N+1)+36M} \mid x \text{ satisfies (17.1) with given } V_0 \text{ and } v_0 = V_0 V_0^{\mathsf{H}} \right\}$$

where $N+1$ is the number of nodes and $M := |E|$ is the number of lines in G. We say that two sets A and B are *equivalent*, denoted by $A \equiv B$, if there is a bijection between them. When assumption C17.1 holds and shunt admittance matrices $y^m_{jk} = y^m_{kj} = 0 \in \mathbb{C}^{3\times 3}$, the branch flow model (17.1) reduces to (17.4) for radial networks and (17.9) reduces to (17.10) for general networks. It therefore suffices to prove the equivalence of $\mathbb{V}$, $\tilde{\mathbb{X}}$ and $\mathbb{X}_{\text{tree}}$.

The following theorem generalizes Theorem 5.2 of Chapter 5.2 from single-phase to unbalanced three-phase networks.

Theorem 17.1. Suppose the network G is connected and V_0 at the reference bus 0 is given.

1 Then $\mathbb{V} \equiv \tilde{\mathbb{X}}$.
2 If G is a tree then $\tilde{\mathbb{X}} \equiv \mathbb{X}_{\text{tree}}$.

Proof *Part 1:* $\mathbb{V} \equiv \tilde{\mathbb{X}}$. Fix any $(s,V) \in \mathbb{V}$. We will construct an $\tilde{x} := (s,V,I,\ell,S) \in \tilde{\mathbb{X}}$, i.e. $\tilde{x}$ satisfies (17.9). Define (I,ℓ,S) in terms of V by (17.9b)(17.9c)(17.9d). Therefore, to show that $\tilde{x} \in \tilde{\mathbb{X}}$, it suffices to show that $\tilde{x}$ also satisfies (17.9a). Since (s,V) satisfies (17.8) we have

$$s_j = \sum_{k:j\sim k} \operatorname{diag}\left(V_j (\tilde{y}_{jk} V_j - y^s_{jk} V_k)^{\mathsf{H}} \right) = \sum_{k:j\sim k} \operatorname{diag}\left(S_{jk} \right)$$

where the second equality follows from (17.9b)(17.9d). Therefore $\tilde{x}$ also satisfies (17.9) and hence is in $\tilde{\mathbb{X}}$. Conversely, if $\tilde{x} := (s,V,I,\ell,S)$ satisfies (17.9) then substituting (17.9b)(17.9d) into (17.9a) yields (17.8). Hence $(s,V) \in \mathbb{V}$.

Part 2: $\tilde{\mathbb{X}} \equiv \mathbb{X}_{tree}$. We explicitly construct a bijection between these two sets. Fix any $\tilde{x} := (s,V,I,\ell,S)$ with the given V_0 that satisfies (17.9). The mapping $\tilde{x} \mapsto x$ is defined by $x := (s,v,\ell,S)$ where $v := (v_j, j \in \overline{N})$ with

$$v_j := V_j V_j^{\mathsf{H}} \tag{17.11}$$

We first show that x satisfies (17.1). Then we show that the mapping $\tilde{x} \mapsto x$ defined by (17.11) is injective (when the network is connected and V_0 is given) and surjective (when the network is a tree and the linear cycle condition (17.1f)(17.1g) is satisfied). It is therefore a bijection between $\tilde{\mathbb{X}}$ and $\mathbb{X}_{\text{tree}}$.

First x clearly satisfies (17.1a). To prove (17.1b), we have from (17.9b) and (17.9c)

$$\ell_{jk} = \tilde{y}_{jk} v_j \tilde{y}^{\mathsf{H}}_{jk} + y^s_{jk} v_k y^{s\mathsf{H}}_{jk} - 2\operatorname{Re}\left(\tilde{y}_{jk} V_j V_k^{\mathsf{H}} y^{s\mathsf{H}}_{jk} \right) \tag{17.12}$$

We have from (17.9b) and (17.9d) $S_{jk} = v_j \tilde{y}_{jk}^{\mathsf{H}} - V_j V_k^{\mathsf{H}} y_{jk}^{s\mathsf{H}}$ and hence

$$\tilde{y}_{jk} V_j V_k^{\mathsf{H}} y_{jk}^{s\mathsf{H}} \;=\; \tilde{y}_{jk} v_j \tilde{y}_{jk}^{\mathsf{H}} - \tilde{y}_{jk} S_{jk}$$

Substituting into (17.12) yields

$$\ell_{jk} \;=\; \tilde{y}_{jk} v_j \tilde{y}_{jk}^{\mathsf{H}} + y_{jk}^{s} v_k y_{jk}^{s\mathsf{H}} - 2\mathrm{Re}\left(\tilde{y}_{jk} v_j \tilde{y}_{jk}^{\mathsf{H}} - \tilde{y}_{jk} S_{jk}\right) \;=\; -\tilde{y}_{jk} v_j \tilde{y}_{jk}^{\mathsf{H}} + y_{jk}^{s} v_k y_{jk}^{s\mathsf{H}} + 2\mathrm{Re}\left(\tilde{y}_{jk} S_{j}\right.$$

which is (17.1b). Similarly (17.1c) follows from (17.9b)(17.9c)(17.9d). To prove the cycle condition (17.1f)(17.1g), use again $S_{jk} = v_j \tilde{y}_{jk}^{\mathsf{H}} - V_j V_k^{\mathsf{H}} y_{jk}^{s\mathsf{H}}$ and $S_{kj} = v_k \tilde{y}_{kj}^{\mathsf{H}} - V_k V_j^{\mathsf{H}} y_{kj}^{s\mathsf{H}}$ to obtain

$$\left(v_j \tilde{y}_{jk}^{\mathsf{H}} - S_{jk}\right)^{\mathsf{H}} \;=\; y_{jk}^{s} V_k V_j^{\mathsf{H}}, \qquad \left(v_k \tilde{y}_{kj}^{\mathsf{H}} - S_{kj}\right)^{\mathsf{H}} \;=\; y_{kj}^{s} V_j V_k^{\mathsf{H}}$$

$$y_{kj}^{s}\left(v_j \tilde{y}_{jk}^{\mathsf{H}} - S_{jk}\right) \;=\; y_{kj}^{s} V_j V_k^{\mathsf{H}} y_{jk}^{s\mathsf{H}}, \qquad y_{jk}^{s}\left(v_k \tilde{y}_{kj}^{\mathsf{H}} - S_{kj}\right) \;=\; y_{jk}^{s} V_k V_j^{\mathsf{H}} y_{kj}^{s\mathsf{H}}$$

which implies the cycle conditions (17.1g) and (17.1f) respectively. Finally (17.11) and (17.9c)(17.9d) implies

$$\begin{bmatrix} v_j & S_{jk} \\ S_{jk}^{\mathsf{H}} & \ell_{jk} \end{bmatrix} \;=\; \begin{bmatrix} V_j \\ I_{jk} \end{bmatrix} \begin{bmatrix} V_j^{\mathsf{H}} & I_{jk}^{\mathsf{H}} \end{bmatrix}, \qquad \begin{bmatrix} v_k & S_{kj} \\ S_{kj}^{\mathsf{H}} & \ell_{kj} \end{bmatrix} \;=\; \begin{bmatrix} V_k \\ I_{kj} \end{bmatrix} \begin{bmatrix} V_k^{\mathsf{H}} & I_{kj}^{\mathsf{H}} \end{bmatrix}$$

which implies the psd and rank-1 conditions (17.1d)(17.1e). This completes the proof that x satisfies (17.1).

When the network is a (connected) tree and V_0 is fixed (not just $\angle V_0^a$ say), Lemma 17.2 shows that the mapping $\tilde{x} \mapsto x$ is injective. Lemma 17.3 shows that the mapping is surjective, i.e., given $x := (s, v, \ell, S)$ that satisfies (17.1) (in particular the cycle condition), there is a $\tilde{x}$ that satisfies (17.9). Hence it is bijective.

□

17.2.3 Tree topology, cycle condition, angle recovery

Consider the mapping $\tilde{x} := (s, V, I, \ell, S) \mapsto x := (s, v, \ell, S)$ from $\tilde{\mathbb{X}}$ to $\mathbb{X}_{\text{tree}}$ defined through (17.11). Suppose V_0 is given. In this subsection we show that the mapping is injective when the network graph G is connected, it is surjective when G is a (connected) tree and $x \in \mathbb{X}_{\text{tree}}$ satisfies the linear cycle condition (17.1f)(17.1g), and the linear cycle condition becomes vacuous when assumption C17.1 holds and shunt admittances are zero, i.e., any x satisfying (17.4) also satisfies (17.17) for a radial network.

Connected graph G and uniqueness of (V, I).

Lemma 17.2 (Injectivity)**.** Suppose the network graph G is a (connected) tree and V_0 is given. Given $x := (s, v, \ell, S) \in \mathbb{X}_{\text{tree}}$, if there exists $(V(x), I(x))$ such that $\tilde{x} := (s, V(x), I(x), \ell, S) \in \tilde{\mathbb{X}}$, then $(V(x), I(x))$ is unique.

Proof For the sake of contradiction, suppose both $\tilde{x} := (\tilde{s}, \tilde{V}, \tilde{I}, \tilde{\ell}, \tilde{S}) \in \tilde{\mathbb{X}}$ and $\hat{x} = (\hat{s}, \hat{V}, \hat{I}, \hat{\ell}, \hat{S}) \in \tilde{\mathbb{X}}$, with $\tilde{V}_0 = \hat{V}_0$, are mapped to $x = (s, v, \ell, S) \in \mathbb{X}_{\text{tree}}$ through (17.11). By definition of $\tilde{x} \mapsto x$ we have $\tilde{s} = s = \hat{s}$, $\tilde{\ell} = \ell = \hat{\ell}$, $\tilde{S} = S = \hat{S}$. Moreover $\tilde{V}_j \tilde{V}_j^{\mathsf{H}} = v_j = \hat{V}_j \hat{V}_j^{\mathsf{H}}$ for all $j \in \overline{N}$. We have to show that $\tilde{V} = \hat{V}$ and $\tilde{I} = \hat{I}$. Since the psd rank-1 decomposition (17.2) is unique up to an arbitrary phase, $(\tilde{V}_j, \tilde{I}_{jk})$ and $(\hat{V}_j, \hat{I}_{jk})$ can differ only by an arbitrary phase shift φ_{jk} for each (j,k), and $(\tilde{V}_k, \tilde{I}_{kj})$ and $(\hat{V}_k, \hat{I}_{kj})$ can differ only by an arbitrary phase shift φ_{kj} for each (j,k). We argue that φ_{jk} and φ_{kj} must be the same for all lines $(j,k) \in E$ as long as the network is connected. Moreover $\hat{V}_0 = \tilde{V}_0$ implies that $\varphi_{jk} = \varphi_{kj} = 0$ for all $(j,k) \in E$.

It is convenient to assume (only) in this proof, without loss of generality, a graph orientation and, since the graph is a tree, assume that all lines point *towards* bus 0. Start from a leaf node i and consider a line $i \to j \in E$. Let

$$\hat{V}_i \;=\; \tilde{V}_i\, e^{\mathbf{i}\varphi_{ij}}, \qquad \hat{I}_{ij} \;=\; \tilde{I}_{ij}\, e^{\mathbf{i}\varphi_{ij}} \tag{17.13}$$

Similarly, for all lines $j \to k$ connected to $j \neq 0$, we have $\hat{V}_j \;=\; \tilde{V}_j\, e^{\mathbf{i}\varphi_{jk}}$. Substituting $\hat{V}_i, \hat{V}_j, \hat{I}_{ij}$ into (17.9b) for $j \neq 0$ yields

$$\tilde{I}_{ij} \;=\; \tilde{y}_{ij}\tilde{V}_i - y^s_{ij}\tilde{V}_j\, e^{\mathbf{i}(\varphi_{jk} - \varphi_{ij})}$$

which, together with (17.9b), implies $\varphi_{jk} = \varphi_{ij}$ for all directed lines $j \to k$ (we assume without loss of generality that all angles are projected to $(-\pi, \pi]$). When $j = 0$ (i.e., there is no line $j \to k$), we have $\tilde{I}_{i0} = \tilde{y}_{i0}\tilde{V}_i - y^s_{i0}\tilde{V}_0 e^{-\mathbf{i}\varphi_{i0}}$ since $\hat{V}_0 = \tilde{V}_0$ by assumption, and hence (17.9b) implies that $\varphi_{i0} = 0$. Propagating towards bus 0 in a reverse breadth-first search order, we conclude that $\varphi_{jk} = \varphi_{i0} = 0$ on all directed lines $j \to k \in E$ since the network is connected. This implies in particular that $\hat{V}_j = \tilde{V}_j$ for all $j \in \overline{N}$. For each directed line $j \to k \in E$, since $(\tilde{V}_k, \tilde{I}_{kj})$ and $(\hat{V}_k, \hat{I}_{kj})$ in the opposite direction can differ only by a phase shift φ_{kj} for each (j,k), $\hat{V}_k = \tilde{V}_k$ for all k implies that $\varphi_{kj} = 0$ for all $j \to k \in E$. Hence $\tilde{x} = \hat{x}$ and the mapping $\tilde{x} \mapsto x$ is injective. □

Tree graph G, cycle condition and existence of (V, I).

We now explain that a phasor $(V(x), I(x))$ and hence $\tilde{x} := (s, V(x), I(x), \ell, S) \in \tilde{\mathbb{X}}$ can be recovered from an $x := (s, v, \ell, S) \in \mathbb{X}_{\text{tree}}$ if and only if the network is radial, as well as the role the linear cycle condition (17.1f)(17.1g) plays in the recovery.

We recall the following simple property of psd rank-1 matrices. Given a psd rank-1 matrix $A \in \mathbb{C}^{n\times n}$, there exists a vector $x \in \mathbb{C}^n$, unique up to an arbitrary angle, such that $xx^{\mathsf{H}} = A$, i.e., $A_{ij} = |x_i||x_j|e^{\mathbf{i}(\theta_i - \theta_j)}$. Therefore x can be determined explicitly from the given A as follows. Let $x =: |x_i|e^{\mathbf{i}\theta_i}$. Then $|x_i| = \sqrt{A_{ii}}$. To determine θ_i, define a graph $G := (N, E)$ induced by A with n nodes and m directed lines (with arbitrary graph orientation) where there is a line $i \to j$ if and only if $A_{ij} \neq 0$. Let C denote the $n \times m$ incidence matrix of G. Let $\beta_{jk}(A) := \angle A_{jk}$ for $j \to k \in E$ and let $\beta(A) := (\beta_{jk}(A), j \to k \in E)$. Then

$$\beta(A) \;=\; C^{\mathsf{T}}\theta \qquad \text{and} \qquad x_i := \sqrt{A_{ii}}\, e^{\mathbf{i}\theta_i} \tag{17.14}$$

i.e., if $\beta(A) = C^{\mathsf{T}}\theta$ for some θ (i.e., $\beta(A)$ is the row space of the incidence matrix C), then x given by (17.14) is the rank-1 decomposition of the psd rank-1 matrix $A = xx^{\mathsf{H}}$, unique up to a reference angle.

Fix a power flow solution $x \in \mathbb{X}_{\text{tree}}$. First note that if an $\tilde{x} \in \tilde{\mathbb{X}}$ exists with $v_j = V_j V_j^{\mathsf{H}}$, then (17.9b) and (17.9d) implies $S_{jk} = v_j \tilde{y}_{jk}^{\mathsf{H}} - V_j V_k^{\mathsf{H}} y_{jk}^{s\mathsf{H}}$ and hence

$$V_j V_k^{\mathsf{H}} \;=\; \left(v_j \tilde{y}_{jk}^{\mathsf{H}} - S_{jk}\right)\left(y_{jk}^{s\mathsf{H}}\right)^{\dagger} + \xi_{jk} \;\in\; \mathbb{C}^{3\times 3}, \qquad (j,k) \in E \tag{17.15a}$$

where $\left(y_{jk}^{s\mathsf{H}}\right)^{\dagger}$ denotes the pseudo-inverse of $y_{jk}^{s\mathsf{H}}$ and $\xi_{jk}^{\mathsf{H}} \in \mathbb{C}^{3\times 3}$ is the component of $V_k V_j^{\mathsf{H}}$ in $\text{null}(y_{jk}^{s})$ so that $\xi_{jk} y_{jk}^{s\mathsf{H}} = 0 \in \mathbb{C}^{3\times 3}$ (see Theorem A.19 on psudo-inverse in Appendix A.7).[3] Similarly in the opposite direction we have

$$V_k V_j^{\mathsf{H}} \;=\; \left(v_k \tilde{y}_{kj}^{\mathsf{H}} - S_{kj}\right)\left(y_{kj}^{s\mathsf{H}}\right)^{\dagger} + \xi_{kj} \;\in\; \mathbb{C}^{3\times 3}, \qquad (j,k) \in E \tag{17.15c}$$

for some $\xi_{kj} \in \mathbb{C}^{3\times 3}$, dependent on $V_k V_j^{\mathsf{H}}$, such that $\text{col}(\xi_{kj}^{\mathsf{H}}) \subseteq \text{null}(y_{kj}^{s})$.

Motivated by (17.15) when $\tilde{x} \in \tilde{\mathbb{X}}$ exists with $v_j = V_j V_j^{\mathsf{H}}$, define the $3(N+1)\times 3(N+1)$ matrix $b(x)$ by

$$[b(x)]_{jj} \;=\; v_j, \qquad j \in \overline{N} \tag{17.16a}$$

$$[b(x)]_{jk} \;=\; \left(v_j \tilde{y}_{jk}^{\mathsf{H}} - S_{jk}\right)\left(y_{jk}^{s\mathsf{H}}\right)^{\dagger}, \qquad (j,k) \in E \tag{17.16b}$$

$$[b(x)]_{kj} \;=\; \left(v_k \tilde{y}_{kj}^{\mathsf{H}} - S_{kj}\right)\left(y_{kj}^{s\mathsf{H}}\right)^{\dagger}, \qquad (j,k) \in E \tag{17.16c}$$

$$[b(x)]_{jk} \;=\; 0 \in \mathbb{C}^{3\times 3}, \qquad [b(x)]_{kj} \;=\; 0 \in \mathbb{C}^{3\times 3}, \qquad (j,k) \notin E \tag{17.16d}$$

Then, since $\tilde{x} \in \tilde{\mathbb{X}}$ with $v_j = V_j V_j^{\mathsf{H}}$, (17.15) implies that

$$b(x) \;=\; VV^{\mathsf{H}} \tag{17.16e}$$

Conversely (17.14) says that, if $b(x)$ is psd rank-1, then its rank-1 decomposition V can be uniquely recovered from $b(x)$, given V_0. Specifically let $\beta_{jk}(x) := \text{diag}\left(\angle\,[b(x)]_{jk}\right)$ for $(j,k) \in E$ and $\beta(x) := (\beta_{jk}(x), (j,k) \in E)$. Note that $b(x)$ being psd implies that

[3] Let $y_{jk}^{s} =: \begin{bmatrix} U_r \\ U_{3-r} \end{bmatrix}\begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} W_r^{\mathsf{H}} \\ W_{3-r}^{\mathsf{H}} \end{bmatrix}$. Then Theorem A.19 implies

$$\left(y_{jk}^{s}\right)^{\dagger} y_{jk}^{s}\left(V_k V_j^{\mathsf{H}}\right) \;=\; V_k V_j^{\mathsf{H}} - \underbrace{W_{3-r} W_{3-r}^{\mathsf{H}}\left(V_k V_j^{\mathsf{H}}\right)}_{\xi_{jk}^{\mathsf{H}}} \tag{17.15b}$$

and hence $\text{col}\left(\xi_{jk}^{\mathsf{H}}\right) \subseteq \text{range}(W_{3-r}) = \text{null}(y_{jk}^{s})$. Indeed ξ_{jk}^{H} is exactly the component of $V_k V_j^{\mathsf{H}}$ that is in $\text{null}(y_{jk}^{s})$. Recall that the pseudo-inverse $\left(y_{jk}^{s}\right)^{\dagger}$ is a bijection from $\text{range}(y_{jk}^{s})$ to $\text{range}(y_{jk}^{s\mathsf{H}})$ and $\left(y_{jk}^{s}\right)^{\dagger} y_{jk}^{s}\left(V_k V_j^{\mathsf{H}}\right)$ projects $V_k V_j^{\mathsf{H}}$, which may contain components in $\text{null}(y_{jk}^{s})$, onto the orthogonal subspace $\text{range}(y_{jk}^{s\mathsf{H}})$ of $\text{null}(y_{jk}^{s})$. Therefore (17.15b) says that the effect of $\left(y_{jk}^{s}\right)^{\dagger} y_{jk}^{s}$ is to remove the component ξ_{jk}^{H} of $V_k V_j^{\mathsf{H}}$ in $\text{null}(y_{jk}^{s})$. Of course if y_{jk}^{s} is nonsingular then $\xi_{jk} = 0$.

$[b(x)]_{kj} = \left([b(x)]_{jk}\right)^{\mathsf{H}}$ in (17.16), and hence $\beta_{kj}(x) := \operatorname{diag}\left(\angle\,[b(x)]_{kj}\right)$ satisfies $\beta_{kj}(x) = -\beta_{kj}(x)$. Suppose there exists $\theta_j := (\theta_j^a, \theta_j^b, \theta_j^c)$ for $j \in \overline{N}$ and $\theta := (\theta_j, j \in \overline{N}) \in (-\pi, \pi]^{3(N+1)}$ that satisfy

$$\beta(x) \;=\; \left(C^{\mathsf{T}} \otimes \mathbb{I}\right)\theta \tag{17.17}$$

where C is the $(N+1) \times M$ incidence matrix whose rank is $M = N$ ($C \otimes \mathbb{I}$ is the $3(N+1) \times 3M$ incidence matrix of the single-phase equivalent circuit). Then, given V_0, the unique rank-1 decomposition V of $b(x)$ is given by, from (17.14),

$$V_j^{\phi} \;:=\; \sqrt{v_j^{\phi\phi}}\, e^{\mathbf{i}\theta_j^{\phi}(x)}, \qquad \phi \in \{a,b,c\},\; j \in \overline{N} \tag{17.18a}$$

where $\theta(x)$ is a solution of (17.17). This is abbreviated as

$$V_j(x) \;:=\; \operatorname{diag}\left(\sqrt{v_j}\right) \odot e^{\mathbf{i}\theta_j(x)}, \qquad j \in \overline{N} \tag{17.18b}$$

where $\odot$ denotes componentwise product. Define $(I_{jk}(x), I_{kj}(x))$ in terms of x and V_0:

$$I_{jk}^{\phi}(x) \;:=\; \sqrt{\ell_{jk}^{\phi\phi}}\, e^{\mathbf{i}\left(\theta_j^{\phi'}(x) - \angle S_{jk}^{\phi'\phi}\right)}, \qquad \phi \in \{a,b,c\},\; (j,k) \in E \tag{17.19a}$$

$$I_{kj}^{\phi}(x) \;:=\; \sqrt{\ell_{kj}^{\phi\phi}}\, e^{\mathbf{i}\left(\theta_k^{\phi'}(x) - \angle S_{kj}^{\phi'\phi}\right)}, \qquad \phi \in \{a,b,c\},\; (j,k) \in E \tag{17.19b}$$

The ϕ' in (17.19) can be any phase in $\{a,b,c\}$ because $S_{jk}^{\phi'\phi} = V_j^{\phi'} \bar{I}_{jk}^{\phi}$.

We now show that, given $x \in \mathbb{X}_{\text{tree}}$, for the existence of $\tilde{x} \in \tilde{\mathbb{X}}$, it is not only necessary but also sufficient that $x \in \mathbb{X}_{\text{tree}}$ satisfies (17.17). The is the nonlinear cycle condition that generalizes (5.21e) from single-phase to three-phase networks. The $\theta(x)$ guaranteed by (17.17) can be used to define voltage and current angles and hence the phasors $(V(x), I(x))$ according to (17.18)(17.19). Moreover (17.17) reduces to the linear cycle condition (17.1f)(17.1g) when the network graph is a tree. We will see later that it becomes vacuous when assumption C17.1 holds and shunt admittances are zero.

Lemma 17.3 (Surjectivity and tree graph G)**.** Consider a general network graph G possibly with cycles and suppose V_0 is given.

1. Suppose an arbitrary $x := (s, v, \ell, S)$ satisfies (17.1a)–(17.1e) (without the linear cycle condition (17.1f)(17.1g)) and the nonlinear cycle condition (17.17). Let $\theta(x)$ denote a solution of (17.17) and construct $(V(x), I(x))$ according to (17.18)(17.19). Then $\tilde{x} := (s, V(x), I(x), \ell, S) \in \tilde{\mathbb{X}}$.
2. Suppose G is a (connected) tree. If $x \in \mathbb{X}_{\text{tree}}$ (in particular x satisfies the linear cycle condition (17.1f)(17.1g)), then x satisfies (17.17) and hence $\tilde{x} \in \tilde{\mathbb{X}}$.

Proof *Part 1.* Fix an arbitrary x that satisfies (17.1a)–(17.1e) and the nonlinear cycle condition (17.17). Construct $(V(x), I(x))$ from x and V_0 according to (17.18)(17.19). The psd rank-1 conditions (17.1d)(17.1e) means that the matrices in (17.1d)(17.1e) has a unique rank-1 decomposition (17.2), given V_0. Moreover the decomposition

$(V(x), I(x))$ is determined according to (17.14). In particular v_j and (ℓ_{jk}, ℓ_{kj}) are psd rank-1 matrices and therefore

$$v_j = V_j(x)V_j^{\mathsf{H}}(x), \quad j \in \overline{N}, \quad \ell_{jk} = I_{jk}(x)I_{jk}^{\mathsf{H}}(x), \quad \ell_{kj} = I_{kj}(x)I_{kj}^{\mathsf{H}}(x) \quad (j,k) \in E$$

where $V(x)$ and $(I_{jk}(x), I_{kj}(x))$ are given by (17.18)(17.19) respectively. Moreover (17.2) means that $S_{jk} = V_j(x)I_{jk}^{\mathsf{H}}(x)$ and $S_{kj} = V_k(x)I_{kj}^{\mathsf{H}}(x)$, specifically $S_{jk}^{\phi_1\phi_2} = \sqrt{v_j^{\phi_1\phi_1}\ell_{jk}^{\phi_2\phi_2}}\exp\left(\mathbf{i}\angle S_{jk}^{\phi_1\phi_2}\right)$.

Therefore to show that $\tilde{x} := (s, V(x), I(x), \ell, S) \in \tilde{\mathbb{X}}$, it suffices to show that $\tilde{x}$ satisfies (17.9b). Define $\hat{I}_{jk} := \tilde{y}_{jk}V_j(x) - y_{jk}^s V_k(x)$ and $\hat{I}_{kj}$ in the opposite direction in terms of $V(x)$. Let $\hat{I} := (\hat{I}_{jk}, \hat{I}_{kj}, (j,k) \in E)$. We will show that

$$\ell_{jk} = \hat{I}_{jk}\hat{I}_{jk}^{\mathsf{H}}, \qquad S_{jk} = V_j(x)\hat{I}_{jk}^{\mathsf{H}}, \qquad (j,k) \in E$$

and similarly in the opposite direction. Since the rank-1 decomposition (17.2) is unique given V_0, this implies $I_{jk}(x) = \hat{I}_{jk}$ and $I_{kj}(x) = \hat{I}_{kj}$, proving (17.9b).

To show $S_{jk} = V_j(x)\hat{I}_{jk}^{\mathsf{H}}$ we have

$$V_j(x)\hat{I}_{jk}^{\mathsf{H}} = v_j\tilde{y}_{jk}^{\mathsf{H}} - V_jV_k^{\mathsf{H}}y_{jk}^{s\mathsf{H}} \tag{17.20}$$

Recall from (17.16) that V satisfies $V(x)V^{\mathsf{H}}(x) = b(x)$ and hence, in view of the discussion following (17.15a), we have

$$V_jV_k^{\mathsf{H}}y_{jk}^{s\mathsf{H}} = v_j\tilde{y}_{jk}^{\mathsf{H}} - S_{jk} + \eta_{jk}$$

where $\eta_{jk} \in \mathbb{C}^{3\times 3}$ has the property $\eta_{jk}y_{jk}^s = 0 \in \mathbb{C}^{3\times 3}$ (see Theorem A.19 on psudo-inverse in Appendix A.7). Substituting into (17.20) yields $S_{jk} = V_j(x)\hat{I}_{jk}^{\mathsf{H}} - \eta_{jk}$.

Given the psd rank-1 matrix ℓ_{jk}, we will show that $\hat{I}_{jk} := \tilde{y}_{jk}V_j(x) - y_{jk}^s V_k(x)$ satisfies $\ell_{jk} = \hat{I}_{jk}(x)\hat{I}_{jk}^{\mathsf{H}}(x)$ using (17.1b).

Since the rank-1 decomposition of ℓ_{jk} is unique given V_0, $I_{jk}(x)$ must be equal to $\hat{I}_{jk}(x)$, proving (17.9b). We have

$$\begin{aligned}\hat{I}_{jk}\hat{I}_{jk}^{\mathsf{H}} &= \left(\tilde{y}_{jk}V_j(x) - y_{jk}^s V_k(x)\right)\left(\tilde{y}_{jk}V_j(x) - y_{jk}^s V_k(x)\right)^{\mathsf{H}} \\ &= \tilde{y}_{jk}v_j\tilde{y}_{jk}^{\mathsf{H}} + y_{jk}^s v_k y_{jk}^{s\mathsf{H}} - 2\mathrm{Re}\left(\tilde{y}_{jk}V_j(x)V_k^{\mathsf{H}}(x)y_{jk}^{s\mathsf{H}}\right)\end{aligned}$$

Note that $\tilde{y}_{jk}v_j\tilde{y}_{jk}^{\mathsf{H}}$ is a real matrix (since it equals its real part), and hence

$$\begin{aligned}\hat{I}_{jk}\hat{I}_{jk}^{\mathsf{H}} &= -\tilde{y}_{jk}v_j\tilde{y}_{jk}^{\mathsf{H}} + y_{jk}^s v_k y_{jk}^{s\mathsf{H}} + 2\mathrm{Re}\left(\tilde{y}_{jk}v_j\tilde{y}_{jk}^{\mathsf{H}} - \tilde{y}_{jk}V_j(x)V_k^{\mathsf{H}}(x)y_{jk}^{s\mathsf{H}}\right) \\ &= -\tilde{y}_{jk}v_j\tilde{y}_{jk}^{\mathsf{H}} + y_{jk}^s v_k y_{jk}^{s\mathsf{H}} + 2\mathrm{Re}\left(\tilde{y}_{jk}V_j(x)\left(\tilde{y}_{jk}V_j(x) - y_{jk}^s V_k(x)\right)^{\mathsf{H}}\right) \\ &= -\tilde{y}_{jk}v_j\tilde{y}_{jk}^{\mathsf{H}} + y_{jk}^s v_k y_{jk}^{s\mathsf{H}} + 2\mathrm{Re}\left(\tilde{y}_{jk}V_j(x)\hat{I}_{jk}^{\mathsf{H}}\right)\end{aligned}$$

We have

$$\begin{aligned}\left\|I_{jk}(x)-\tilde{y}_{jk}V_j(x)+y^s_{jk}V_k(x)\right\|_2^2 &= \ell_{jk}+\tilde{y}_{jk}v_j\tilde{y}^{\mathsf{H}}_{jk}+y^s_{jk}v_k y^{s\mathsf{H}}_{jk}-2\mathrm{Re}\left(\tilde{y}_{jk}S_{jk}\right)+2\mathrm{Re}\left(y^s_{jk}V_k(x)I^{\mathsf{H}}_{j}\right.\\ &\quad -2\mathrm{Re}\left(\tilde{y}_{jk}V_j(x)V^{\mathsf{H}}_k(x)y^{s\mathsf{H}}_{jk}\right)\\ &= 2y^s_{jk}v_k y^{s\mathsf{H}}_{jk}+2\mathrm{Re}\left(y^s_{jk}V_k(x)I^{\mathsf{H}}_{jk}(x)\right)-2\mathrm{Re}\left(\tilde{y}_{jk}V_j(x)V^{\mathsf{H}}_k(x)y^{s\mathsf{H}}_{j}\right.\end{aligned}$$

where the second equality follows from (17.1b).

Since x satisfies the psd rank-1 conditions (17.1d)(17.1e), there exists (V_j, I_{jk}) that satisfies (17.2) and that is unique up to an arbitrary reference angle for each $(j,k)\in E$.

Clearly $\tilde{x}$ satisfies (17.9a). The construction (**??**) and the psd rank-1 condition (17.1d)(17.1e) imply that $\tilde{x}$ also satisfies (17.9c)(17.9d). Finally to prove (17.9b), recall the derivation above of (17.1b) from (17.9b)(17.9c)(17.9d) using (17.12). The argument in the reverse direction implies that, since x satisfies (17.1b), the (I_{jk}, I_{kj}) obtained from the psd rank-1 decomposition

rank-1 decompositions $\ell_{jk} = I_{jk}I^{\mathsf{H}}_{jk}$ and $\ell_{kj} = I_{kj}I^{\mathsf{H}}_{kj}$ are unique given V_0.

$\tilde{x}$ must satisfies (17.9b) when (I_{jk}, I_{kj}) are given by (17.9b), since the rank-1 decompositions $\ell_{jk} = I_{jk}I^{\mathsf{H}}_{jk}$ and $\ell_{kj} = I_{kj}I^{\mathsf{H}}_{kj}$ are unique given V_0.

To show that the mapping $\tilde{x}\mapsto x$ is surjective, we show that for any $x := (s, v, \ell, S)$ that satisfies (17.1) there is a $\tilde{x}$ that satisfies (17.9). Fix such an $x := (s, v, \ell, S) \in \mathbb{X}_{\text{tree}}$. , when G is a tree, if (v_j, ℓ_{jk}, S_{jk}) satisfy psd and rank-1 conditions (17.1d)(17.1e), then there exist (V_j, I_{jk}, I_{kj}) that satisfy the rank-1 decomposition (**??**). Moreover they are unique since V_0 is fixed. Let $\tilde{x} := (s, V, I, \ell, S)$ where (V, I) are obtained from x uniquely through the rank-1 decomposition. We now show that $\tilde{x}$ satisfies (17.9) and hence the mapping $\tilde{x}\mapsto x$ through (17.11) is surjective.

□

We now show that the cycle condition is vacuous for radial networks, i.e., any x satisfying (17.4) also satisfies (17.17) when the network is radial.

Partition C into its first row c_0^{T} and an $N\times M$ matrix $\hat{C}$ of the remaining rows so that

$$C^{\mathsf{T}} =: \begin{bmatrix} c_0 & \hat{C}^{\mathsf{T}}\end{bmatrix}$$

Similarly partition $\theta =: (\theta_0, \hat{\theta}) \in \mathbb{R}^{3(N+1)}$. Suppose G is a (connected) tree with $M = N$. Then $\hat{C}^{\mathsf{T}}$ is $N\times N$ and of full rank. Therefore $c_0 = \hat{C}^{\mathsf{T}}\eta$ for some $\eta\in\mathbb{C}^N$. It is proved

in Exercise 17.1 that $\left(\hat{C}^{\mathsf{T}}\eta\right)\otimes\mathbb{I} = \left(\hat{C}^{\mathsf{T}}\otimes\mathbb{I}\right)(\eta\otimes\mathbb{I})$. Hence (17.17) becomes

$$\beta(x) \;=\; \left(c_0^{\mathsf{T}}\otimes\mathbb{I}\right)\theta_0 \;+\; \left(\hat{C}^{\mathsf{T}}\otimes\mathbb{I}\right)\hat{\theta} \;=\; \left(\hat{C}^{\mathsf{T}}\otimes\mathbb{I}\right)\left(\hat{\theta}+(\eta\otimes\theta_0)\right) \tag{17.21a}$$

where we have used $(\eta\otimes\mathbb{I})\,\theta_0 = \eta\otimes\theta_0$. Since $\hat{C}^{\mathsf{T}}$ and hence $\left(\hat{C}^{\mathsf{T}}\otimes\mathbb{I}\right)$ are invertible, for any x satisfying (17.4), there always exists an $\theta = \left(\theta_0,\hat{\theta}\right) \in \mathbb{R}^{3(N+1)}$ that satisfies (17.21a). Indeed the solution θ of (17.21a) is not unique. Given any $\theta_0 \in \mathbb{C}^3$, there is always a (unique) $\hat{\theta} := \left(\left(\hat{C}^{\mathsf{T}}\right)^{-1}\otimes\mathbb{I}\right)\beta(x) - \eta\otimes\theta_0$ that satisfies (17.21a).[4]

1in Since x satisfies the cycle condition (17.1f) (17.1g)

$$y^s_{kj}\left(v_j\tilde{y}^{\mathsf{H}}_{jk} - S_{jk}\right) \;=\; \left(v_k\tilde{y}^{\mathsf{H}}_{kj} - S_{kj}\right)^{\mathsf{H}}\left(y^s_{jk}\right)^{\mathsf{H}}, \qquad (j,k)\in E$$

xxx

If G contains cycles, on the other hand, then $M > N$ and the $3M\times 3(N+1)$ matrix $\left(C^{\mathsf{T}}\otimes\mathbb{I}\right)$ in (17.17) has a column rank of $3N < 3M$ since rank $(A\otimes B) =$ rank $A\cdot$ rank B from Lemma 16.6. This means that the column space of $\left(C^{\mathsf{T}}\otimes\mathbb{I}\right)$ does not span $\mathbb{R}^{3M}$ and hence there may be $\beta(x)$ for which no θ exists that satisfies (17.17), regardless of whether θ_0 is given. A power flow model for a meshed network consists of (17.4) augmented with the cycle condition (17.17).

Angle recovery.

As explained in the proof of Theorem 17.1, given V_0, a unique (V,I) can be obtained from a power flow solution x of (17.1) using the psd rank-1 decomposition (17.2). The existence of such a (V,I) guaranteed by Lemma **??**. We next describe an explicit construction of (V,I) from x.

Fix a power flow solution x of (17.1). We will focus on line variables $\left(I_{jk},\ell_{jk},S_{jk}\right)$ in the direction j to k; line variables in the opposite direction have the same properties. Given matrices $\left(v_j,\ell_{jk},S_{jk}\right)$ from the power flow solution x, the vectors $\left(V_j,I_{jk}\right)$ from the psd rank-1 decomposition (17.2) is unique up to a reference angle φ_{jk}, one for each $(j,k)\in E$. When the network graph is a tree, φ_{jk} are the same for all lines $(j,k)\in E$. In this case a given reference angle, e.g., $\angle V_0^a = 0°$, will fix the angles of all variables in x. See Example 17.1 in Chapter 17.3.

If V_0 is given, not just (say) V_0^a, then we can explicitly construct an $\tilde{x} := (s,V,I,\ell,S) \in \tilde{\mathbb{X}}$

the power solution $x := (s,v,\ell,S)$ of (17.1)

that satisfies (17.4), an

[4] Here the vector θ_0 can be arbitrary to satisfy (17.21a) whereas a single angle e.g. θ_0^a fixes all other angles in (17.2). This is because (17.2) uses the matrix v_j whereas (17.21) uses only the diagonal entries of v_j.

can be explicitly constructed using the iterative Algorithm 5 from [104] that makes use of the tree topology. The basic idea in Step 5 of the algorithm is to compute the phasors V_i and I_{ij} recursively, starting from bus 0 when V_0 is given: since $S_{ij} = V_i I_{ij}^{\mathsf{H}}$, taking the Hermitian transpose and multiplying both sides by V_i, we have

$$V_i I_{ij}^{\mathsf{H}} = S_{ij} \;\Rightarrow\; I_{ij}\left(V_i^{\mathsf{H}} V_i\right) = S_{ij}^{\mathsf{H}} V_i \;\Rightarrow\; I_{ij} = \frac{1}{\mathrm{tr}(v_i)} S_{ij}^{\mathsf{H}} V_i \quad (17.22)$$

Algorithm 5: Recover $\tilde{x} = (s, V, I, \ell, S)$ from $x = (s, v, \ell, S)$.
Down orientation where all lines point away from root bus 0.

Input: $x = (s, v, \ell, S) \in \mathbb{X}$; $V_0 \in \mathbb{C}^3$.
Output: $\tilde{x} = (\tilde{s}, \tilde{V}, \tilde{I}, \tilde{\ell}, \tilde{S}) \in \tilde{\mathbb{X}}$

1: $\tilde{s} \leftarrow s$; $\tilde{\ell} \leftarrow \ell$; $\tilde{S} \leftarrow S$;
2: $N_{\text{visit}} \leftarrow \{0\}$;
3: **while** $N_{\text{visit}} \neq \overline{N}$ **do**
4: find $i \rightarrow j$ such that $i \in N_{\text{visit}}$ and $j \notin N_{\text{visit}}$;
5: compute

$$\begin{aligned} \tilde{I}_{ij} &\leftarrow \frac{1}{\mathrm{tr}\,(v_i)} S_{ij}^{\mathsf{H}} \tilde{V}_i \\ \tilde{V}_j &\leftarrow \tilde{V}_i - z_{ij} \tilde{I}_{ij} \\ N_{\text{visit}} &\leftarrow N_{\text{visit}} \cup \{j\} \end{aligned}$$

6: **end while**

17.3 Overall model and examples

17.3.1 Overall model

Suppose assumption C17.1 holds. The overall model of a network of three-phase devices connected by three-phase lines, its specification and analysis are similar to that in the bus injection model discuss in Chapter 16.2. The only difference is that the power flow equations are those for BFM rather than BIM. Specifically the overall model consists of:

1 A network model that relates terminal voltage, current, and power (V, I, s). Any equivalent model can be used, whichever is convenient for the problem under study, including:
- the BFM (17.4) for radial networks; or

- the BFM (17.10) for general networks.

2 A device model for each three-phase device j. For ideal devices, this can either be:
   - Its internal model (14.29) and the conversion rules (14.8) and (14.9)(14.10); or
   - Its external model summarized in Tables 14.3 and 14.4 when only terminal quantities are needed.

   For non-ideal devices, this can either be:
   - Its internal model summarized in Table 14.2 and the conversion rules (14.8) and (14.9)(14.10); or
   - Its external model summarized in Table 14.2 when only terminal quantities are needed.

Unlike the models of Chapter 16.1.5 where, if only voltage sources, current sources and impedances are involved, then the overall model is linear, consisting of the nodal current balance equation (16.5)(16.6) and linear device models. Here the BFM equations (17.4) and (17.10) are quadratic, leading to a nonlinear overall model even if power sources are absent.

A typical three-phase analysis problem can be specified and analyzed the same way as described in Chapter 16.2 for BIM. A solution typically takes the following steps:

1 Write down the models of the given collection of three-phase devices, either their internal models and conversion rules or their external models (if internal variables are not required).
2 Write down a network equation that relates the terminal variables, either the current balance equation or a power flow equation.
3 Steps 1 and 2 specify a system of nonlinear equations that relate relevant external and internal variables as well as given parameters. It generally needs to be solved numerically. We will describe in Chapter 17.4 such an algorithm for radial networks, the three-phase backward-forward sweep (BFS).
4 Usually we first compute the terminal variables (V_j, I_j, s_j) using network equations, together with some of (γ_j, β_j), and then determine the internal variables $\left(V_j^{Y/\Delta}, I_j^{Y/\Delta}, s_j^{Y/\Delta}\right)$ using the conversion rules.

17.3.2 Examples

We now illustrate with examples three-phase BFMs and the analysis procedure. Suppose assumption C17.1 holds.

Example 17.1 (Power source in Y configuration)**.** Consider the system in Figure 17.1 where a constant-power source $\sigma_j^Y \in \mathbb{C}^3$ is connected through a three-phase line to an impedance load z_k^Y, both in Y configuration. For simplicity we assume that both

neutrals are directly grounded and all voltages are defined with respect to the ground, so that the neutral voltages $\gamma_j := V_j^n = \gamma_k := V_k^n = 0$. Suppose the following are given:

- The constant-power source $\sigma_j^Y := \left(\sigma_j^{an}, \sigma_j^{bn}, \sigma_j^{cn}\right)$ with $\angle V_0^a := 0°$.
- The impedance load $z_k^Y := \text{diag}\left(z_k^{an}, z_k^{bn}, z_k^{cn}\right)$.
- The series impedance matrix $z_{jk} \in \mathbb{C}^{3\times 3}$ of the line. Its shunt admittance matrices are assumed zero.

Derive the $\left(s_k^Y, v_k, \ell_{jk}, S_{jk}\right)$ in terms of the given parameters.

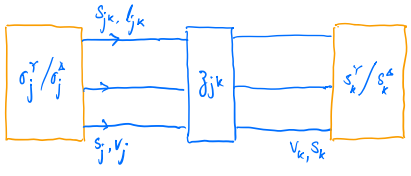


Figure 17.1 Example 17.1.

Solution. The system is specified by:

1 *Netowrk model*: The power flow equation (17.4) that relates terminal variables, specialized to the two-bus system in Figure 17.1, is:

$$\text{diag}(S_{jk}) = s_j, \qquad \text{diag}\left(S_{jk} - z_{jk}\ell_{jk}\right) = -s_k \tag{17.23a}$$

$$v_j - v_k = \left(z_{jk} S_{jk}^{\mathsf{H}} + S_{jk} z_{jk}^{\mathsf{H}}\right) - z_{jk}\,\ell_{jk}\, z_{jk}^{\mathsf{H}} \tag{17.23b}$$

$$\begin{bmatrix} v_j & S_{jk} \\ S_{jk}^{\mathsf{H}} & \ell_{jk} \end{bmatrix} \geq 0, \qquad \text{rank} \begin{bmatrix} v_j & S_{jk} \\ S_{jk}^{\mathsf{H}} & \ell_{jk} \end{bmatrix} = 1 \tag{17.23c}$$

2 *Device model*: The internal model of Y-configured impedance is (since $V_k = V_k^Y = z_k^Y I_k^Y$ and $I_k^Y = I_{jk}$):

$$v_k = z_k^Y \ell_{jk}\, z_k^{Y\mathsf{H}}, \qquad s_k^Y = \text{diag}\left(z_k^Y \ell_{jk}\right) \tag{17.24a}$$

and the conversion rule (14.8) between internal and terminal variables is:

$$s_j = -\left(\sigma_j^Y + V_j^n \bar{I}_{jk}\right) = -\sigma_j^Y, \quad s_k = -\left(s_k^Y + V_k^n(-\bar{I}_{jk})\right) = -s_k^Y \tag{17.24b}$$

The system of quadratic equations (17.23)(17.24) cannot generally be solved in closed form, but can be solved numerically for $\left(s_k^Y, v_k, \ell_{jk}, S_{jk}\right)$ (see Chapter 17.4).

To better appreciate the structure of the three-phase model we now reduce

(17.23)(17.24) to three quadratic equations in three unknowns $I_{jk} \in \mathbb{C}^3$. Relate ℓ_{jk} to σ_j^Y by eliminating the terminal powers (s_j, s_k), line power S_{jk} and internal power s_k^Y from (17.23a) (17.24):

$$-\sigma_j^Y \;=\; \operatorname{diag}\left(\left(z_k^Y + z_{jk}\right)\ell_{jk}\right) \tag{17.25}$$

This is a system of three complex quadratic equations in three unknown line currents $I_{jk} := \left(I_{jk}^a, I_{jk}^b, I_{jk}^c\right)$ because (17.23c) means that ℓ_{jk} has a rank-1 decomposition $\ell_{jk} = I_{jk} I_{jk}^{\mathsf{H}}$ (from (17.2)). Let $Z_k^Y := z_k^Y + z_{jk}$. Then (17.25) is explicitly:

$$-\sigma_j^Y \;=\; \operatorname{diag}\left(\begin{bmatrix} Z_k^{aa} & Z_k^{ab} & Z_k^{ac} \\ Z_k^{ba} & Z_k^{bb} & Z_k^{bc} \\ Z_k^{ca} & Z_k^{cb} & Z_k^{cc} \end{bmatrix} \begin{bmatrix} I_{jk}^a \\ I_{jk}^b \\ I_{jk}^c \end{bmatrix} \begin{bmatrix} I_{jk}^{a\mathsf{H}} & I_{jk}^{b\mathsf{H}} & I_{jk}^{c\mathsf{H}} \end{bmatrix}\right)$$

or

$$\begin{aligned} -\sigma_j^{an} &= Z_k^{aa} I_{jk}^a I_{jk}^{a\mathsf{H}} + Z_k^{ab} I_{jk}^b I_{jk}^{a\mathsf{H}} + Z_k^{ac} I_{jk}^c I_{jk}^{a\mathsf{H}} \\ -\sigma_j^{bn} &= Z_k^{ba} I_{jk}^a I_{jk}^{b\mathsf{H}} + Z_k^{bb} I_{jk}^b I_{jk}^{b\mathsf{H}} + Z_k^{bc} I_{jk}^c I_{jk}^{b\mathsf{H}} \\ -\sigma_j^{cn} &= Z_k^{ca} I_{jk}^a I_{jk}^{c\mathsf{H}} + Z_k^{cb} I_{jk}^b I_{jk}^{c\mathsf{H}} + Z_k^{cc} I_{jk}^c I_{jk}^{c\mathsf{H}} \end{aligned}$$

There is a power flow solution for (17.23)(17.24) if and only if (17.25) has a solution for I_{jk}, up to an angle to be determined (from the given $\angle V_0^a = 0°$).

Once I_{jk} and hence ℓ_{jk} are determined from (17.25), all other variables can be obtained. Specifically since $V_k = V_k^Y + V_k^n = V_k^Y$ by assumption, the load voltage and power are given by (17.24a):

$$v_k \;=\; v_k^Y \;=\; z_k^Y \ell_{jk} z_k^{Y\mathsf{H}} \;=\; \left(z_k^Y I_{jk}\right)\left(z_k^Y I_{jk}\right)^{\mathsf{H}}, \qquad s_k^Y \;=\; \operatorname{diag}\left(z_k^Y \ell_{jk}\right)$$

Since v_k has a rank-1 decomposition due to (17.23c), $V_k := \left(V_k^a, V_k^b, V_k^c\right)$ can be obtained from the first equation as $V_k = z_k^Y I_{jk}$, up to an angle to be determined. Finally we obtain V_j from $-\sigma_j^Y = s_j = \operatorname{diag}\left(V_j I_{jk}^{\mathsf{H}}\right)$ due to (17.24b) and then $S_{jk} = V_j I_{jk}^{\mathsf{H}}$. The given $\angle V_j^a = 0°$ then fixes the angles of $\left(V_j, V_k, I_{jk}\right)$. □

The next example illustrates two solution approaches for constant-power source in Δ configuration. Both relate the terminal variables of each device to its parameters and then relates these terminal variables by the power flow equation. The first approach boils down to computing the internal current I_j^Δ from a system of quadratic equations, which then yields $\left(I_j, \beta_j\right)$ and all other variables. The second approach boils down to computing the terminal current and its zero-sequence component $\left(I_j, \beta_j\right)$ and then other variables.

As for Example 16.8, only γ_j of the source needs to be given. All other variables including $\left(\beta_j, \gamma_k, \beta_k\right)$ can then be determined. The solution method of these two examples is similar because the overall models in these examples differ only in their power flow equations, BIM (16.12) versus BFM (17.10). The positive definite and

rank-1 condition in (17.23c) leads to the equivalence of BFM (17.10) to (17.23) and BIM (16.12) (Theorems **??** and 17.1).

Example 17.2 (Power source in Δ configuration)**.** Consider a three-phase power source and an impedance, both in Δ configuration, connected by a three-phase line (as in Example 16.8) with the following given parameters:

- The constant-power source $\left(\sigma_j^\Delta, \gamma_j\right)$ with $\angle V_j^{ab} := 0°$.
- The impedance load z_k^Δ. (Note that β_k need not be specified for an impedance and can be derived.)
- The series impedance matrix z_{jk} of the line. Its shunt admittance matrices are assumed zero.

Solve for the remaining variables.

Solution 1: compute I_j^Δ. The system is specified by:

1 *Netowrk model*: The power flow equation that relates terminal variables remains (17.23).
2 *Device model for power source σ_j^Δ*: At bus j we use the model (14.25b) and the conversion rule that relates the terminal variables $\left(V_j, I_j, s_j\right)$ to internal power σ_j^Δ and internal current I_j^Δ:

$$s_j := \operatorname{diag}\left(V_j I_j^{\mathsf{H}}\right) \tag{17.26a}$$

$$\sigma_j^\Delta := \operatorname{diag}\left(V_j^\Delta I_j^{\Delta\mathsf{H}}\right) = \operatorname{diag}\left(\Gamma V_j I_j^{\Delta\mathsf{H}}\right), \qquad I_j = -\Gamma^{\mathsf{T}} I_j^\Delta \tag{17.26b}$$

3 *Device model for impedance z_k^Δ*: At bus k the external model in Table 14.4 relates the terminal variables (V_k, I_k, s_k) to impedance z_k^Δ through the admittance matrix Z_k^Δ defined in (14.27b):[5]

$$s_k := \operatorname{diag}\left(V_k I_k^{\mathsf{H}}\right), \qquad V_k = -Z^\Delta I_k + \gamma_k \mathbf{1}, \qquad \mathbf{1}^{\mathsf{T}} I_k = 0 \tag{17.26c}$$

The device models (17.26) relate terminal variables $\left(V_j, I_j, s_j\right)$ and (V_k, I_k, s_k) to the internal parameters $\left(\sigma_j^\Delta, z_k^\Delta\right)$ of the devices through γ_k (which is to be determined). The power flow equation (17.23) relates these terminal variables.

The rank-1 condition (17.23c) (as well as KCL) connects these terminal variables

[5] Using the equivalent impedance model in terms of the impedance matrix Y_k^Δ defined in (14.27a) here does not . in which case (17.26c) is replaced by:

$$s_k := \operatorname{diag}\left(V_k I_k^{\mathsf{H}}\right), \qquad I_k = -Y^\Delta V_k$$

and the variables $(v_j, v_k, \ell_{jk}, S_{jk})$ of (17.23):

$$I_j = I_{jk} = -I_k, \qquad S_{jk} = V_j I_{jk}^{\mathsf{H}} \tag{17.27a}$$

$$\ell_{jk} = I_{jk} I_{jk}^{\mathsf{H}}, \qquad v_j = V_j V_j^{\mathsf{H}}, \qquad v_k = V_k V_k^{\mathsf{H}} \tag{17.27b}$$

The equations (17.23)(17.26)(17.27) are a system of quadratic equations in variables $\left(V_j, I_j, s_j, I_j^\Delta\right)$, $(V_k, I_k, s_k, \gamma_k)$, and $\left(I_{jk}, v_j, v_k, \ell_{jk}, S_{jk}\right)$. They can be solved numerically. Once these terminal variables are determined, the internal variables $\left(\beta_j, V_k^\Delta, I_k^\Delta, s_k^\Delta, \beta_k\right)$ can be determined. In particular once V_k is determined from the network equations we can obtain $V_k^\Delta = \Gamma V_k$ and then $I_k^\Delta = z_{jk}^{-1} V_k^\Delta$ and hence β_k.

To better appreciate the structure of this model we now reduce (17.23)(17.26)(17.27) to 3 quadratic equations in 3 variables I_{jk}^Δ for each link $j \to k \in E$. Theorem **??** implies the equivalence of BFMs (17.23) and (17.10). In particular (from (17.10b))

$$V_j - V_k = z_{jk} I_{jk}$$

which can also be derived by substituting (17.27) into (17.23b). Substitute V_k from (17.26c) and $I_k = -I_{jk}$ into this equation to eliminate V_k:

$$V_j = \hat{Z}_k^\Delta I_{jk} + \gamma_k \mathbf{1}, \qquad \mathbf{1}^{\mathsf{T}} I_{jk} = 0 \tag{17.28}$$

where $\hat{Z}_k^\Delta := Z_k^\Delta + z_{jk}$ is the equivalent of the line impedance in series with the load impedance. Substituting $I_{jk} = I_j = -\Gamma^{\mathsf{T}} I_j^\Delta$ into (17.28) and substituting the resulting V_j into (17.26b), we obtain a quadratic equation in I_j^Δ (using $\Gamma \mathbf{1} = 0$):

$$\sigma_j^\Delta := -\mathrm{diag}\left(\left(\Gamma \hat{Z}_k^\Delta \Gamma^{\mathsf{T}}\right) I_j^\Delta I_j^{\Delta\mathsf{H}}\right), \qquad j \in \overline{N} \tag{17.29}$$

There is a power flow solution to (17.23)(17.26)(17.27) if and only if (17.29) has a solution for I_j^Δ. Once I_j^Δ is determined it yields $I_{jk} = I_j = -\Gamma^{\mathsf{T}} I_j^\Delta$ and $\beta_j := \frac{1}{3}\mathbf{1}^{\mathsf{T}} I_j^\Delta$. Since (17.29) is the same equation as (16.24) in Example 16.8, we can follow the same procedure there to derive all variables (V_j, I_j, s_j, β_j) and $(V_k, I_k, s_k, \gamma_k)$. Then we can obtain internal variables $\left(V_k^\Delta, I_k^\Delta, s_k^\Delta, \beta_k\right)$ and the BFM variables $\left(I_{jk}, v_j, v_k, \ell_{jk}, S_{jk}\right)$ from (17.27). In particular, V_k yields V_k^Δ and hence I_k^Δ and β_k. (To get more insight on its solution, see the solution of the balanced case in Exercise 16.10.)

Solution 2: compute I_j. Instead of the power source model (17.26b), we can also use the external model in Table 14.4 to relate the terminal current I_j direclty to the internal power σ_j^Δ:

$$\sigma_j^\Delta := \mathrm{diag}\left(V_j^\Delta I_j^{\Delta\mathsf{H}}\right) = -\mathrm{diag}\left(\Gamma\left(V_j I_j^{\mathsf{H}}\right)\Gamma^\dagger\right) + \overline{\beta}_j \Gamma V_j, \quad \mathbf{1}^{\mathsf{T}} I_j = 0 \tag{17.30}$$

where the internal variable β_j is to be determined. Substituting (17.28) into (17.30) and noting $I_j = I_{jk}$ we have

$$\sigma_j^\Delta = -\frac{1}{3}\mathrm{diag}\left(\Gamma \hat{Z}_k^\Delta I_{jk} I_{jk}^{\mathsf{H}} \Gamma^{\mathsf{T}}\right) + \overline{\beta}_j \Gamma \hat{Z}_k^\Delta I_{jk}, \qquad \mathbf{1}^{\mathsf{T}} I_{jk} = 0 \tag{17.31}$$

There is a power flow solution to (17.23)(17.26)(17.27) if and only if there is a solution

$I_{jk} := I_{jk}(\sigma_j^\Delta)$ and $\beta_j := \beta_j(\sigma_j^\Delta)$ to (17.31). Given a solution (I_{jk}, β_j) and hence I_{jk}^Δ, all other variables can be derived as in Solution 1. □

Remark 17.1. Even though the analysis in Example 17.2 makes heavy use of BFM (17.23) with phasor variables such as (V_j, I_{jk}) instead of variables of BFM (17.10) such as (v_j, ℓ_{jk}, S_{jk}), the model (17.10) is useful for solving optimal power flow problems through semidefinite relaxation; see Chapter . □

17.4 Backward forward sweep

In this section we extend the backward forward sweep (BFS) of Chapter 5.3 for the computation of power flow solutions from single-phase radial networks to three-phase radial networks. As explained in Chapter 5.3.1 BFS can be interpreted as a Gauss-Siedel algorithm that computes a fixed point of BFM equations. It has two special structures that exploit the tree topology of the network. First it partitions the power flow variable into two vectors x and y and updates them iteratively in an outer loop. Typically x consists of branch variables, e.g., branch currents or powers, and y consists of nodal variables, e.g., nodal voltages. Second, for each outer iteration, it computes iteratively each component of (x, y) in an inner loop that makes use of a spatially recursive structure enabled by the tree topology. Specifically it computes the components of x iteratively from leaf nodes towards the root of the tree (backward sweep) and then computes the components of y iteratively from the root towards the leaf nodes (forward sweep). The design of BFS involves the choice of power flow equations and variables (x, y) based on what information is given in a power flow problem. These choices are not unique and may have different convergence properties. The general algorithmic structure described in Chapter 5.3.1 applies to three-phase as well as single-phase radial networks. We have presented two BFS algorithms in Chapters 5.3.2 and 5.3.3 that use different branch flow models. In this section we describe an algorithm that extends both single-phase algorithms to the three-phase setting. As we will see, the main addition is the computation of internal variables associated with each three-phase device.

Recall that we assume C17.1 holds throughout this chapter.

17.4.1 Complex form BFM

Consider a radial network modeled as a directed graph G, rooted at bus 0 and with each line pointing *away* from the root bus 0. Each line is characterized by 3×3 admittance matrices $\left(y_{jk}^s, y_{jk}^m, y_{kj}^m\right)$. Suppose there is exactly one three-phase power source at each bus j either in Y or Δ configuration. At every non-root bus $j \in N$, the internal power $\sigma_j^{Y/\Delta} \in \mathbb{C}^3$ of the power source is given and its terminal voltage and current (V_j, I_j)

are to be determined.[6] At bus 0, $V_0 \in \mathbb{C}^3$ is given and the current injection I_0 and the internal power injection $s_0^{Y/\Delta}$ are to be determined. We assume for simplicity that C14.1 with $z_j^n = 0$ holds at every bus $j \in \overline{N}$ that has a Y-configured power source so that $V_j^n = 0$ (see Remark 17.3 on the case when $z_j^n \neq 0$ so that $V_j^n = -z_j^n \left(\mathbf{1}^\mathsf{T} I_j\right)$).

As for the single-phase BFS, let $\left(I_{jk}^s,\ j \to k \in E\right)$ be the branch current through the series admittance matrix $y_{jk}^s \in \mathbb{C}^{3\times 3}$ (see Exercise 17.3 for a BFS algorithm that computes the sending-end current I_{jk} instead). The receiving current at bus j from its parent i is $\left(I_{ij}^s - y_{ji}^m V_j\right) \in \mathbb{C}^3$ (see Figure 17.2). The current balance equation is then

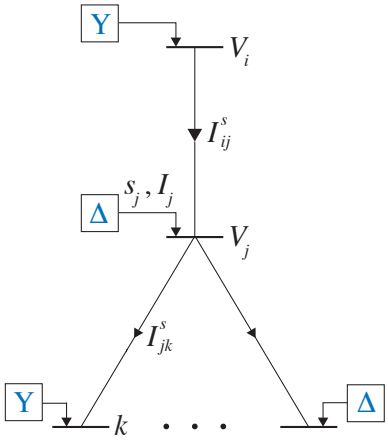


Figure 17.2 Notation for BFS on unbalanced three-phase radial networks.

$$I_j + \left(I_{ij}^s - y_{ji}^m V_j\right) = \sum_{k:j\to k} \left(I_{jk}^s + y_{jk}^m V_j\right)$$

Rewriting this in a form suitable for backward sweep, we obtain the following three-phase branch flow model in terms of branch variables $\left(I_{jk}^s, j \to k \in E\right)$ and nodal variables $\left(V_j, I_j,\ j \in \overline{N}\right)$:

$$I_{ij}^s = \sum_{k:j\to k} I_{jk}^s - \left(I_j - y_{jj}^m V_j\right), \qquad j \in N \tag{17.32a}$$

$$V_j = V_i - z_{ij}^s I_{ij}^s, \qquad j \in N \tag{17.32b}$$

where $y_{jj}^m := y_{ji}^m + \sum_{k:j\to k} y_{jk}^m$ are the total shunt admittances incident on j and $z_{ij}^s := \left(y_{ij}^s\right)^{-1}$ are the series impedances. These network equations relate the branch currents I_{jk}^s as well as the terminal voltages and currents $\left(V_j, I_j\right)$ at buses across the network.

6

Each terminal variable (V_j, I_j) is related to the internal power $\sigma_j^{Y/\Delta}$ through a three-phase device model. We adopt the following device models for reasons discussed in Remark 17.2 (from (14.17b) and (14.25b) and recall that $V_j^n = 0$ by assumption):

$$Y \text{ configuration:} \qquad \sigma_j^Y = \operatorname{diag}\left(V_j I_j^{Y\mathsf{H}}\right), \quad I_j = -I_j^Y \tag{17.33a}$$

$$\Delta \text{ configuration:} \qquad \sigma_j^\Delta = \operatorname{diag}\left(\Gamma V_j I_j^{\Delta\mathsf{H}}\right), \quad I_j = -\Gamma^{\mathsf{T}} I_j^\Delta \tag{17.33b}$$

Hence, for a non-root bus j, the given internal power $\sigma_j^{Y/\Delta}$ determines, through its internal current $I_j^{Y/\Delta}$, its terminal voltage and current (V_j, I_j) according to (17.33). These terminal variables interact across the network according to the network equations (17.32). Given V_j, the forward sweep function g_j in (**??**) to update $\left(I_j^{Y/\Delta}, I_j\right)$ is:

$$Y: \qquad I_j^Y = \left(\operatorname{diag} \bar{V}_j\right)^{-1} \bar{\sigma}_j^Y, \qquad I_j = -I_j^Y, \qquad j \in N \tag{17.34a}$$

$$\Delta: \qquad I_j^\Delta = \left(\operatorname{diag}\left(\Gamma \bar{V}_j\right)\right)^{-1} \bar{\sigma}_j^\Delta, \qquad I_j = -\Gamma^{\mathsf{T}} I_j^\Delta, \qquad j \in N \tag{17.34b}$$

where $\bar{v}$ denotes the componentwise complex conjugate of a vector v. Here, we have used, for vectors $v, w \in \mathbb{C}^n$, $\operatorname{diag}(v w^{\mathsf{H}}) = \operatorname{diag}(v)\bar{w} = \operatorname{diag}(\bar{w})v \in \mathbb{C}^n$ where $\operatorname{diag}(v)$ is the diagonal matrix whose diagonal is the vector v.

To construct the backward forward sweep, identify lines $j \to k \in E$ by the non-root buses $k \in N$. Given V_0 and $\sigma := \left(\sigma_j^{Y/\Delta}, j \in N\right)$, the BFS will compute the following branch and nodal variables respectively:

$$x := \left(I_{ij}^s,\ j \in N\right), \qquad y := \left(V_j, I_j, I_j^{Y/\Delta},\ j \in N\right)$$

All other variables, such as injections $I_0, s_0, s_0^{Y/\Delta} \in \mathbb{C}^3$, branch flow matrices $S_{jk} \in \mathbb{C}^{3\times 3}$, and $(\gamma_j, \beta_j) \in \mathbb{C}^2$ of power sources σ_j^Δ, can be computed once (x, y) are determined. The update function f in the backward sweep to update x is defined by (17.32a) and the update function g in the forward sweep to update y is defined by (17.32b) and (17.34). The function f is jointly linear in (x, y). The function g is linear in x but nonlinear in y because of the power source model (17.34).

The boundary conditions are

$$V_0 \in \mathbb{C}^3 \text{ is given}, \qquad I_{jk}^s := 0 \text{ for all leaf nodes } j, \qquad V_j(0) := V_0,\ j \in N \tag{17.35a}$$

In addition, given the initial voltages $\left(V_j(0), j \in N\right)$, the terminal and internal currents $\left(I_j(0), I_j^{Y/\Delta(0)}, j \in N\right)$ are determined using (17.34):

$$Y: \qquad I_j^Y(0) = \left(\operatorname{diag} \bar{V}_j(0)\right)^{-1} \bar{\sigma}_j^Y, \qquad I_j(0) = -I_j^Y(0), \qquad j \in N \tag{17.35b}$$

$$\Delta: \qquad I_j^\Delta(0) = \left(\operatorname{diag}\left(\Gamma \bar{V}_j(0)\right)\right)^{-1} \bar{\sigma}_j^\Delta, \qquad I_j(0) = -\Gamma^{\mathsf{T}} I_j^\Delta(0), \qquad j \in N \tag{17.35c}$$

Specifically the BFS algorithm defined by (17.32) (17.34) (17.35) proceeds as follows.

0. *Input*: voltage V_0 pu and internal power $\left(\sigma_j^{Y/\Delta}, j \in N\right)$.
1. *Initialization.*
   - $I^s_{jk}(t) := 0$ for all leaf nodes j for all iterations $t = 1, 2, \ldots$.
   - $V_0(t) := V_0$ for all $t = 0, 1, \ldots$.
   - $V_j(0) := V_0$ at all buses $j \in N$. Compute $\left(I_j(0), I_j^{Y/\Delta}(0)\right)$ using (17.35b)(17.35c).
2. *Backward forward sweep.* Iterate for $t = 1, 2, \ldots$ until a stopping criterion (see below) is satisfied:
   1. *Backward sweep.* Starting from the leaf nodes and iterating towards bus 0, compute

$$I^s_{ij}(t) \leftarrow \sum_{k: j \to k} I^s_{jk}(t) \; - \; \left(I_j(t-1) - y^m_{jj} V_j(t-1)\right), \quad i \to j \in E \tag{17.36a}$$

   where $y^m_{jj} := y^m_{ji} + \sum_{k: j \sim k} y^m_{jk}$.
   2. *Forward sweep.* Starting from bus 0 and iterating towards the leaf nodes, compute for $j \in N$

$$V_j(t) \; \leftarrow \; V_i(t) \; - \; z^s_{ij} I^s_{ij}(t) \tag{17.36b}$$

$$Y: \quad I^Y_j(t) \; \leftarrow \; \left(\operatorname{diag} \bar{V}_j(t)\right)^{-1} \bar{\sigma}^Y_j, \qquad I_j(t) \; \leftarrow \; -I^Y_j(t) \tag{17.36c}$$

$$\Delta: \quad I^\Delta_j(t) \; \leftarrow \; \left(\operatorname{diag}\left(\Gamma \bar{V}_j(t)\right)\right)^{-1} \bar{\sigma}^\Delta_j, \quad I_j(t) \; \leftarrow \; -\Gamma^{\mathsf{T}} I^\Delta_j(t) \tag{17.36d}$$

   where $z^s_{ij} := \left(y^s_{ij}\right)^{-1}$.
3. *Output*: branch variable $x := \left(I^s_{ij}(t),\ j \in N\right)$ and nodal variable $y := \left(V_j(t), I_j(t), I_j^{Y/\Delta(t)},\ j \in N\right)$.

A stopping criterion can be based on the discrepancy between the given internal powers $\sigma_j^{Y/\Delta}$ and those implied by the nodal variable $\left(V_j(t), I_j(t), I_j^{Y/\Delta(t)},\ j \in N\right)$ in each iteration t. From the device model (17.34), let

$$\hat{\sigma}_j(t) \; := \; \begin{cases} \operatorname{diag}\left(V_j(t) I_j^{Y\mathsf{H}}(t)\right) & \text{for} \quad Y \text{ configuration} \\ \operatorname{diag}\left(\Gamma V_j(t) I_j^{\Delta\mathsf{H}}(t)\right) & \text{for} \quad \Delta \text{ configuration} \end{cases}$$

Then a stopping criterion can be

$$\|\hat{\sigma}(t) - \sigma^{Y/\Delta}\|_2^2 \; := \; \sum_{j \in N} \left(\hat{\sigma}_j(t) - \sigma_j^{Y/\Delta}\right)^2 < \epsilon$$

for a given tolerance $\epsilon > 0$.

Remark 17.2 (Choice of variables)**.** 1 We have used the current balance equation (17.32a) to relate terminal voltages and currents (V_j, I_j) across the network. This leads to a linear update function (17.32a) for x in backward sweep.

Nonlinearity shows up in the device model (17.34) for the nodal variable $y := \left(V_j, I_j, I_j^{Y/\Delta},\ j \in N\right)$ in the forward sweep (together with (17.32b)).

2 A direct extension of the single-phase BFS in [30] to the three-phase setting is the approach in [45] which substitutes I_j in (17.32a) by $I_j = \left(\text{diag}\bar{V}_j\right)^{-1} \bar{s}_j$ to obtain a nonlinear update function for x:

$$I_{ij}^s \;=\; \sum_{k:j\to k} I_{jk}^s \;-\; \left(\left(\text{diag}\;\bar{V}_j\right)^{-1} \bar{s}_j - y_{jj}^m V_j\right), \qquad j \in N \tag{17.37a}$$

In this case the nodal variable becomes $y := \left(V_j, s_j, I_j^{Y/\Delta},\ j \in N\right)$ and the update functions (17.34) become

$$Y: \quad I_j^Y \;=\; \left(\text{diag}\;\bar{V}_j\right)^{-1} \bar{\sigma}_j^Y, \qquad s_j \;=\; -\sigma_j^Y, \qquad j \in N \tag{17.37b}$$

$$\Delta: \quad I_j^\Delta \;=\; \left(\text{diag}\left(\Gamma\bar{V}_j\right)\right)^{-1} \bar{\sigma}_j^\Delta, \quad s_j \;=\; -\text{diag}\left(V_j I_j^{\Delta\mathsf{H}}\Gamma\right), \quad j \in N \tag{17.37c}$$

The three-phase BFS of [45] includes only Y-configured power sources and therefore its update functions simplifies to only (17.37a) (17.32b), with $s_j = -\sigma_j^Y$ that is fixed and given. The addition of Δ-configured power sources requires the nodal variable I_j^Δ and update function (17.37c).

3 For Δ configuration, the device model (17.34) relates σ^Δ to (V_j, I_j) through I_j^Δ. Since $\left(V_j, I_j^\Delta\right)$ are determined directly from the overall model, the quantities (γ_j, β_j) can be computed and need not be specified. Note however that V_0 is given.

□

Remark 17.3 (Nonzero z_j^n)**.** If we had assumed C14.1 with $z_j^n \neq 0$ so that $V_j^n = -z_j^n\left(\mathbf{1}^\mathsf{T} I\right)$, then the device model (17.34a) for a Y-configured power source becomes nonlinear in I_j (from (14.17b)):

$$Y: \qquad V_j \;=\; -\left(\text{diag}\left(\bar{I}_j\right)\right)^{-1} \sigma_j^Y \;-\; z_j^n\left(\mathbf{1}\mathbf{1}^\mathsf{T}\right) I_j, \qquad j \in N$$

Given voltage V_j this is a system of three quadratic equations in three unknowns $I_j \in \mathbb{C}^3$:

$$z_j^n\left(\mathbf{1}^\mathsf{T} I_j\right)\bar{I}_j \;+\; \text{diag}\left(V_j\right)\bar{I}_j \;+\; \sigma_j^Y \;=\; 0$$

The linear update functions (17.34a) (17.35b) then become nonlinear. Moreover the update of I_j is defined only implicitly by a solution of this system of quadratic equations.

□

Remark 17.4 (Specification)**.** Unlike in Examples 17.1 and 17.2, the BFS method here does not required γ_j be specified, but it requires that V_0 be specified. □

17.4.2 DistFlow model

Consider a three-phase radial network modeled by a directed graph with every link $k \to j \in E$ points *away from* the root bus 0. Assume for simplicity zero shunt admittances, $y^m_{jk} = y^m_{kj} = 0$. The three-phase DistFlow equations for the down orientation are (17.4). Given V_0, hence $v_0 := V_0 V_0^{\mathsf{H}}$, and internal power $\sigma := \left(\sigma_j^{Y/\Delta}, j \in N\right)$, we wish to compute the other variables from (17.4).

The nonlinear equation $v_j \ell_{jk} = |S_{jk}|^2$ in (**??**) for the single-phase model is replaced by (17.4c)(**??**) in the three-phase model, reproduced here

$$\begin{bmatrix} v_j & S_{jk} \\ S_{jk}^{\mathsf{H}} & \ell_{jk} \end{bmatrix} \geq 0, \qquad \operatorname{rank} \begin{bmatrix} v_j & S_{jk} \\ S_{jk}^{\mathsf{H}} & \ell_{jk} \end{bmatrix} = 1$$

These equations are an implicit description and do not directly yield an update equation for a BFS algorithm, as $v_j \ell_{jk} = |S_{jk}|^2$ does in the single-phase model. Instead, they imply that there exist voltage and current phasors $(V, \tilde{I})$ that satisfy the rank-1 decomposition in (17.2). In order to compute DistFlow variables (v, ℓ, S) we have to compute iteratively the voltages V_j and (sending-end) line currents $\tilde{I}_{jk}$ in the process. Here we use $\tilde{I}_{jk}$ to denote a line current to differentiate it from the terminal current I_j in a device model (see below). Therefore, instead of designing an BFS algorithm based on (17.4), we will use the following network equations derived from (17.4) to compute $(V, I, \tilde{I})$:

$$\tilde{I}_{ij} = -I_j + \sum_{k: j \to k} \tilde{I}_{jk} \tag{17.38a}$$

$$V_k = V_j - z_{jk} \tilde{I}_{jk} \tag{17.38b}$$

All other terminal variables such as $v_j = V_j V_j^{\mathsf{H}}$, $\ell_{jk} = \tilde{I}_{jk} \tilde{I}_{jk}^{\mathsf{H}}$, and $S_{ij} = V_i \tilde{I}_{ij}^{\mathsf{H}}$, can then be derived. Note that we have replaced the power balance equation (17.4a) by the current balance equation in (17.38a). The network equation (17.38) is the same as (17.32) with $\tilde{I}_{jk} = I^s_{jk}$ when $y^m_{jk} = y^m_{kj} = 0$. Hence the three-phase DistFlow model can be solved using the BFS algorithm of Chapter 17.4.1.

17.5 Linear model

17.5.1 Three-phase LinDistFlow

Model.

We generalize the linear DistFlow model from single-phase to unbalanced multiphase radial networks. The key assumptions in our linear approximation are:

1 The real and reactive line losses $z_{jk}\ell_{jk}$ are much smaller than line flows S_{jk} on each line $j \to k$, so that we can assume $\ell_{jk} = 0$ in (17.4).
2 The voltages are approximately balanced, so that we can assume

$$\frac{V_j^a}{V_j^b} \;=\; \frac{V_j^b}{V_j^c} \;=\; \frac{V_j^c}{V_j^a} \;=\; e^{\mathrm{i}2\pi/3}$$

Recall that we adopt, without loss of generality, the graph orientation in which all lines point away from bus 0. Then, as for the single-phase model, we set $\ell_{jk} = 0$ in (17.4a)(17.4b) to obtain

$$\sum_{k:j\to k} \operatorname{diag}(S_{jk}) = \operatorname{diag}\left(S_{ij}\right) + s_j, \qquad j \in \overline{N}$$

$$v_j - v_k = z_{jk}\, S_{jk}^{\mathsf{H}} \;+\; S_{jk}\, z_{jk}^{\mathsf{H}} \qquad j \to k \in E$$

where bus $i := i(j)$ is the unique parent of bus j. Given injections s_j for all non-slack buses $j \in N$, the first set of equations determines uniquely s_0 and the diagonal entries of S_{jk}, but not the off-diagonal entries of S_{jk}. The second assumption of balanced voltage is needed to determine the off-diagonal entries of S_{jk}. Specifically the assumption means that the vector V_j is determined by a scalar (say) V_j^a. Let

$$\alpha \;:=\; e^{-\mathrm{i}2\pi/3}, \qquad \alpha_+ \;:=\; \begin{bmatrix} 1 \\ \alpha \\ \alpha^2 \end{bmatrix} \tag{17.39a}$$

Then, assuming positive sequence,

$$V_j = V_j^a \begin{bmatrix} 1 \\ \alpha \\ \alpha^2 \end{bmatrix} \;=\; V_j^a\, \alpha_+ \tag{17.39b}$$

This makes it possible to determine the off-diagonal entries of S_{jk} from its diagonal entries, as follows. Let λ_{jk} denote the vector consisting of the diagonal entries of S_{jk}:

$$\lambda_{jk} \;:=\; \operatorname{diag}(S_{jk}) := \begin{bmatrix} V_j^a \bar{I}_{jk}^a \\ V_j^b \bar{I}_{jk}^b \\ V_j^c \bar{I}_{jk}^c \end{bmatrix}$$

Using (17.39), the 3×3 line flow matrix S_{jk} is given by:

$$S_{jk} := V_j I_{jk}^{\mathsf{H}} \;=\; V_j^a\, \alpha_+ \begin{bmatrix} \bar{I}_{jk}^a & \bar{I}_{jk}^b & \bar{I}_{jk}^c \end{bmatrix}$$

This expression says that the columns of S_{jk} are in $\operatorname{span}(\alpha_+)$. The first column of the right-hand side is

$$\alpha_+ \underbrace{V_j^a \bar{I}_{jk}^a}_{[S_{jk}]_{11}} = \alpha_+ \left[\lambda_{jk}\right]_1$$

The second column is (noting $\alpha^{-1} = \alpha^2 = \bar{\alpha}$)

$$\alpha_+ V_j^a \bar{I}_{jk}^b = \frac{1}{\alpha}\alpha_+ \left(\alpha V_j^a\right) \bar{I}_{jk}^b \;=\; \frac{1}{\alpha}\alpha_+ \underbrace{V_j^b \bar{I}_{jk}^b}_{[S_{jk}]_{22}} \;=\; \bar{\alpha}\,\alpha_+ \left[\lambda_{jk}\right]_2$$

The third column is (noting $\alpha^{-2} = \alpha = \bar{\alpha}^2$)

$$\alpha_+ V_j^a \bar{I}_{jk}^c = \frac{1}{\alpha^2}\alpha_+ \left(\alpha^2 V_j^a\right) \bar{I}_{jk}^c \;=\; \frac{1}{\alpha^2}\alpha_+ \underbrace{V_j^c \bar{I}_{jk}^c}_{[S_{jk}]_{33}} \;=\; \bar{\alpha}^2\,\alpha_+ \left[\lambda_{jk}\right]_3$$

Putting all this together define

$$\gamma \;:=\; \alpha_+ \alpha_+^{\mathsf{H}} \;=\; \begin{bmatrix} 1 & \alpha^2 & \alpha \\ \alpha & 1 & \alpha^2 \\ \alpha^2 & \alpha & 1 \end{bmatrix}$$

and we can determine the line flow matrix S_{jk} in terms of its diagonal entries:

$$S_{jk} \;=\; \gamma\,\mathrm{diag}\left(\lambda_{jk}\right) \;=\; \left(\alpha_+ \alpha_+^{\mathsf{H}}\right)\mathrm{diag}\left(\lambda_{jk}\right), \qquad j \to k \in E$$

where $\mathrm{diag}(x)$ is a diagonal matrix whose diagonal consists of entries of vector x. Then the linear model that generalizes the single phase linear DistFlow model to three-phase radial networks is (graph is oriented so that all lines point away from bus 0):

$$\sum_{k:\, j \to k} \lambda_{jk} \;=\; \lambda_{ij} \;+\; s_j, \qquad j \in \overline{N} \tag{17.40a}$$

$$S_{jk} \;=\; \gamma\,\mathrm{diag}\left(\lambda_{jk}\right) \;:=\; \alpha_+ \alpha_+^{\mathsf{H}} \mathrm{diag}\left(\lambda_{jk}\right), \qquad j \to k \in E \tag{17.40b}$$

$$v_j - v_k \;=\; z_{jk} S_{jk}^{\mathsf{H}} \;+\; S_{jk} z_{jk}^{\mathsf{H}}, \qquad j \to k \in E \tag{17.40c}$$

where $i := i(j)$ is the unique parent node of j, assuming positive sequence.

Solution.

Given $\left(v_0, s_j, j \in N\right)$, (17.40) can be used to determine explicitly $\left(s_0, v_j, j \in N\right)$ and $\left(\lambda_{jk}, S_{jk}, j \to k \in E\right)$, as follows (Exercise 17.5):

$$\begin{aligned} s_0 \;&=\; -\sum_{j \in N} s_j \\ \lambda_{ij} \;&=\; -\sum_{k \in \mathsf{T}_j} s_k, \qquad S_{ij} \;=\; \gamma\,\mathrm{diag}\left(\lambda_{ij}\right) \;:=\; \alpha_+ \alpha_+^{\mathsf{H}} \mathrm{diag}\left(\lambda_{ij}\right), \qquad i \to j \in E \\ v_j \;&=\; v_0 \;-\; \sum_{(i,k) \in \mathsf{P}_j} \left(z_{ik} S_{ik}^{\mathsf{H}} \;+\; S_{jk} z_{ik}^{\mathsf{H}}\right), \qquad j \in N \end{aligned}$$

where T_j is the subtree rooted at bus j, including j, and P_k is the set of lines on the unique path from bus 0 to bus k; see Figure 17.3. In general the 3×3 solution matrices v_j are not of rank 1 even if $v_0 = |V_0^a|^2 \alpha_+ \alpha_+^{\mathsf{H}}$ is of rank 1 (and even if all lines are symmetric whose series impedances z_{jk} satisfy (15.9a)). This is because

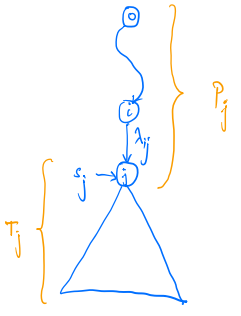


Figure 17.3 Linear solution of branch flow model for unbalanced three-phase radial networks.

the linear model is an approximation and its solution may not satisfy the Kirchhoff's laws. If $v_0 = |V_0^a|^2 \alpha_+ \alpha_+^{\mathsf{H}}$ then v_j is Hermitian and hence has a spectral decomposition. An approximate solution for the voltage phasor V_j can be taken to be largest spectral component of v_0, i.e., if $v_j = \sum_i \rho_i u_i u_i^{\mathsf{H}}$ where ρ_i are real eigenvalues and u_i are eigenvectors of v_j with $|\rho_1| \geq |\rho_2| \geq |\rho_3|$, then $V_j = \sqrt{\rho_1} u_1$ if $\rho_1 > 0$ or $V_j = -\sqrt{\rho_1} u_1$ if $\rho_1 < 0$.

17.5.2 Application example

We describe a voltage regulation algorithm adapted from [175] to illustrate the three-phase linear model.

17.6 Bibliographical notes

Algorithms for solving power flows in three-phase radial networks are developed in [41, 42, 43, 45, 47, 50]. For backward forward sweep methods for radial networks, both single-phase and three-phase networks, see bibliographical notes in Chapter 5.5.

17.7 Problems

Chapter 17.1.

Exercise 17.1. Show that $(Ab) \otimes \mathbb{I} = (A \otimes \mathbb{I})(b \otimes \mathbb{I})$ where $A \in \mathbb{C}^{n \times n}$, $b \in \mathbb{C}^n$, and $\mathbb{I}$ is the identity matrix of size 3. (Hint: Use Lemma 16.6.)

Exercise 17.2 (BFM without shunt admittances)**.** Derive the generalized three-phase DistFlow equations (17.4) by substituting $\tilde{y}_{jk} = \tilde{y}_{kj} = y^s_{jk} = y^s_{kj}$, $y^m_{jk} = y^m_{kj} = 0$, and (17.3) into (17.1).

Chapter 17.2.

Chapter 17.3.

Chapter 17.4.

Exercise 17.3 (Backward forward sweep)**.** This exercise solves the same overall model as the BFS described in Chapter 17.4.1, but here, instead of $I^s_{jk} \in \mathbb{C}^3$ over the series impedance, we are to derive a BFS algorithm to compute the sending-end current $I_{jk} \in \mathbb{C}^3$ for every line $j \to k$, as well as the nodal variable $y := \left(V_j, I_j, I^\Delta_j \; j \in N\right)$. It extends Exercise 5.6 from single-phase radial networks to three-phase radial networks.

Exercise 17.4 (Backward forward sweep)**.** Extend the BFS described in Chapter 5.3.3 from single-phase to three-phase radial networks. This allows the inclusion of PV buses where real power and voltage magnitudes are given instead of internal powers.

Chapter 17.5.

Exercise 17.5 (Three-phase BFM linear solution)**.** Given $\left(v_0, s_j, j \in N\right)$, show that an explicit solution $\left(s_0, v_j, j \in N, S_{jk}, j \to k \in E\right)$ of (17.40) is

$$s_0 = -\sum_{j \in N} s_j, \qquad \lambda_{ij} = -\sum_{k \in \mathsf{T}_j} s_k, \qquad i \to j \in E \tag{17.41a}$$

$$S_{ij} = \gamma \operatorname{diag}\left(\lambda_{ij}\right) := \alpha_+ \alpha_+^{\mathsf{H}} \operatorname{diag}\left(\lambda_{ij}\right), \qquad i \to j \in E \tag{17.41b}$$

$$v_j = v_0 - \sum_{(i,k) \in \mathsf{P}_j} \left(z_{ik} S_{ik}^{\mathsf{H}} + S_{jk} z_{ik}^{\mathsf{H}}\right), \qquad j \in N \tag{17.41c}$$

where T_j is the subtree rooted at bus j, including j, and P_k is the set of lines on the unique path from bus 0 to bus k.

18 Power flow optimization

In this chapter we study optimal power flow (OPF) problems for unbalanced three-phase networks. As for single-phase networks studied in Chapter 9, OPF is a constrained optimization that takes the form

$$\min_{u,x} \; c(u,x) \qquad \text{subject to} \qquad f(u,x) = 0, \;\; g(u,x) \leq 0$$

The cost function c may represent generation cost, voltage deviation, power loss, or user disutility. The variable u collects control decisions such as generator commitment, generation setpoints, transformer taps, capacitor switch status, electric vehicle charging levels, thermostatic settings, or inverter reactive power. The variable x collects network state such as voltage levels, line currents, or power flows. The constraint functions f, g describe current or power balance, generation or consumption limits, voltage or line limits, and stability and security constraints, as well as other operational requirements. OPF is a fundamental problem because it underlies numerous power system operation and planning applications. While the structure of OPF remains the same as for single-phase networks, in this chapter, the cost function c and constraint functions f, g models three-phase devices and networks.

In Chapter 18.1 we formulate OPF in both the bus injection model and the branch flow model. In Chapters 18.2 and 18.3 we derive their semidefinite relaxations. Finally we illustrate in Chapter 18.4 these results in example applications.

18.1 Three-phase OPF

In Chapter 18.1.1 we describe device models that will be used in both the bus injection model and the branch flow model. We formulate in Chapter 18.1.2 OPF in the bus injection model and show in Chapter 18.1.3 that it is equivalent to a nonconvex quadratically constrained quadratic program (QCQP), generalizing OPF in Chapter 9.1 from a single-phase to three-phase setting. In Chapter 18.1.4 we formulate OPF in the branch flow model for radial networks.

18.1.1 Three-phase devices

A key assumption underlying our OPF formulation is that all controllable devices are the single-phase devices that make up three-phase devices. Therefore internal variables u_j are optimization variables (i.e., $V_j^{Y/\Delta}$ for voltage sources, $I_j^{Y/\Delta}$ for current sources, $\left(s_j^{Y/\Delta}, I_j^{\Delta}\right)$ for power sources). Their values determine the terminal variables (V_j, I_j, s_j) through conversion rules. These terminal variables interact over the network through either the current balance equation $I = YV$ or the power balance equation, but they are typically not directly controllable. In this chapter we use the power balance equation to relate the terminal voltages and power injections (V, s). A device model therefore consists of:

- A conversion rule (and external models of impedances) from Chapter 14.3 that relates an internal variable u_j device j to its terminal voltage and power (V_j, s_j).
- Operational constraints on the internal variable u_j. These constraints are local to j.

We describe each of them next.

Conversion rules.

1 *Voltage source* $u_j := V_j^{Y/\Delta}$: For an ideal voltage source its internal voltage $V_j^{Y/\Delta} \in \mathbb{C}^3$ is an optimization variable. It is related to the terminal voltage V_j through a linear constraint (from the conversion rules (14.8) and (14.9a)):

$$Y \text{ configuration:} \quad V_j = V_j^Y + \gamma_j^Y \mathbf{1} \tag{18.1a}$$

$$\Delta \text{ configuration:} \quad \Gamma V_j = V_j^{\Delta} \tag{18.1b}$$

We assume here that the neutral voltage $\gamma_j^Y := V_j^n$ of a Y-configured device is a given parameter. For example, $\gamma_j^Y = 0$ if the neutral of the Y-configured device directly grounded and all voltages are defined with respect to the ground.

2 *Current source* $u_j := I_j^{Y/\Delta}$: For an ideal current source its internal current $I_j^{Y/\Delta} \in \mathbb{C}^3$ is an optimization variable. It is related to the terminal variables $\left(V_j, s_j\right)$ through a quadratic constraint (from the conversion rules (14.8) and (14.10c)):

$$Y \text{ configuration:} \quad s_j = -\operatorname{diag}\left(V_j I_j^{Y\mathsf{H}}\right) \tag{18.1c}$$

$$\Delta \text{ configuration:} \quad s_j = -\operatorname{diag}\left(V_j I_j^{\Delta\mathsf{H}} \Gamma\right) \tag{18.1d}$$

3 *Power source* $u_j := \left(s_j^{Y/\Delta}, I_j^{Y/\Delta}\right)$: For an ideal power source we assume that the internal power and current $\left(s_j^{Y/\Delta}, I_j^{Y/\Delta}\right)$ are optimization variables. We assume

the neutral voltage $\gamma_j^Y := V_j^n$ of a Y-configured power source is a given parameter, e.g., $\gamma_j^Y = 0$ if the neutral is directly grounded and all voltages are defined with respect to ground. They are related to the terminal variables (V_j, s_j) according to the conversion rules (14.8) and (14.10c):

$$Y \text{ configuration:} \quad s_j = -\operatorname{diag}\left(V_j I_j^{Y\mathsf{H}}\right), \quad s_j^Y = -s_j - \gamma_j^Y \bar{I}_j^Y \tag{18.1e}$$

$$\Delta \text{ configuration:} \quad s_j = -\operatorname{diag}\left(V_j I_j^{\Delta\mathsf{H}} \Gamma\right), \quad s_j^\Delta = \operatorname{diag}\left(\Gamma V_j I_j^{\Delta\mathsf{H}}\right) \tag{18.1f}$$

For a Y-configured power source, if $\gamma_j^Y = 0$, then the optimization variable is s_j^Y and the conversion rule reduces to

$$Y \text{ configuration:} \quad s_j = -s_j^Y$$

It is possible to formulate OPF in which a power source is characterized only by its internal power $u_j := s_j^{Y/\Delta}$ instead of $u_j := \left(s_j^{Y/\Delta}, I_j^{Y/\Delta}\right)$, but the formulation is more complicated; see Exercise 18.5.

4 *Impedance* $\left(z_j^Y, \gamma_j^Y\right)$ *or* z_j^Δ: An impedance, if not controllable, does not introduce any addition optimization variable but imposes an additional constraint on the terminal variables (V_j, s_j) (from (14.19a) and Theorem 14.4):

$$Y \text{ configuration:} \quad s_j = -\operatorname{diag}\left(V_j \left(V_j - \gamma_j^Y \mathbf{1}\right)^{\mathsf{H}} y_j^{Y\mathsf{H}}\right) \tag{18.1g}$$

$$\Delta \text{ configuration:} \quad s_j = -\operatorname{diag}\left(V_j V_j^{\mathsf{H}} Y_j^{\Delta\mathsf{H}}\right) \tag{18.1h}$$

where $y_j^{Y/\Delta} := \left(z_j^{Y/\Delta}\right)^{-1}$, $Y_j^\Delta := \Gamma^{\mathsf{T}} y^\Delta \Gamma$. The neutral voltage $\gamma_j^Y := V_j^n$ is usually a fixed parameter, e.g. $\gamma_j^Y = 0$.

The conversion rule (18.1) takes the form $f_j^{Y/\Delta}\left(u_j, V_j, s_j\right) = 0$ and is local to each bus j. Note the structural similarity between Y and Δ configurations when $\gamma_j^Y := V_j^n = 0$: (18.1) reduces to

Voltage source:	$V_j = V_j^Y$,	$\Gamma V_j = V_j^\Delta$
Current source:	$s_j = -\operatorname{diag}\left(V_j I_j^{Y\mathsf{H}}\right)$,	$s_j = -\operatorname{diag}\left(V_j I_j^{\Delta\mathsf{H}} \Gamma\right)$
Power source:	$s_j = -\operatorname{diag}\left(V_j I_j^{Y\mathsf{H}}\right)$,	$s_j = -\operatorname{diag}\left(V_j I_j^{\Delta\mathsf{H}} \Gamma\right)$
	$s_j^Y = -s_j$,	$s_j^\Delta = \operatorname{diag}\left(\Gamma V_j I_j^{\Delta\mathsf{H}}\right)$
Impedance:	$s_j = -\operatorname{diag}\left(V_j V_j^{\mathsf{H}} y_j^{Y\mathsf{H}}\right)$,	$s_j = -\operatorname{diag}\left(V_j V_j^{\mathsf{H}} Y_j^{\Delta\mathsf{H}}\right)$

Once an optimal solution $\left(u_j^{\text{opt}}, V_j^{\text{opt}}, s_j^{\text{opt}}\right)$ of an OPF problem is chosen, other internal variables for each device j can be derived (possibly requiring additional information e.g. β_j of an ideal voltage source).

Remark 18.1 (Implicit optimization over (γ_j, β_j))**.** The constraint (18.1b) for a Δ-configured device does not determined the terminal voltage V_j uniquely and therefore an optimal V_j also determines an optimal zero-sequence voltage $\gamma_j^\Delta := \frac{1}{3}\mathbf{1}^\mathsf{T} V_j$. If γ_j^Δ is given instead, then (18.1b) should be replaced by $V_j = \Gamma^\dagger V_j^\Delta + \gamma_j \mathbf{1}$. Similarly for other devices, e.g., Δ-configured impedance.

Optimization over I_j^Δ in current source and power source implicitly chooses an optimal zero-sequence current $\beta_j := \frac{1}{3}\mathbf{1}^\mathsf{T}\beta_j^\Delta$. If β_j is given then it imposes an additional constraint through the conversion rule $I_j^\Delta = -\frac{1}{3}\Gamma I_j + \beta_j \mathbf{1}$ (and express I_j in terms of (V_j, s_j)). □

Device constraints.

The operational constraints on the devices are also local to each bus j and are inequality constraints on the internal variables u_j only, of the form $g_j^{Y/\Delta}(u_j) \le 0$:

1 *Voltage source* $u_j := V_j^{Y/\Delta}$:

$$v_j^{Y/\Delta\min} \le \operatorname{diag}\left(u_j u_j^\mathsf{H}\right) \le v_j^{Y/\Delta\max} \tag{18.2a}$$

2 *Current source* $u_j := I_j^{Y/\Delta}$:

$$\operatorname{diag}\left(u_j u_j^\mathsf{H}\right) \le \ell_j^{Y/\Delta\max} \tag{18.2b}$$

3 *Power source* $u_j := (u_{j1}, u_{j2}) := \left(s_j^{Y/\Delta}, I_j^{Y/\Delta}\right)$:

$$s_j^{Y/\Delta\min} \le u_{j1} \le s_j^{Y/\Delta\max}, \qquad \operatorname{diag}\left(u_{j2} u_{j2}^\mathsf{H}\right) \le \ell_j^{Y/\Delta\max} \tag{18.2c}$$

18.1.2 Bus injection model

Consider a three-phase network modeled as an undirected graph $G := (\overline{N}, E)$ where there are $N+1$ buses $j \in \overline{N}$ and M lines in E. Each line $(j,k) \in E$ is characterized by 3×3 admittance matrices $\left(y_{jk}^s, y_{jk}^m\right) \in \mathbb{C}^6$ and $\left(y_{kj}^s, y_{kj}^m\right) \in \mathbb{C}^6$. We now explain the variables, power flow equations, cost function, and constraints that define an OPF problem. As we will see the OPF formulation (18.5) below does not require the assumption $y_{jk}^s = y_{kj}^s$ (C16.1 for BIM and C17.1 for BFM). It can therefore accommodate standard three-phase transformers, e.g., in ΔY and $Y\Delta$ configurations. As for the single-phase OPF we studied in Chapter 9.1.2 we assume there is exactly one three-phase device at each bus j. We will then interchangeably refer to j as a bus or a device. See Chapter 9.1.2 on how to relax this assumption. We now describe the optimization variables, network equations and operational constraints, as well as a cost function that define a three-phase OPF.

Optimization variables.

As mentioned above a key assumption underlying our formulation is that all controllable devices are the single-phase devices that make up three-phase devices. There are therefore two types of optimization variables (u,x). The internal variable $u := (u_j, j \in \overline{N})$ represents controllable quantities of the three-phase devices discussed in Chapter 18.1.1. The terminal variable $x := \left(V_j, s_j, j \in \overline{N}\right)$ represents the terminal voltages and power injections. The conversion rule relates u to x which interact over the network through either the current balance equation $I = YV$ or the power balance equation. The terminal variables are typically not directly controllable (even though they are optimization variables).

Network constraints.

The power flow equations relate the terminal variables $x := (V, s)$, from (16.12):

$$s_j \;=\; \sum_{k:j\sim k} \operatorname{diag}\left(V_j (V_j - V_k)^{\mathsf{H}} \left(y^s_{jk}\right)^{\mathsf{H}} \;+\; V_j V_j^{\mathsf{H}} \left(y^m_{jk}\right)^{\mathsf{H}}\right), \qquad j \in \overline{N} \tag{18.3}$$

which directly extend the single-phase equations (9.3). This constraint is global as it couples voltages and powers (V_j, s_j) at all neighboring buses.

The operational constraints on $x := (V, s)$ are the same as (9.4) for single-phase OPF, except that the variables and their bounds are 3-dimensional vectors, rather than scalars, for three-phase networks:

$$\text{injection limits:} \quad s^{\min}_j \;\le\; s_j \;\le\; s^{\max}_j, \qquad j \in \overline{N} \tag{18.4a}$$

$$\text{voltage limits:} \quad v^{\min}_j \;\le\; \operatorname{diag}\left(V_j V_j^{\mathsf{H}}\right) \;\le\; v^{\max}_j, \qquad j \in \overline{N} \tag{18.4b}$$

$$\text{line limits:} \quad \operatorname{diag}\left(I_{jk}(V)\, I^{\mathsf{H}}_{jk}(V)\right) \;\le\; \ell^{\max}_{jk}, \quad \operatorname{diag}\left(I_{kj}(V)\, I^{\mathsf{H}}_{kj}(V)\right) \;\le\; \ell^{\max}_{kj}, \quad (j,k) \in E \tag{18.4c}$$

where $\left(I_{jk}(V), I_{kj}(V)\right)$ in (18.4c) are given by (16.1) reproduced here:

$$I_{jk}(V) \;=\; y^s_{jk}\left(V_j - V_k\right) \;+\; y^m_{jk} V_j, \qquad I_{kj}(V) \;=\; y^s_{kj}\left(V_k - V_j\right) \;+\; y^m_{kj} V_k$$

The constraint (18.4a) can be due to limits on the busbar to which the three-phase device is connected. The constraints (18.4a)(18.4b) are local at each bus j but (18.4c) is global.

Cost function.

As for single-phase OPF, the cost function $C(u,x)$ may represent generation cost, real power loss, estimation error, voltage deviations, or user disutility, depending on

applications. For instance to minimize the cost of real power generations we can use

$$C(u,x) \;:=\; C(u,V,s) \;:=\; \sum_{\text{gens. } j} c_j \mathbf{1}^{\mathsf{T}} \mathrm{Re}\left(s_j^{Y/\Delta}\right)$$

Other example costs include estimation error in state estimation and user disutility in demand response.

OPF.

Define the feasible set

$$\mathbb{V}_{3p} \;:=\; \{(u,x) := (u,V,s) \mid (u,x) \text{ satisfies (18.1)(18.2)(18.3)(18.4)}\} \tag{18.5a}$$

Then the simple OPF formulation in the three-phase setting is

$$\min_{(u,x)} \;\; C(u,x) \;\text{ s.t. } (u,x) \in \mathbb{V}_{3p} \tag{18.5b}$$

The constraint (18.2) describes local operational constraints on the internal variables u of the three-phase devices, (18.3)(18.4) describe the network equation and operational constraints on the terminal variable $x := (V,s)$, and the conversion rule (18.1) relates u and x and is also a local constraint. Since the constraints (18.3)(18.4c) do not require assumption C16.1 that $y^s_{jk} = y^s_{kj}$, the OPF formulation (18.5) can accommodate three-phase transformers whose admittance matrices Y are not block symmetric, e.g., transformers in ΔY and $Y\Delta$ configurations.

Remark 18.2 (Uncontrollable parameters)**.** As for single-phase OPF, the formulation (18.5) allows the case where a quantity is not an optimization variable but a given parameter. For instance a given uncontrollable constant-power load or a given renewable generation at bus j can be represented by setting $s_j^{Y/\Delta} = s_j^{Y/\Delta\,\min} = s_j^{Y/\Delta\,\max}$ to the specified value. □

Structurally the three-phase OPF (18.5) takes the form with $x := (V,s)$:

$$\min_{(u,x)} \;\; C(u,x) \tag{18.6a}$$

$$\text{s.t.} \;\; f_j^{Y/\Delta}(u_j,V_j,s_j) \;=\; 0, \quad g_j^{Y/\Delta}(u_j) \;\le\; 0, \qquad j \in \overline{N} \tag{18.6b}$$

$$f(V,s) \;=\; 0, \quad g(V,s) \;\le\; 0 \tag{18.6c}$$

where the local constraint (18.6b) represents operational constraints (18.2) on the internal variables u_j of device j and the conversion rules (18.1) that relate u_j to its terminal variables x_j, and the global constraint (18.6c) represents the power flow equation (18.3) and operational constraint (18.4) on the terminal variable x. The local constraint (18.6b) generalizes (9.6) from single-phase systems to three-phase systems.

18.1.3 Three-phase OPF as QCQP

The three-phase OPF (18.5) can be written as a QCQP in (V,u), following the same procedure studied in Chapter 9.1.3 for single-phase OPF.

Device constraints as quadratic forms.

We start by writing the local device constraints (18.2), represented by $g_j^{Y/\Delta}(u_j) \le 0$ in (18.6b), as quadratic forms. Let

$$e^a := (1,0,0), \quad e^b := (0,1,0), \quad e^c := (0,0,1), \quad E^\phi := e^\phi e^{\phi\mathsf{T}} \in \mathbb{C}^{3\times 3}, \quad \phi \in \{a,b,c\} \tag{18.7}$$

Then the device constraints (18.2) become the quadratic forms local to each bus j:

1 *Voltage source* $u_j := V_j^{Y/\Delta}$:

$$v_j^{(Y/\Delta)\phi\min} \le u_j^{\mathsf{H}} E^\phi u_j \le v_j^{(Y/\Delta)\phi\max} \tag{18.8a}$$

2 *Current source* $u_j := I_j^{Y/\Delta}$:

$$u_j^{\mathsf{H}} E^\phi u_j \le \ell_j^{(Y/\Delta)\phi\max} \tag{18.8b}$$

3 *Power source* $u_j := (u_{j1}, u_{j2}) := \left(s_j^{Y/\Delta}, I_j^{Y/\Delta}\right)$:

$$s_j^{Y/\Delta\min} \le u_{j1} \le s_j^{Y/\Delta\max}, \qquad u_{j2}^{\mathsf{H}} E^\phi u_{j2} \le \ell_j^{(Y/\Delta)\phi\max} \tag{18.8c}$$

Network constraints as quadratic forms.

Next we eliminate the power flow equation (18.3), represented by $f(V,s) = 0$ in (18.6c), by substituting $s_j(V)$ as functions of V into the network constraint (18.4) on the terminal variables represented by $f(V,s) = 0$ in (18.6c). This reduces (18.6c) to a single inequality constraint of the form

$$g(V, s(V)) \le 0$$

where components of g are quadratic forms in V. The conversion into quadratic forms follows the same derivation in Chapter 9.1.3, but applied to the single-phase equivalent circuit.

Let

$$e_j \in \{0,1\}^{N+1}, \quad e_j^\phi \in \{0,1\}^{3(N+1)}, \quad E_j^\phi := e_j^\phi \left(e_j^\phi\right)^{\mathsf{H}}, \quad \phi \in \{a,b,c\} \tag{18.9}$$

where e_j is of size $N+1$ and has a single 1 in its jth position, e_j^ϕ is of size $3(N+1)$ and has a single 1 in its $j\phi$th position, and E_j^ϕ is the $3(N+1)\times 3(N+1)$ diagonal Hermitian matrix with a single 1 in the $(j\phi, j\phi)$th entry and 0 everywhere else.

1 *Injection limits*: Let $Y \in \mathbb{C}^{3(N+1)\times 3(N+1)}$ denote the single-phase equivalent admittance matrix. Define the matrix $Y_j^{\phi} := e_j^{\phi}\, e_j^{\phi \mathsf{H}} Y$ where $e_j^{\phi} \in \{0,1\}^{3(N+1)}$ is the unit vector with a single 1 at the (j,ϕ)th entry and 0 elsewhere. Define the Hermitian and skew Hermitian components of $Y_j^{\phi \mathsf{H}}$:

$$\Phi_j^{\phi} \;:=\; \frac{1}{2}\left(Y_j^{\phi \mathsf{H}} + Y_j^{\phi}\right) \quad \text{and} \quad \Psi_j^{\phi} \;:=\; \frac{1}{2\mathbf{i}}\left(Y_j^{\phi \mathsf{H}} - Y_j^{\phi}\right) \tag{18.10a}$$

Then

$$p_j^{\phi} \;:=\; \mathrm{Re}\left(s_j^{\phi}\right) \;=\; V^{\mathsf{H}} \Phi_j^{\phi} V \quad \text{and} \quad q_j^{\phi} \;:=\; \mathrm{Im}(s_j^{\phi}) \;=\; V^{\mathsf{H}} \Psi_j^{\phi} V$$

and the injection limits become

$$p_j^{\phi\,\min} \;\le\; V^{\mathsf{H}} \Phi_j^{\phi} V \;\le\; p_j^{\phi\,\max}, \quad q_j^{\phi\,\min} \;\le\; V^{\mathsf{H}} \Psi_j^{\phi} V \;\le\; q_j^{\phi\,\max}, \quad j \in \overline{N} \tag{18.10b}$$

2 *Voltage limits*: The terminal voltage limits are

$$v_j^{\phi\,\min} \;\le\; V^{\mathsf{H}} E_j^{\phi} V \;\le\; v_j^{\phi\,\max}, \qquad j \in \overline{N} \tag{18.10c}$$

where E_j^{ϕ} is defined in (18.9).

3 *Line limits*: The same derivation as that for single-phase OPF shows that the limit on the sending-end current I_{jk}^{ϕ} in the phase-a line is (Exercise 18.1)

$$\left|I_{jk}^{\phi}\right|^2 \;:=\; V^{\mathsf{H}} \hat{Y}_{jk}^{\phi} V \;\le\; \ell_{jk}^{\phi\,\max}, \qquad (j,k) \in E \tag{18.10d}$$

where $\hat{Y}_{jk}^{\phi} := \tilde{Y}_{jk}^{\mathsf{H}} E^{\phi} \tilde{Y}_{jk}$ is a $3(N+1)\times 3(N+1)$ matrix and $\tilde{Y}_{jk}$ is a $3\times 3(N+1)$ matrix given by

$$\tilde{Y}_{jk} \;:=\; \left((e_j - e_k)^{\mathsf{T}} \otimes y_{jk}^{s} \;+\; e_j^{\mathsf{T}} \otimes y_{jk}^{m}\right)$$

(Here E^{ϕ} is defined in (18.7) and e_j in (18.9).) The matrix $\hat{Y}_{jk}$ is Hermitian and hence $V^{\mathsf{H}} \hat{Y}_{jk}^{\phi} V$ is indeed a real number. Similarly for $\left|I_{kj}^{\phi}\right|^2$.

Conversion rules as quadratic forms.

Finally we eliminate s_j from the the conversion rule (18.1) for three-phase devices, represented by the local equality constraint $f_j^{Y/\Delta}\left(u_j, V_j, s_j\right) = 0$ in (18.6b). This reduces $f_j^{Y/\Delta}\left(u_j, V_j, s_j\right) = 0$ to an equality constraint of the form

$$f_j^{Y/\Delta}\left(u_j, V, s_j(V)\right) \;=\; 0, \qquad j \in \overline{N}$$

where $f_j^{Y/\Delta}$ is a quadratic form in (u_j, V). It also transforms the original local constraints into global constraints since the function $s_j(V)$ depends on V_k at all neighbors k of j; see (18.11).

Recall that $s_j(V) := \left(s_j^a(V), s_j^b(V), s_j^c(V)\right)$ and

$$s_j^\phi(V) \;=\; V^{\mathsf{H}}\left(Y_j^{\phi\mathsf{H}}\right)V \;=\; V^{\mathsf{H}}\left(\Phi_j^\phi + \mathbf{i}\Psi_j^\phi\right)V, \qquad \phi \in \{a,b,c\},\ j \in \overline{N} \tag{18.11}$$

where $Y_j^\phi := e_j^\phi e_j^{\phi\mathsf{T}} Y$ and Φ_j^ϕ and Ψ_j^ϕ are defined in (18.10a). Then $V_j \in \mathbb{C}^3$ can be written in terms of $V \in \mathbb{C}^{3(N+1)}$ as follows:

$$V_j \;=\; \left(e_j \otimes \mathbb{I}\right)^{\mathsf{H}} V \;=\; \left(e_j^{\mathsf{H}} \otimes \mathbb{I}\right)V, \qquad V_j^\phi \;=\; e_j^{\phi\mathsf{H}} V, \quad \phi \in \{a,b,c\} \tag{18.12}$$

where $\mathbb{I}$ is the identity matrix of size 3.

We now use (18.7)(18.9)(18.11)(18.12) to convert the conversion rule $f_j^{Y/\Delta}$ in (18.1) into inhomogeneous quadratic forms in (u_j, V). They can then be homogenized using the identity (9.15) in Remark 9.4.

1 *Voltage source* $u_j := V_j^{Y/\Delta}$: Application of (18.12) to the conversion rules (18.1a) (18.1b) leads to the following linear constraints in (u_j, V):

Y configuration: $$\left(e_j^{\mathsf{H}} \otimes \mathbb{I}\right)V \;=\; u_j + \gamma_j^Y \mathbf{1} \tag{18.13a}$$

Δ configuration: $$\Gamma\left(e_j^{\mathsf{H}} \otimes \mathbb{I}\right)V \;=\; u_j \tag{18.13b}$$

where $\gamma_j^Y := V_j^n$ is assumed given (e.g., $\gamma_j^Y = 0$). These constraints remain local as they depend on V only through $V_j \in \mathbb{C}^3$.

2 *Current source* $u_j := I_j^{Y/\Delta}$: The conversion rules (18.1c)(18.1d) for a current source are equivalent to the following inhomogeneous quadratic equations in (u_j, V) (Exercise 18.2):

Y configuration: $$s_j^\phi(V) \;=\; -u_j^{\mathsf{H}}\left(e_j^{\mathsf{H}} \otimes E^\phi\right)V \tag{18.13c}$$

Δ configuration: $$s_j^\phi(V) \;=\; -u_j^{\mathsf{H}}\left(e_j^{\mathsf{H}} \otimes (\Gamma E^\phi)\right)V \tag{18.13d}$$

where $s_j(V)$ is given in (18.11). These constraints are global as $s_j(V)$ depend on V_k at neighboring buses k.

3 *Power source* $u_j := \left(s_j^{Y/\Delta}, I_j^{Y/\Delta}\right)$: For a Y-configured power source let $u_j =: (u_{j1}, u_{j2})$ where $u_{j1} := s_j^Y$ and $u_{j2} := I_j^Y$. Then the conversion rule (18.1e) is equivalent to the following inhomogeneous quadratic equations in (u_j, V) (Exercise 18.3):

$$Y: \quad s_j^\phi(V) \;=\; -u_{j2}^{\mathsf{H}}\left(e_j^{\mathsf{H}} \otimes E^\phi\right)V, \quad s_j(V) \;=\; -u_{j1} - \gamma_j^Y \bar{u}_{j2}, \quad \phi \in \{a,b,c\} \tag{18.13e}$$

where $s_j(V)$ is given in (18.11) and $\gamma_j^Y := V_j^n$ is assumed given (e.g., $\gamma_j^Y = 0$).

For a Δ-configured power source let $u_j =: (u_{j1}, u_{j2})$ where $u_{j1} := s_j^\Delta$ and $u_{j2} := I_j^\Delta$. Then the conversion rule (18.1f) is equivalent to the following inhomogeneous quadratic equations in (u_j, V) (Exercise 18.3):

$$\Delta: \quad s_j^\phi(V) = -u_{j2}^{\mathsf{H}} \left(e_j^{\mathsf{H}} \otimes (\Gamma E^\phi) \right) V, \quad u_{j1}^{\phi\varphi} = u_{j2}^{\mathsf{H}} \left(e_j^{\mathsf{H}} \otimes \left(E^\phi \Gamma \right) \right) V, \quad \phi\varphi \in \{ab, bc, ca\} \tag{18.13f}$$

where $s_j(V)$ is given in (18.11).

4 *Impedance* $\left(z_j^Y, \gamma_j^Y \right)$ *or* z_j^Δ: The equality constraint (18.1g) or (18.1h) imposed by an impedance $\left(z_j^Y, \gamma_j^Y \right)$ or z_j^Δ respectively is equivalent to the following inhomogeneous quadratic equation in V (Exercise 18.4):

$$Y \text{ configuration:} \quad s_j^\phi(V) = V^{\mathsf{H}} \left(\left(e_j e_j^{\mathsf{H}} \right) \otimes \left(y_j^{Y\mathsf{H}} E^\phi \right) \right) V - \bar{\gamma}_j \left(e_j^{\mathsf{H}} \otimes \left(\mathbf{1}^{\mathsf{H}} y_j^{Y\mathsf{H}} E^\phi \right) \right) V \tag{18.13g}$$

$$\Delta \text{ configuration:} \quad s_j^\phi(V) = -V^{\mathsf{H}} \left(\left(e_j e_j^{\mathsf{H}} \right) \otimes \left(Y_j^{\Delta\mathsf{H}} E^\phi \right) \right) V \tag{18.13h}$$

where $\gamma_j^Y := V_j^n$ is assumed given (e.g., $\gamma_j^Y = 0$), $Y_j^\Delta := \Gamma^{\mathsf{T}} y_j^\Delta \Gamma$ and $y_j^\Delta := \left(z_j^\Delta \right)^{-1}$.

Note the structural similarity between Y and Δ configurations when $\gamma_j^Y := V_j^n = 0$

Three-phase OPF as QCQP.

We have thus eliminated the power flow equation $f(V, s) = 0$ in (18.6), and expressed the local device constraints $g_j^{Y/\Delta}(u_j) \le 0$, the network constraints $g(V, s(V)) \le 0$, and conversion rules $f_j^{Y/\Delta}(u_j, V_j, s_j(V)) = 0$ as quadratic forms in (u, V). Therefore (18.6) is equivalent to the inhomogeneous QCQP (assuming C is also expressed as a quadratic form in (u, V)):

$$\min_{(u,V)} \quad C(u, V, s(V)) \tag{18.14a}$$

$$\text{s.t.} \quad (18.8)\ (18.10)\ (18.13) \tag{18.14b}$$

where $s(V)$ is given by (18.11). The inhomogeneous quadratic constraints in (18.14b) can be homogenized (see (9.15) in Remark 9.4).

18.1.4 Branch flow model: radial networks

Since the branch flow model is most useful for radial networks, we make the same assumptions as in the single-phase setting studied in Chapter 9.2:

- $z_{jk}^s = z_{kj}^s$, or equivalently $y_{jk}^s = y_{kj}^s$, for every line (j, k) (assumption C17.1).

- $y^m_{jk} = y^m_{kj} = 0$ for every line (j,k). This is a reasonable assumption on distribution lines where y^m_{jk} and y^m_{kj} are typically much smaller in magnitude than the series admittance y^s_{jk}.

Consider a three-phase radial network $G = (\overline{N}, E)$ with $N+1$ buses and $M = N$ lines. The assumptions allow us to adopt a directed graph $G = (\overline{N}, E)$ and include branch variables in only one direction. We denote a line in E from bus j to bus k either by $(j,k) \in E$ or $j \to k$. It is characterized by its series impedance z^s_{jk}. Without loss of generality we take bus 0 as the root of the tree. We now describe the three-phase optimization variables, device models, power flow equations, operational constraints, and the cost function that define a three-phase OPF problem.

Optimization variables.

As in BIM, we assume that only the single-phase devices that make up three-phase devices are directly controllable. There are therefore two types of optimization variables (u,x). The internal variable $u := (u_j, j \in \overline{N})$ represents controllable quantities of the three-phase devices, as in BIM. The variable x represents both the terminal variables (e.g., a nodal voltage V_j) as well as the line variables (e.g., a line power S_{jk}). The variables x interact over the network through the power balance equation. Both BIM and BFM use the same device models and their operational constraints. Their difference lies in the power flow equations that, for BFM, include line variables as well.

Device constraints.

The device models are described in Chapter 18.1.1. The internal variables u_j, $j \in \overline{N}$, their conversion rules (18.1) and operational constraints (18.2) on u_j are the same as for the bus injection model.

Network constraints.

Power flow equations relate the following terminal variables and line variables (see Chapter 17.1.3 for three-phase branch flow model):

$$\begin{array}{llll} s_j \in \mathbb{C}^3, & v_j \in \mathbb{S}^3_+, & V_j \in \mathbb{C}^3, & j \in \overline{N} \\ \ell_{jk} \in \mathbb{S}^3_+, & S_{jk} \in \mathbb{C}^{3\times 3}, & \tilde{I}_{jk} \in \mathbb{C}^3, & j \to k \in E \end{array}$$

where $\mathbb{S}^n_+ \subseteq \mathbb{C}^{n\times n}$ is the set of of $n \times n$ complex (Hermitian and) positive semidefinte matrices. Let $s := (s_j, j \in \overline{N}), v := (v_j, j \in \overline{N}), \ell := (\ell_{jk}, (j,k) \in E), S := (S_{jk}, (j,k) \in E)$. Here (s, v, ℓ, S) directly generalize the corresponding variables in the single-phase model. The voltage phasor $V := \left(V_j, j \in \overline{N}\right)$ is introduced here in order to express the conversion rule (18.1) for three-phase devices and the line current phasor $\tilde{I} := \left(\tilde{I}_{jk}, j \to k \in E\right)$ is introduced for convenience. Let $x := (s, v, \ell, S, V, \tilde{I})$.

The power flow equations we use are (17.4) in Chapter 17.1, reproduced here:

$$\sum_{k:j\to k} \operatorname{diag}(S_{jk}) \;=\; \sum_{i:i\to j} \operatorname{diag}\left(S_{ij} - z_{ij}^{s}\ell_{ij}\right) + s_j, \qquad j \in \overline{N} \tag{18.15a}$$

$$v_j - v_k \;=\; \left(z_{jk}^{s}\, S_{jk}^{\mathsf{H}} + S_{jk} z_{jk}^{s\mathsf{H}}\right) - z_{jk}^{s}\,\ell_{jk}\, z_{jk}^{s\mathsf{H}}, \qquad j \to k \in E \tag{18.15b}$$

$$\begin{bmatrix} v_j & S_{jk} \\ S_{jk}^{\mathsf{H}} & \ell_{jk} \end{bmatrix} \geq 0, \qquad \operatorname{rank}\begin{bmatrix} v_j & S_{jk} \\ S_{jk}^{\mathsf{H}} & \ell_{jk} \end{bmatrix} = 1, \qquad j \to k \in E \tag{18.15c}$$

$$v_j \;=\; V_j V_j^{\mathsf{H}}, \qquad \ell_{jk} \;=\; \tilde{I}_{jk}\tilde{I}_{jk}^{\mathsf{H}}, \qquad S_{jk} \;=\; V_j \tilde{I}_{jk}^{\mathsf{H}}, \qquad j \to k \in E \tag{18.15d}$$

where $V_0 \in \mathbb{C}^3$ is given and bus $i := i(j)$ is the unique parent of bus j in (18.15a). Given matrices (v_j, ℓ_{jk}, S_{jk}), the vectors $(V_j, \tilde{I}_{jk})$, $j \in \overline{N}$, $j \to k \in E$, are determined uniquely up to a reference angle. These constraints are global.

The operational constraints on x are similar to those (18.4) in the bus injection model:

$$\text{injection limits:} \qquad s_j^{\min} \;\leq\; s_j \;\leq\; s_j^{\max}, \qquad j \in \overline{N} \tag{18.16a}$$

$$\text{voltage limits:} \qquad v_j^{\min} \;\leq\; \operatorname{diag}(v_j) \;\leq\; v_j^{\max}, \qquad j \in \overline{N} \tag{18.16b}$$

$$\text{line limits:} \qquad \operatorname{diag}(\ell_{jk}) \;\leq\; \ell_{jk}^{\max}, \qquad (j,k) \in E \tag{18.16c}$$

The constraint (18.16a) can be due to limits on the busbar to which the three-phase device is connected. All constraints in (18.16) are local at each bus j or on each line (j,k). While the voltage and line limits (18.4b)(18.4c) in BIM are generally nonconvex, these limits (18.16b)(18.16c) in BFM are linear in x.

Cost function.

Let $C(u,x)$ denote the cost function. For instance to minimize the thermal loss in the network we can use

$$C(u,x) \;:=\; \sum_{(j,k)\in E} \operatorname{diag}^{\mathsf{T}}\left(\operatorname{Re}(z_{jk})\right) \operatorname{diag}\left(\ell_{jk}\right)$$

OPF.

We assume $V_0 \in \mathbb{C}^3$ is given and impose $v_0 = V_0 V_0^{\mathsf{H}}$. Let the feasible set be

$$\mathbb{T}_{3p} \;:=\; \left\{(u,x) := (u,s,v,\ell,S,V,\tilde{I}) \mid (u,x) \text{ satisfies (18.1)(18.2)(18.15)(18.16)}, v_0 = V_0 V_0^{\mathsf{H}}\right\} \tag{18.17a}$$

Then the three-phase OPF problem is:

$$\min_{u,x} \; C(u,x) \;\text{ subject to }\; (u,x) \in \mathbb{T}_{3p} \tag{18.17b}$$

As for the bus injection model, the local constraint (18.2) describes the operational constraint on the internal variable u_j of the three-phase device j, the global constraint (18.15)(18.16) describes the network equation and operational constraint on the terminal variable $x := (V, s)$. The conversion rule (18.1) is local and relates u_j and x_j at each bus j. By Theorems 17.1, the feasible set $\mathbb{T}_{3p}$ in (18.17) is equivalent to the feasible set $\mathbb{V}_{3p}$ of the three-phase OPF (18.5) in BIM. Hence these problems are equivalent, provided their cost functions are the same. OPF (18.17) in the branch flow model can also be reformulated as QCQP using a similar method described in Chapter 18.1.3 for the bus injection model (see Chapter 18.3.2).

18.2 Semidefinite relaxation: BIM

Consider the three-phase OPF (18.5) in the bus injection model. In Chapter 18.2.1 we reformulate the constraints in (18.5) as semidefinite and rank constraints and in Chapter 18.2.2 we derive an SDP relaxation of three-phase OPF. Finally in Chapter 18.2.3 we show that if the three-phase network is radial then the relaxation is equivalent to a chordal relaxation because the single-phase equivalent of the network is a chordal graph.

18.2.1 Reformulation

The conversion rule (18.1) and the local operational constraint (18.2) in the device model are expressed in terms of the internal variable $u_j := V_j^{Y/\Delta}$ for a voltage source and $u_j := \left(s_j^{Y/\Delta}, I_j^{Y/\Delta}\right)$ for a power source. The network equation and constraint (18.3)(18.4) are expressed in terms of the terminal variable $x := (V, s)$. We will reformulate these constraints as semidefinite and rank constraints using a different set of variables; see Table 18.1. We first reformulate the network equation and constraint (18.3)(18.4) and then reformulate the device model (18.1)(18.2).

Network equations and constraints.

The power flow equations (18.3) are reproduced here:

$$s_j \;=\; \sum_{k:j\sim k} \operatorname{diag}\left(V_j (V_j - V_k)^{\mathsf{H}} \left(y_{jk}^{s}\right)^{\mathsf{H}} \;+\; V_j V_j^{\mathsf{H}} \left(y_{jk}^{m}\right)^{\mathsf{H}}\right), \qquad j \in \overline{N} \tag{18.18}$$

OPF	int vars: voltage	int vars: power	dev model	terminal vars	net model
(18.5)	$V_j^{Y/\Delta} \in \mathbb{C}^3$	$\left(s_j^{Y/\Delta}, I_j^{Y/\Delta}\right) \in \mathbb{C}^6$	(18.1)(18.2)	$(V,s) \in \mathbb{C}^{6(N+1)}$	(18.3)(18.4)
(18.23)	$W_j^{Y/\Delta} \in \mathbb{C}^{3\times 3}$	$s_j^{Y/\Delta} \in \mathbb{C}^3$ $X_j^{\Delta} \in \mathbb{C}^{3\times 3}$ $\ell_j^{\Delta} \in \mathbb{C}^{3\times 3}$	(18.22)	$W \in \mathbb{C}^{3(N+1)\times 3(N+1)}$ $s \in \mathbb{C}^{3(N+1)}$	(18.20)(18.21

Table 18.1 Internal and terminal variables of voltage and power sources for OPF (18.5) and its equivalent semidefinite reformulation (18.23).

Consider the $3(N+1)\times 3(N+1)$ matrix $W = VV^{\mathsf{H}}$ and its 3×3 submatrices W_{jj} and W_{jk} defined by:

$$W_{jj} \;=\; V_j V_j^{\mathsf{H}}, \qquad j \in \overline{N}, \qquad\qquad W_{jk} \;=\; V_j V_k^{\mathsf{H}}, \qquad (j,k) \in E \tag{18.19}$$

Then (18.18) is equivalent to the following equation that is linear in W:

$$s_j \;=\; \sum_{k:j\sim k} \operatorname{diag}\left(\left(W_{jj} - W_{jk}\right) \left(y_{jk}^s\right)^{\mathsf{H}} \;+\; W_{jj} \left(y_{jk}^m\right)^{\mathsf{H}} \right), \qquad j \in \overline{N} \tag{18.20a}$$

The network constraints (18.4) can be expressed also as linear functions of (s, W):

$$\begin{aligned}
&\text{injection limits:} && s_j^{\min} \;\le\; s_j \;\le\; s_j^{\max}, && j \in \overline{N} && (18.20b)\\
&\text{voltage limits:} && v_j^{\min} \;\le\; \operatorname{diag}\left(W_{jj}\right) \;\le\; v_j^{\max}, && j \in \overline{N} && (18.20c)\\
&\text{line limits:} && \operatorname{diag}\left(\ell_{jk}\left(W_{jj}, W_{jk}, W_{kk}\right)\right) \;\le\; \ell_{jk}^{\max}, && (j,k) \in E && (18.20d)\\
& && \operatorname{diag}\left(\ell_{kj}\left(W_{jj}, W_{kj}, W_{kk}\right)\right) \;\le\; \ell_{kj}^{\max}, && (j,k) \in E && (18.20e)
\end{aligned}$$

where, motivated by $I_{jk}(V) = \left(y_{jk}^s + y_{jk}^m\right) V_j - y_{jk}^s V_k$ and $I_{kj}(V) = \left(y_{kj}^s + y_{kj}^m\right) V_k - y_{kj}^s V_j$, we define the 3×3 matrices:

$$\begin{aligned}
\ell_{jk}\left(W_{jj}, W_{jk}, W_{kk}\right) \;&:=\; \left(y_{jk}^s + y_{jk}^m\right) W_{jj} \left(y_{jk}^s + y_{jk}^m\right)^{\mathsf{H}} - 2\mathrm{Re}\left(\left(y_{jk}^s + y_{jk}^m\right) W_{jk}\, y_{jk}^{s\mathsf{H}}\right) + y_{jk}^s W_{kk}\, y_{jk}^{s\mathsf{H}}\\
\ell_{kj}\left(W_{jj}, W_{kj}, W_{kk}\right) \;&:=\; \left(y_{kj}^s + y_{kj}^m\right) W_{kk} \left(y_{kj}^s + y_{kj}^m\right)^{\mathsf{H}} - 2\mathrm{Re}\left(\left(y_{kj}^s + y_{kj}^m\right) W_{kj}\, y_{kj}^{s\mathsf{H}}\right) + y_{kj}^s W_{jj}\, y_{kj}^{s\mathsf{H}}
\end{aligned}$$

Here the lower and upper bounds in (18.20b) – (18.20e) are 3-dimensional complex or real vectors. Instead of the quadratic equations (18.19) we use the following equivalent specification that is easy to convexify:

$$W \;\succeq\; 0, \qquad \operatorname{rank}(W) \;=\; 1 \tag{18.21}$$

Therefore the power flow equations and constraints (18.3)(18.4) are equivalent to the linear constraints (18.20) and the convex and nonconvex constraints in (18.21). These constraints are global. The semidefinite relaxation of the three-phase OPF (18.5) is obtained by omitting the nonconvex rank-1 constraint in (18.21).

Conversion rules and device constraints.

We apply the same method to reformulate the device models (18.1)(18.2). To simplify notation we assume:

- Only three-phase voltage and power sources are included, in Y or Δ configurations.
- The neutrals of all Y-configured devices are directly grounded and all voltages are defined with respect to the ground, so that all neutral voltages $\gamma_j^Y := V_j^n = 0$.

The conversion rules (18.1) are:

1 *Voltage source* $V_j^{Y/\Delta} \in \mathbb{C}^3$:

$$\begin{aligned} Y \text{ configuration:} \quad & V_j = V_j^Y \\ \Delta \text{ configuration:} \quad & \Gamma V_j = V_j^\Delta \end{aligned}$$

We reformulate this using a matrix variable $u_j := W_j^{Y/\Delta} \in \mathbb{C}^{3\times 3}$, as follows:

$$Y \text{ configuration:} \quad W_{jj} = W_j^Y, \quad W_j^Y \succeq 0, \quad \operatorname{rank}\left(W_j^Y\right) = 1 \tag{18.22a}$$

$$\Delta \text{ configuration:} \quad \Gamma W_{jj} \Gamma^{\mathsf{T}} = W_j^\Delta, \quad W_j^\Delta \succeq 0, \quad \operatorname{rank}\left(W_j^\Delta\right) = 1 \tag{18.22b}$$

Note that W_{jj} is the 3×3 principal submatrix of the $3(N+1)\times 3(N+1)$ matrix W defined in (18.21) associated with the vector V of terminal voltages while $W_j^{Y/\Delta}$ is a 3×3 matrix associated with the internal voltage $V_j^{Y/\Delta}$ of device j. The conditions (18.22a)(18.22b) ensure that there exists an internal voltage $V_j^{Y/\Delta}$, unique up to a rotation, so that $W_j^{Y/\Delta} = V_j^{Y/\Delta}\left(V_j^{Y/\Delta}\right)^{\mathsf{H}}$.

The device constraints (18.2a) on the internal voltage magnitudes can be expressed as a linear function of the internal variable $u_j := W_j^{Y/\Delta}$:

$$v_j^{Y/\Delta\min} \le \operatorname{diag}(u_j) := \operatorname{diag}\left(W_j^{Y/\Delta}\right) \le v_j^{Y/\Delta\max} \tag{18.22c}$$

where the lower and upper bounds $\left(v_j^{Y/\Delta\min}, v_j^{Y/\Delta\max}\right) \in \mathbb{C}^6$ are given vectors.

2 *Power source* $\left(s_j^{Y/\Delta}, I_j^{Y/\Delta}\right) \in \mathbb{C}^6$:

$$\begin{aligned} Y \text{ configuration:} \quad & s_j = -\operatorname{diag}\left(V_j I_j^{Y\mathsf{H}}\right), \qquad s_j = -s_j^Y \\ \Delta \text{ configuration:} \quad & s_j = -\operatorname{diag}\left(V_j I_j^{\Delta\mathsf{H}} \Gamma\right), \qquad s_j^\Delta = \operatorname{diag}\left(\Gamma V_j I_j^{\Delta\mathsf{H}}\right) \end{aligned}$$

We reformulate this using an internal variable $u_j := \left(s_j^{Y/\Delta}, X_j^\Delta, \ell_j^\Delta\right)$ where $s_j^{Y/\Delta} \in \mathbb{C}^3$

is the vector of terminal power injections and $X_j^\Delta, \ell_j^\Delta$ are 3×3 matrices for a Δ-configured power source, as follows:

$$Y \text{ configuration:} \quad s_j = -s_j^Y \tag{18.22d}$$

$$\Delta \text{ configuration:} \quad s_j = -\mathrm{diag}\left(X_j^\Delta \Gamma\right), \quad s_j^\Delta = \mathrm{diag}\left(\Gamma X_j^\Delta\right) \tag{18.22e}$$

$$0 \preceq \begin{bmatrix} W_{jj} & X_j^\Delta \\ X_j^{\Delta\mathsf{H}} & \ell_j^\Delta \end{bmatrix}, \quad 1 = \mathrm{rank} \begin{bmatrix} W_{jj} & X_j^\Delta \\ X_j^{\Delta\mathsf{H}} & \ell_j^\Delta \end{bmatrix} \tag{18.22f}$$

For a Δ-configured power source, the conditions (18.22e)(18.22f) ensure that there exist a terminal voltage V_j and an internal current I_j^Δ so that $W_{jj} = V_j V_j^\mathsf{H}$, $\ell_j^\Delta = I_j^\Delta I_j^{\Delta\mathsf{H}}$, and $X_j^\Delta = V_j I_j^{\Delta\mathsf{H}}$.

The device constraints (18.2c) on the internal powers and currents can be expressed as linear functions of the internal variable $u_j := \left(s_j^{Y/\Delta}, X_j^\Delta, \ell_j^\Delta\right)$:

$$s_j^{Y/\Delta\min} \le s_j^{Y/\Delta} \le s_j^{Y/\Delta\max}, \qquad \mathrm{diag}\left(\ell_j^\Delta\right) \le \ell_j^{\Delta\max} \tag{18.22g}$$

where the lower and upper bounds are given vectors.

Therefore the conversion rule (18.1) and the device constraint (18.2) are equivalent to the constraint (18.22) in terms of the new set of internal variables u_j and terminal variables (W, s), as summarized in Table 18.1. These constraints are local at each bus j. The rank-1 constraints in (18.22a)(18.22b)(18.22f) are nonconvex and the other constraints are convex (or linear). These rank-1 constraints will be omitted to derive a SDP relaxation of the three-phase OPF (18.5).

Equivalent OPF.

In summary, let $s \in \mathbb{C}^{N+1}$ denote the terminal power injections and $W \in \mathbb{C}^{3(N+1)\times 3(N+1)}$ denote the terminal variable associated with terminal voltages. Let $u := (u_j, j \in \overline{N})$ denote the internal variables defined by

$$u_j := \begin{cases} W_j^{Y/\Delta} & \text{if} \quad \text{device } j \text{ is a voltage source} \\ \left(s_j^{Y/\Delta}, X_j^\Delta, \ell_j^\Delta\right) & \text{if} \quad \text{device } j \text{ is a power source} \end{cases}$$

Finally we assume the terminal voltage V_0 at bus 0 is given and imposes the constraint $W_{00} = V_0 V_0^\mathsf{H}$. Putting all this together the three-phase OPF (18.5) is equivalent to

$$\min_{(u,s,W)} \quad C(u,s,W) \quad \text{s.t.} \quad W_{00} = V_0 V_0^\mathsf{H}, \ (18.20)(18.21)(18.22) \tag{18.23}$$

where $V_0 \in \mathbb{C}^3$ is given.

18.2.2 SDP relaxation

Define the matrix $M(A,B,D) \in \mathbb{C}^{6\times 3}$ as a function of 3×3 Hermitian matrices A, D, and a 3×3 arbitrary matrix B:

$$M(A,B,D) := \begin{bmatrix} A & B \\ B^{\mathsf{H}} & D \end{bmatrix} \tag{18.24}$$

Then $M(A,B,D)$ is Hermitian. For instance the matrix in (18.22f) is $M\left(W_{jj}, X_j^{\Delta}, \ell_j^{\Delta}\right)$.

Let N_v^Y and N_v^{Δ} denote the set of voltage sources in Y and Δ configuration respectively, and N_p^Y and N_p^{Δ} the set of power sources in Y and Δ configuration respectively. Omitting the rank-1 constraints in (18.22a)(18.22b)(18.22f) yields an SDP relaxation of (18.23):

$$\min_{(u,s,W)} \quad C(u,s,W) \tag{18.25a}$$

$$\text{s.t.} \quad W_{00} = V_0 V_0^{\mathsf{H}}, \ (18.20), \quad W \succeq 0 \tag{18.25b}$$

$$W_{jj} = W_j^Y, \quad W_j^Y \succeq 0, \quad j \in N_v^Y \tag{18.25c}$$

$$\Gamma W_{jj} \Gamma^{\mathsf{T}} = W_j^{\Delta}, \quad W_j^{\Delta} \succeq 0, \quad j \in N_v^{\Delta} \tag{18.25d}$$

$$s_j = -s_j^Y, \quad j \in N_p^Y \tag{18.25e}$$

$$s_j = -\mathrm{diag}\left(X_j^{\Delta}\Gamma\right), \quad j \in N_p^{\Delta} \tag{18.25f}$$

$$s_j^{\Delta} = \mathrm{diag}\left(\Gamma X_j^{\Delta}\right), \quad M\left(W_{jj}, X_j^{\Delta}, \ell_j^{\Delta}\right) \succeq 0, \quad j \in N_p^{\Delta} \tag{18.25g}$$

where $V_0 \in \mathbb{C}^3$ is given and $M\left(W_{jj}, X_j^{\Delta}, \ell_j^{\Delta}\right)$ is defined in (18.24). Let $\left(u^{\mathrm{opt}}, s^{\mathrm{opt}}, W^{\mathrm{opt}}\right)$ denote an optimal solution of the SDP relaxation (18.25). We say (18.25) is *exact* if the psd matrices of *every* optimal solution $\left(u^{\mathrm{opt}}, s^{\mathrm{opt}}, W^{\mathrm{opt}}\right)$ are of rank 1, i.e., $\mathrm{rank}\left(W^{\mathrm{opt}}\right) = 1$ and

$$\mathrm{rank}\left(W_j^{Y\,\mathrm{opt}}\right) = 1, \quad \mathrm{rank}\left(W_j^{\Delta\mathrm{opt}}\right) = 1, \quad \mathrm{rank}\left(M\left(W_{jj}^{\mathrm{opt}}, X_j^{\Delta\mathrm{opt}}, \ell_j^{\Delta\mathrm{opt}}\right)\right) = 1 \tag{18.26}$$

If $\mathrm{rank}\left(W^{\mathrm{opt}}\right) = 1$ then all its principal submatrices W_{jj}^{opt} are of rank 1 and therefore, by (18.25c)(18.25d), $W_j^{Y\,\mathrm{opt}}$ and $W_j^{\Delta\mathrm{opt}}$ are of rank 1 as well. The following result implies that the network matrix W^{opt} being psd rank-1 is insufficient to ensure exact relaxation. It is necessary for exact relaxation that all Δ-configured power sources must satisfy $\mathrm{rank}\left(M\left(W_{jj}^{\mathrm{opt}}, X_j^{\Delta\mathrm{opt}}, \ell_j^{\Delta\mathrm{opt}}\right)\right) = 1$.

Lemma 18.1 ([136])**.** Suppose the matrix $M(A,B,D) \in \mathbb{C}^{6\times 3}$ defined in (18.24) is positive semidefinite and $\operatorname{rank}(A) = 1$. Then A is psd rank-1, B is rank-1, and D is psd.

Lemma 18.1 says that B is rank-1 but may not be psd, and D is psd but may not be rank-1. Indeed it implies that the matrix $M(A,B,D) \succeq 0$ takes the form

$$M(A,B,D) \;=\; \begin{bmatrix} x \\ z \end{bmatrix} \begin{bmatrix} x^{\mathsf{H}} & z^{\mathsf{H}} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & KK^{\mathsf{H}} \end{bmatrix} \tag{18.27}$$

for some vectors x, z and matrix K. The structure (18.27) have three implications on the exactness of SDP relaxation (18.25). First if there are no Δ-configured power sources, then (18.25) is exact if, for every optimal solution $\left(u^{\text{opt}}, s^{\text{opt}}, W^{\text{opt}}\right)$ of (18.25), the network matrix W^{opt} is of rank 1. Second if there are Δ-configured power sources in N_p^{Δ}, then $\operatorname{rank}\left(W^{\text{opt}}\right) = 1$ is insufficient to guarantee exactness because the last condition in (18.26) may not be satisfied. Third, however, any optimal optimal solution $\left(u^{\text{opt}}, s^{\text{opt}}, W^{\text{opt}}\right)$ with $\operatorname{rank}\left(W^{\text{opt}}\right) = 1$ is sufficient for recovering an optimal solution of OPF (18.23), even if $\ell_j^{\Delta\text{opt}}$ in u_j^{opt} and hence $M\left(W_{jj}^{\text{opt}}, X_j^{\Delta\text{opt}}, \ell_j^{\Delta\text{opt}}\right)$ may not be of rank 1, provided that the cost $C(u,s,W)$ does not depend on ℓ_j^{Δ} (e.g., C depends only on $\left(s_j, s_j^{Y/\Delta}\right)$) [136, Theorem 1]. This is because Lemma 18.1 guarantees that there exists vectors $\left(V_j^{\text{opt}}, I_j^{\Delta\text{opt}}\right) \in \mathbb{C}^6$ such that, since W^{opt} is psd rank-1 and $M\left(W_{jj}^{\text{opt}}, X_j^{\Delta\text{opt}}, \ell_j^{\Delta\text{opt}}\right) \succeq 0$,

$$W_{jj}^{\text{opt}} \;=\; V_j^{\text{opt}} \left(V_j^{\text{opt}}\right)^{\mathsf{H}}, \qquad X_j^{\Delta\text{opt}} \;=\; V_j^{\text{opt}} \left(I_j^{\Delta\text{opt}}\right)^{\mathsf{H}}, \qquad j \in N_p^{\Delta} \tag{18.28a}$$

Then consider the point $\left(\tilde{u}, s^{\text{opt}}, W^{\text{opt}}\right)$ obtained from $\left(u^{\text{opt}}, s^{\text{opt}}, W^{\text{opt}}\right)$ by replacing $\ell_j^{\Delta\text{opt}}$ in u_j^{opt} by

$$\tilde{\ell}_j^{\Delta} \;:=\; I_j^{\Delta\text{opt}} \left(I_j^{\Delta\text{opt}}\right)^{\mathsf{H}}, \qquad j \in N_p^{\Delta} \tag{18.28b}$$

It can then be checked that $\left(\tilde{u}, s^{\text{opt}}, W^{\text{opt}}\right)$ is feasible for OPF (18.23). Since the cost C is independent of $\tilde{\ell}_j^{\Delta}$, $\left(\tilde{u}, s^{\text{opt}}, W^{\text{opt}}\right)$ is also optimal for OPF (18.23).

Remark 18.3 (Strong exactness)**.** As discussed in Remarks 10.3 and 10.4, even when a relaxation is not exact under our definition, an optimal solution of the original OPF problem may still be recoverable from an optimal solution of its relaxation under certain conditions. Theorems 10.6 and 10.9 provide two such conditions for single-phase radial network. The discussion above shows that $\operatorname{rank}\left(W^{\text{opt}}\right) = 1$ is sufficient for recovering an optimal solution of the original three-phase OPF (18.23) from an optimal solution of its SDP relaxation (18.25), provided that the cost C is independent of ℓ_j^{Δ} for Δ-configured power sources. □

The method (18.28) to recover an optimal solution $\left(\tilde{u}, s^{\text{opt}}, W^{\text{opt}}\right)$ of OPF (18.23) from an optimal solution of its relaxation may not work well in practice because of inevitable numerical errors. Even if W_{jj}^{opt} is close to being rank-1, i.e., its second largest

eigenvalue is several orders of magnitude smaller than its largest eigenvalue, X_j^Δ can be far from being rank-1, e.g., it can have multiple large eigenvalues of the same magnitude (see [136, Remark 1]). In this case $I_j^{\Delta\text{opt}}$ may not be obtained from $X_j^{\Delta\text{opt}}$ using (18.28a). Two methods are suggested in [136] to address this numerical issue. The first method substitutes V_j^{opt} obtained from $W_{jj}^{\text{opt}} = V_j^{\text{opt}} \left(V_j^{\text{opt}}\right)^{\mathsf{H}}$ into (18.25g):

$$s_j^{\Delta\text{opt}} = \operatorname{diag}\left(\left(\Gamma V_j^{\Delta\text{opt}}\right)\left(I_j^{\Delta\text{opt}}\right)^{\mathsf{H}}\right) \quad \Longrightarrow \quad I_j^{\Delta\text{opt}} := \left(\operatorname{diag}\left(\Gamma \bar{V}_j^{\Delta\text{opt}}\right)\right)^{-1} \bar{s}_j^{\Delta\text{opt}}$$

where $\bar{x}$ is the componentwise complex conjugate of a vector x. The second method adds $\lambda \sum_j \operatorname{tr}\left(\ell_j^\Delta\right)$ to the cost function of the SDP relaxation (18.25) for a positive but small weight $\lambda > 0$. This produces an optimal solution in which $\ell_j^{\Delta\text{opt}}$ tends to be of low rank.

18.2.3 Radial network

A special case that is particularly simple is a network where

- all three-phase devices are either voltage or power sources in Y configuration;
- all voltages are defined with respect to the ground and the neutral voltages $\gamma_j^Y := V_j^n$ of all these Y-configured devices are $\gamma_j^Y := 0$.

In this case the internal variables can be simply expressed in terms of terminal variables, $V_j^Y = V_j$, $I_j^Y = -I_j$, and $s_j^Y = -s_j$, and the operational constraints $g_j^{Y/\Delta}(u_j) \le 0$ on u_j are included in the network constraints (18.10). Hence the internal variable u can be eliminated from the QCQP (18.14) which then consists of only network constraints (18.10) and no device models, as follows:

$$\min_V \quad C(V, s(V)) \qquad \text{s.t.} \qquad (18.10) \tag{18.29}$$

We now study the semidefinite relaxation of (18.29) when the network graph G is a tree.

Consider a network graph $G := (\overline{N}, E)$ with $N+1$ buses. Suppose each line $(j,k) \in E$ is characterized by three 3×3 admittance matrices $\left(y_{jk}^s, y_{jk}^m, y_{kj}^m\right)$, i.e., we assume $y_{jk}^s = y_{kj}^s$. Recall its single-phase equivalent circuit described in Chapter 16.1.2 by a graph $G^{3\phi} := \left(\overline{N}^{3\phi}, E^{3\phi}\right)$ where $\overline{N}^{3\phi}$ contains $3(N+1)$ nodes identified by $j\phi$, $j \in \overline{N}$, $\phi \in \{a,b,c\}$. There is a link $(j\phi, k\phi')$ in $E^{3\phi}$ if and only if the $(j\phi, k\phi')$ entry $Y_{jk}^{\phi\phi'}$ of the three-phase admittance matrix Y is nonzero.

Even when G is a tree (i.e., the three-phase network is radial), its single-phase equivalent $G^{3\phi}$ contains cycles. The key observation is that $G^{3\phi}$ is a chordal graph. To see this, note that $G^{3\phi}$ has a maximal clique with 6 nodes consisting of the set

$\left\{ j\phi, k\phi' \in \overline{N}^{3\phi} : \phi, \phi' \in \{a,b,c\} \right\}$ of buses if and only if (j,k) is a line in G. See Figure 18.1 for an example. Two nodes $j\phi$ and $k\phi'$ in the equivalent circuit $G^{3\phi}$ are

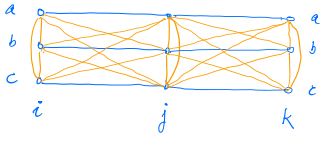


Figure 18.1 The graph $G^{3\phi}$ of the single-phase equivalent circuit of a radial network with three buses i, j, k connected by (three-wire) three-phase lines.

adjacent either because of a physical line between buses j and k in the graph G (in which case $\phi = \phi'$) or because of electromagnetic interactions across phases ϕ and ϕ' (in which case $\phi \neq \phi'$). Indeed $G^{3\phi}$ consists of a macro tree in which every link in the macro tree is such a clique and these are the only cliques in $G^{3\phi}$. This means that $G^{3\phi}$ is a chordal graph.

Theorem 10.4 suggests solving the chordal relaxation of (18.29). It computes a $(N+1)\times(N+1)$ Hermitian partial matrix $W_{G^{3\phi}}$:

$$W_{G^{3\phi}} := \left(\, [W_{G^{3\phi}}]_{jj}^{\phi\phi}, j\phi \in \overline{N}^{3\phi}, \;\; [W_{G^{3\phi}}]_{jk}^{\phi\phi'}, (j\phi, k\phi') \in E^{3\phi} \, \right)$$

The set of maximal cliques of $G^{3\phi}$ correspond to the following 6×6 principal submatrices of $W_{G^{3\phi}}$:

$$W_{G^{3\phi}}(j,k) = \begin{bmatrix} w_{jj} & w_{jk} \\ w_{kj} & w_{kk} \end{bmatrix} \in \mathbb{C}^{6\times 6}, \quad (j,k) \in E$$

where

$$w_{jj} := \begin{bmatrix} [W_{G^{3\phi}}]_{jj}^{aa} & [W_{G^{3\phi}}]_{jj}^{ab} & [W_{G^{3\phi}}]_{jj}^{ac} \\ [W_{G^{3\phi}}]_{jj}^{ba} & [W_{G^{3\phi}}]_{jj}^{bb} & [W_{G^{3\phi}}]_{jj}^{bc} \\ [W_{G^{3\phi}}]_{jj}^{ca} & [W_{G^{3\phi}}]_{jj}^{cb} & [W_{G^{3\phi}}]_{jj}^{cc} \end{bmatrix}, \quad w_{jk} := \begin{bmatrix} [W_{G^{3\phi}}]_{jk}^{aa} & [W_{G^{3\phi}}]_{jk}^{ab} & [W_{G^{3\phi}}]_{jk}^{ac} \\ [W_{G^{3\phi}}]_{jk}^{ba} & [W_{G^{3\phi}}]_{jk}^{bb} & [W_{G^{3\phi}}]_{jk}^{bc} \\ [W_{G^{3\phi}}]_{jk}^{ca} & [W_{G^{3\phi}}]_{jk}^{cb} & [W_{G^{3\phi}}]_{jk}^{cc} \end{bmatrix}$$

The chordal relaxation of (18.29) is then (using (18.10)):

$$\min_{W_{G^{3\phi}}} \quad \mathrm{tr}\,(C_0 W_{G^{3\phi}}) \tag{18.30a}$$

$$\text{s.t.} \quad p_j^{\phi\,\min} \le \mathrm{tr}\left(\Phi_j^{\phi} W_{G^{3\phi}}\right) \le p_j^{\phi\,\max}, \qquad j \in \overline{N},\ \phi \in \{a,b,c\} \tag{18.30b}$$

$$q_j^{\phi\,\min} \le \mathrm{tr}\left(\Psi_j^{\phi} W_{G^{3\phi}}\right) \le q_j^{\phi\,\max}, \qquad j \in \overline{N},\ \phi \in \{a,b,c\} \tag{18.30c}$$

$$v_j^{\phi\,\min} \le \mathrm{tr}\left(E_j^{\phi} W_{G^{3\phi}}\right) \le v_j^{\phi\,\max}, \qquad j \in \overline{N},\ \phi \in \{a,b,c\} \tag{18.30d}$$

$$\mathrm{tr}\left(\hat{Y}_{jk}^{\phi} W_{G^{3\phi}}\right) \le \ell_{jk}^{\phi\,\max}, \qquad (j,k) \in E,\ \phi \in \{a,b,c\} \tag{18.30e}$$

$$\mathrm{tr}\left(\hat{Y}_{kj}^{\phi} W_{G^{3\phi}}\right) \le \ell_{kj}^{\phi\,\max}, \qquad (j,k) \in E,\ \phi \in \{a,b,c\} \tag{18.30f}$$

$$W_{G^{3\phi}}(j,k) \succeq 0, \qquad (j,k) \in E \tag{18.30g}$$

$$w_{00} = V_0 V_0^{\mathsf{H}} \qquad (V_0 \text{ is given}) \tag{18.30h}$$

Let $W_{G^{3\phi}}^{\mathrm{opt}}$ be an optimal solution of (18.30). If every 6×6 principal submatrix $W_{G^{3\phi}}^{\mathrm{opt}}(j,k)$ of the partial matrix $W_{G^{3\phi}}^{\mathrm{opt}}$ satisfies

$$\mathrm{rank}\left(W_{G^{3\phi}}^{\mathrm{opt}}(j,k)\right) = 1, \qquad (j,k) \in E$$

then an optimal solution V^{opt} of (18.29) can be uniquely recovered from $W_{G^{3\phi}}^{\mathrm{opt}}$ according to Theorem 10.3. This is because a chordal relaxation is exact if and only if the principal submatrix $W_{G^{3\phi}}^{\mathrm{opt}}(q)$ of $W_{G^{3\phi}}^{\mathrm{opt}}$ is psd rank-1 for every clique q of the chordal graph $G^{3\phi}$ (Theorem 10.1) and, as noted above, the only maximal cliques of $G^{3\phi}$ are those 6-node cliques corresponding to lines $(j,k) \in E$.

The method in Chapter 10.1.4 to recover an optimal V^{opt} from $W_{G^{3\phi}}^{\mathrm{opt}}$ applies directly here. Since $\mathrm{rank}\left(W_{G^{3\phi}}^{\mathrm{opt}}(j,k)\right) = 1$ for all $(j,k) \in E$, they satisfy the cycle condition (Theorem 10.1). Take any spanning tree of $G^{3\phi}$ with root at, say, node $0a$. Let $\left|V_j^{\phi}\right| := \sqrt{[W_{G^{3\phi}}^{\mathrm{opt}}]_{jj}^{\phi\phi}}$ for $j \in N, \phi \in \{a,b,c\}$. Let P_j^{ϕ} be the unique path from the root $0a$ to the node $j\phi$ in the spanning tree. A link $(j'\phi', j''\phi'')$ in the path P_j^{ϕ} is denoted by $(j'\phi', j''\phi'') \in \mathsf{P}_j^{\phi}$. Then for all nodes $j\phi$ in the equivalent single-phase network $G^{3\phi}$,

$$\angle V_j^{\phi} := \angle V_0^{a} \; - \sum_{(j'\phi', j''\phi'') \in \mathsf{P}_j^{\phi}} \angle \left[W_{G^{3\phi}}^{\mathrm{opt}}\right]_{j'j''}^{\phi'\phi''} \mod 2\pi$$

18.3 Semidefinite relaxation: BFM

As for the bus injection model we reformulate in Chapter 18.3.1 the three-phase OPF (18.17) in the branch flow model for radial networks, and derive in Chapter 18.3.2 its semidefinite relaxation.

18.3.1 Reformulation

Consider the three-phase OPF (18.17) in BFM for radial networks studied in Chapter 18.1.4, reproduced here:

$$\min_{(u,x)} \; C(u,x) \;\; \text{s. t. } (18.1)(18.2)(18.15)(18.16),\; v_0 = V_0 V_0^{\mathsf{H}} \tag{18.31}$$

where $(u,x) := (u,s,v,\ell,S,V,\tilde{I})$, u denotes the internal variables of three-phase devices and x denotes the terminal variables that interact through power flow equations. The devices are modeled by the conversion rules (18.1) on (u_j,x_j) and the operational constraints (18.2) on u_j. The power flow equation is (18.15) and the operational constraint on x is (18.16).

To simplify notation we consider, as in Chapter 18.2.1, only three-phase voltage and power sources and assume that all neutral voltages $\gamma_j^Y := V_j^n = 0$. Then the internal variables for these devices are $u := (u_j, j \in \overline{N})$ where

$$u_j \;:=\; \begin{cases} v_j^{Y/\Delta} & \text{if} \quad \text{device } j \text{ is a voltage source} \\ \left(s_j^{Y/\Delta}, X_j^{\Delta}, \ell_j^{\Delta}\right) & \text{if} \quad \text{device } j \text{ is a power source} \end{cases} \tag{18.32}$$

The device models (18.1)(18.2) have been reformulated as (18.22) in Chapter 18.2.1, with the 3×3 matrix variables W_{jj} and $W_j^{Y/\Delta}$ in BIM replaced by v_j and $v_j^{Y/\Delta}$ respectively in BFM.

Without voltage sources, we no longer need the variable V_j for the conversion rule that relates V_j to the internal voltage $V_j^{Y/\Delta}$. Hence we will omit $(V_j, \tilde{I}_{jk})$ and the quadratic constraints (18.15d), $v_j = V_j V_j^{\mathsf{H}}$, $\ell_{jk} = \tilde{I}_{jk}\tilde{I}_{jk}^{\mathsf{H}}$, and $S_{jk} = V_j \tilde{I}_{jk}^{\mathsf{H}}$. Let the BFM variables be $x := (s,v,\ell,S)$ where v_j, ℓ_{jk}, S_{jk} is each a 3×3 matrix. Finally we assume the terminal voltage V_0 at bus 0 is given and imposes the constraint $v_0 = V_0 V_0^{\mathsf{H}}$. Then the three-phase OPF (18.31) can be reformulated as follows. Let the feasible set be

$$\mathbb{T}_{3p} \;:=\; \left\{(u,x) := (u,s,v,\ell,S) \mid (u,x) \text{ satisfies } (18.15a)-(18.15c)(18.16)(18.22), v_0 = V_0 V_0^{\mathsf{H}}\right\} \tag{18.33a}$$

where u is defined in (18.32). The three-phase OPF problem (18.31) is equivalent to:

$$\min_{u,x} \; C(u,x) \;\text{ subject to }\; (u,x) \in \mathbb{T}_{3p} \tag{18.33b}$$

18.3.2 Semidefinite relaxation

OPF (18.33) is nonconvex due to the rank-1 constraint (18.15c) in the power flow equations and the rank-1 constraints (18.22a)(18.22b)(18.22f) in the device models. Omitting these rank-1 constraints yields a semidefinite relaxation. Recall the function

$M(A,B,D)$ that constructs a 6×6 matrix from 3×3 matrices A,B,D, defined in (18.24) and reproduced here:

$$M(A,B,D) \;:=\; \begin{bmatrix} A & B \\ B^{\mathsf{H}} & D \end{bmatrix} \tag{18.34}$$

where A,D are Hermitian and B is arbitrary. Then the psd constraints in (18.15c) and in (18.22f) can be written in terms of M as respectively.

$$M\left(v_j, S_{jk}, \ell_{jk}\right) \;=\; \begin{bmatrix} v_j & S_{jk} \\ S_{jk}^{\mathsf{H}} & \ell_{jk} \end{bmatrix} \;\succeq\; 0, \qquad j \to k \in E$$

$$M\left(v_j, X_{jk}^{\Delta}, \ell_{jk}^{\Delta}\right) \;=\; \begin{bmatrix} v_j & X_{jk}^{\Delta} \\ X_{jk}^{\Delta\mathsf{H}} & \ell_{jk}^{\Delta} \end{bmatrix} \;\succeq\; 0, \qquad j \to k \in E$$

The feasible set of the semidefinite relaxation is defined by the following constraints:

$$\text{network:} \qquad v_0 \;=\; V_0 V_0^{\mathsf{H}}, \;\; (18.15a)(18.15b), (18.16), \tag{18.35a}$$

$$0 \;\preceq\; M\left(v_j, S_{jk}, \ell_{jk}\right), \quad (j,k) \in E \tag{18.35b}$$

$$\text{devices:} \qquad v_j \;=\; v_j^Y, \quad v_j^Y \;\succeq\; 0, \qquad j \in N_v^Y \tag{18.35c}$$

$$\Gamma v_j \Gamma^{\mathsf{T}} \;=\; v_j^{\Delta}, \quad v_j^{\Delta} \;\succeq\; 0, \qquad j \in N_v^{\Delta} \tag{18.35d}$$

$$s_j \;=\; -s_j^Y, \qquad j \in N_p^Y \tag{18.35e}$$

$$s_j \;=\; -\mathrm{diag}\left(X_j^{\Delta}\Gamma\right), \;\; s_j^{\Delta} \;=\; \mathrm{diag}\left(\Gamma X_j^{\Delta}\right), \;\; M\left(v_j, X_j^{\Delta}, \ell_j^{\Delta}\right) \;\succeq\; 0, \qquad j \in N_p^{\Delta} \tag{18.35f}$$

where $V_0 \in \mathbb{C}^3$ is given. Define the feasible set as

$$\mathbb{T}_{3p}^{+} \;:=\; \{(u,x) := (u,s,v,\ell,S) \mid (u,x) \text{ satisfies } (18.35)\} \tag{18.36a}$$

where u is defined in (18.32). The set $\mathbb{T}_{3p}^{+}$ is a convex superset of $\mathbb{T}_{3p}$. The semidefinite relaxation of the three-phase OPF problem (18.33) is:

$$\min_{u,x} \; C(u,x) \;\text{ subject to }\; (u,x) \in \mathbb{T}_{3p}^{+} \tag{18.36b}$$

Let $\left(u^{\mathrm{opt}}, x^{\mathrm{opt}}\right)$ denote an optimal solution of the SDP relaxation (18.36). We say (18.36) is *exact* if the psd matrices of *every* optimal solution $\left(u^{\mathrm{opt}}, x^{\mathrm{opt}}\right)$ are of rank 1, i.e.,

$$\mathrm{rank}\left(M\left(v_j^{\mathrm{opt}}, X_j^{\Delta\mathrm{opt}}, \ell_j^{\Delta\mathrm{opt}}\right)\right) \;=\; 1, \quad \mathrm{rank}\left(v_j^{Y/\Delta\mathrm{opt}}\right) \;=\; 1, \qquad j \in \overline{N} \tag{18.37a}$$

$$\mathrm{rank}\left(M\left(v_j^{\mathrm{opt}}, S_{jk}^{\mathrm{opt}}, \ell_{jk}^{\mathrm{opt}}\right)\right) \;=\; 1, \qquad (j,k) \in E \tag{18.37b}$$

This means that $\left(u^{\mathrm{opt}}, x^{\mathrm{opt}}\right)$ is feasible and therefore optimal for the original OPF (18.33).

Suppose the terminal voltage satisfies $\mathrm{rank}\left(v_j^{\mathrm{opt}}\right) = 1$. Then the internal voltage

$v_j^{Y/\Delta\text{opt}}$ is also of rank 1 by (18.35c)(18.35d). Unfortunately $M\left(v_j^{\text{opt}}, X_j^{\Delta\text{opt}}, \ell_j^{\Delta\text{opt}}\right)$ and $M\left(v_j^{\text{opt}}, S_{jk}^{\text{opt}}, \ell_{jk}^{\text{opt}}\right)$ may not be of rank 1 because $\ell_j^{\Delta\text{opt}}$ and ℓ_{jk}^{opt} respectively may not be rank-1; see Lemma 18.1. As discussed after Lemma 18.1, even though the SDP relaxation (18.36) may not be exact, it is still possible to recover an optimal solution of OPF (18.33) from an optimal solution $\left(u^{\text{opt}}, x^{\text{opt}}\right)$ of its relaxation (18.36) when $\text{rank}\left(v_j^{\text{opt}}\right) = 1$ for all $j \in \overline{N}$, provided that the cost function C is independent of ℓ_j^{Δ}.

Equivalence.

When the network graph is a tree, then it can be shown that OPF (18.33) and its relaxation (18.36) in BFM are equivalent to OPF (18.23) and its relaxation (18.25) respectively in BIM (see [136, Proposition 1]).

18.4 Example applications

18.5 Bibliographical notes

As for most chapters, this section is now a placeholder with references collected in a somewhat random fashion during the writing of the text. Major rewrite later.

There has been a great deal of research on OPF since Carpentier's first formulation in 1962 [86]. An early solution appears in [87] and extensive surveys can be found in e.g. [88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 39, 101]. It is nonconvex and has been shown to be NP-hard in general [102, 69, 70].

Many references for 3-phase OPF: e.g. [103, 104, 105]

There are many excellent texts on optimization theory especially for convex problems, e.g., [62, 57, 54]. Optimization texts with power system applications include [106, 107]. In particular Chapter 8.5.3 mostly follows the presentation in [57, Chapter 11]. A popular interior-point solver for OPF problems is [108].

A classic text on computational complexity is [75]. OPF has been shown to be NP-hard in general [102, 69, 70, 72, 74]. [77] surveys combinatorial OPF and proves approximation results and conditions for exactness (when there are no discrete variables). It shows that OPF with discrete injections cannot be efficiently approximated. The hardness results complement those in [73, 68, 69, 70]; see [77, Chapter 5] and its Section 5.6 for comparison.

Chapter **??** on global optimality is taken form [80, 79]

[109] shows that, by dualizing clique tree conversion, a class of nonconvex problems, including OPF problems, the per-iteration cost of an interior-point method is linear

$O(n)$ in time and in memory, so an ϵ-accurate and ϵ-feasible iterate is obtained after $O(\sqrt{n}\log(1/\epsilon))$ iterations in $O(n^{1.5}\log(1/\epsilon))$ time.

18.6 Problems

Chapter 18.1

Exercise 18.1 (3-phase OPF as QCQP: line limit)**.** Derive the line limit (18.10d) in three-phase OPF.

Exercise 18.2 (3-phase OPF as QCQP: current source)**.** Derive the conversion rule (18.13c)(18.13d) for a current source $u_j := I_j^{Y/\Delta}$.

Exercise 18.3 (3-phase OPF as QCQP: power source)**.** Derive the conversion rule (18.13e)(18.13f) for a power source $u_j := (u_{j1}, u_{j2}) := \left(s_j^{Y/\Delta}, I_j^{Y/\Delta}\right)$.

Exercise 18.4 (3-phase OPF as QCQP: impedance)**.** Derive the conversion rule (18.13g)(18.13h) for an impedance $\left(z_j^Y, \gamma_j^Y\right)$ or z_j^Δ.

Exercise 18.5 (3-phase OPF as QCQP: power source)**.** For a power source, we use $u_j := \left(s_j^{Y/\Delta}, I_j^{Y/\Delta}\right)$ as the internal variable. This exercise shows that this approach is simpler for a Y-configured power source than if the optimization variable is taken to be $u_j := s_j^Y$ instead. Consider a Y-configured ideal power source where the optimization variable is the internal power (only) $u_j := s_j^Y$ and its neutral voltage $\gamma_j^Y := V_j^n$ is given. If $\gamma_j^Y = 0$ then $s_j = -s_j^Y$. Suppose $\gamma_j^Y \neq 0$.

1 Show that u_j is related to the terminal voltage and current (V_j, s_j) as:

$$s_j = -\mathrm{diag}\left(\frac{V_j^\phi}{V_j^\phi - \gamma_j^Y}, \phi = a,b,c\right) u_j$$

2 *Y configuration*: Show that the conversion rule in part 1 is equivalent to the following set of inhomogeneous equality constraints on $\left(V, u_j, w_j^\phi, \phi \in \{a,b,c\}\right) \in \mathbb{C}^{12(N+1)+3}$: for each $j \in \overline{N}$,

$$\begin{aligned} V^{\mathsf{H}}\left(\gamma_j^Y Y_j^{\phi\mathsf{H}}\right)V &= \bar{u}_j^{\mathsf{H}}\left(e^\phi e_j^{\phi\mathsf{H}}\right)V + w_j^{\phi\mathsf{H}}\left(Y_j^{\phi\mathsf{H}}\right)V, && \phi \in \{a,b,c\} \\ e_k^{\varphi\mathsf{H}} w_j^\phi &= V^{\mathsf{H}}\left(e_j^\phi e_j^{\varphi\mathsf{H}}\right)V, && k \in \overline{N},\ \phi,\varphi \in \{a,b,c\} \end{aligned}$$

where $w_j^{\phi} \in \mathbb{C}^{3(N+1)}$ is an auxiliary variable, one for each $\phi \in \{a,b,c\}$. For each $j \in \overline{N}$, this is a set of $9(N+1)+3$ quadratic equations in $\left(V, u_j, w_j^{\phi}, \phi \in \{a,b,c\}\right)$.

Chapter 18.2

Exercise 18.6 (SDP relaxation in BIM [136]). 1 Prove Lemma 18.1.
2 Give an example where $M(A,B,D)$ is not of rank 1.

Appendix Linear algebra preliminaries

In this chapter we review some basic concepts in linear algebra and algebraic graph theory that we have used in this book. There are many excellent books on these topics and our goal is *not* to be comprehensive or systematic in coverage, but to collect concepts and properties used in this book in one place for the convenience of the readers who have already had exposures to these topics.

A.1 Vector spaces, basis, rank, nullity

A.1.1 Vector spaces, subspaces, span

This subsection mostly follows [15, Chapter 0]. We restrict ourselves mostly to finite vector spaces. Underlying a vector space is its *field* F, which is a set of *scalars* that is closed under two binary operations, called "addition" $(a+b)$ and "multiplication" (ab). Most often, $F = \mathbb{R}$ or $\mathbb{C}$ for us, but in general F can be the set of rational numbers, or a set of integers modulo a specified prime number, etc. The two operations must be associative and commutative, and each must have an identity element in the set; inverses must exist in the set for all elements under addition and for all elements except the additive identity under multiplication; multiplication must distribute over addition.

Definition A.1 (Vector space)**.** A *vector space* V, or *linear space*, over a field F is a set V of objects, called *vectors*, that is closed under two binary operations:

- *vector addition* $+ : V \times V \to V$ denoted by $x + y$;
- *scalar multiplication* $\cdot : F \times V \to V$ denoted by $a \cdot x =: ax$;

and satisfies the following properties: for all $x, y, z \in V$ and $a, b \in F$,

1 *Associativity of vector addition*: $x + (y + z) = (x + y) + z$.
2 *Commutativity of vector addition*: $x + y = y + x$.
3 *Identity element of vector addition*: There exists $0 \in V$, called the *zero vector*, such that $x + 0 = x$.

4 *Inverse elements of vector addition*: There exists $-x \in V$, called the *additive inverse* of x, such that $x + (-x) = 0$.
5 *Associativity of scalar multiplication*: $a(bx) = (ab)x$.
6 *Identity element of scalar multiplication*: There exists $1 \in F$, called the *multiplicative identity* in F such that $1x = x$.
7 *Distributivity of scalar multiplication over vector addition*: $a(x+y) = ax + by$.
8 *Distributivity of scalar multiplication over field addition*: $(a+b)x = ax + bx$.

A *subspace* of a vector space V over a field F is a subset of V that is itself a vector space over F with the same binary operations as in V. □

If $F = \mathbb{R}$ then V is called a *real vector space*. If $F = \mathbb{C}$ then V is called a *complex vector space*. Given F and an integer n the set $V := F^n$ of n-tuples with components from F forms a vector space over F where the vector addition "+" is defined by componentwise addition: $[x+y]_i = x_i + y_i$. The vector space F^n is important because any finite dimensional vector space can be identified with F^n for some integer n (see Example A.1 and the next subsection for a formal definition). Note that $\mathbb{R}^n$ is a real vector space ($V = \mathbb{R}^n$ over $F = \mathbb{R}$) while $\mathbb{C}^n$ is both a real vector space ($V = \mathbb{C}^n$ over $F = \mathbb{R}$) and a complex vector space ($V = \mathbb{C}^n$ over $F = \mathbb{C}$).

A vector space V is however not restricted to $V = F^n$. An important finite dimensional vector space over F is the set $M_{m,n}(F)$ of $m \times n$ matrices whose entries $[M]_{ij} \in F$ for any finite m and n. We can vectorize $A \in M_{m,n}(F)$ and treat A as a vector in $V = F^{mn}$, but we will mostly treat A as an array of scalars in $V = F^{m\times n}$. Note that matrix multiplication is not involved in the definition of $V = F^{m\times n}$ as a vector space (it can be treated as a composition of linear transformations when a matrix is viewed as a linear transformation from F^n to F^m; see below). If $m = n$ we abbreviate $M_{m,n}(F)$ to $M_m(F)$. If $F = C$ we abbreviate $M_{m,n}(C)$ to $M_{m,n}$.

The components x_i of vectors $x \in V$ may not be from F. Possibly infinite dimensional examples include: the set of polynomials with real or with complex coefficients (of up to a specified degree or of arbitrary degree) is a real or complex vector space respectively; the set of real-valued or complex-valued functions on subsets of $\mathbb{R}$ or $\mathbb{C}$ is a real or complex vector space respectively.

If $S \subseteq V$ is a nonempty subset of the vector space V over a field F then $\text{span}(S)$ is the intersection of all subspaces of V that contain S. It consists of all linear combinations of finitely many vectors in S:

$$\text{span}(S) \;=\; \{a_1x_1 + \cdots + a_kx_k \,:\, x_1,\ldots,x_k \in S,\; a_1,\ldots,a_k \in F,\; k = 1,2,\ldots\}$$

It can be checked that $\text{span}(S)$ is always a subspace whether or not S is a subspace. S is said to *span* V if $\text{span}(S) = V$. Let S_1 and S_2 be subspaces of a vector space over a field F. The *sum* of S_1 and S_2 is the subspace

$$S_1 + S_2 \;:=\; \text{span}\{S_1 \cup S_2\} \;=\; \{x+y \,:\, x \in S_1,\, y \in S_2\}$$

If $S_1 \cap S_2 = \{0\}$ then $S_1 + S_2$ is called a *direct sum* and we write it as $S_1 \oplus S_2$. Every vector $z \in S_1 \oplus S_2$ can be uniquely written as $z = x + y$ with $x \in S_1$ and $y \in S_2$.

Example A.1. Consider $S := \{1, t, t^2, \ldots, t^{n-1}\}$. Even though S is not a vector space its span

$$\text{span}(S) \;=\; \{a_0 + a_1 t + \cdots + a_{n-1} t^{n-1} \;:\; a_0, \ldots, a_{n-1} \in F\}$$

is an n-dimensional vector space V that can be identified with F^n where $x \in V$ is defined by $x_i = a_i$, $i = 0, \ldots, n-1$. □

A.1.2 Basis, dimension, rank and nullity

A finite set of vectors $x_1, \ldots, x_k$ in a vector space V over a field F is *linearly dependent* if and only if there are scalars $a_1, \ldots, a_k \in F$, not all zero, such that $a_1 x_1 + \cdots + a_k x_k = 0 \in V$. The vectors $x_1, \ldots, x_k$ are *linearly independent* if they are not linearly dependent. A linearly independent set $B := \{v_1, v_2, \ldots,\} \subseteq V$ of vectors that spans the vector space V is called a *basis*. Any vector $x \in V$ can be uniquely expressed as a linear combination of the basis, i.e., $x = \sum_i a_i v_i$ for a unique set of scalars $a_i \in F$, $i = 1, 2, \ldots$. If there is a positive integer n such that $B := \{v_1, \ldots, v_n\}$ is a basis of V, then all bases of V consist of exactly n vectors and n is the *dimension* of V, denoted by $\dim(V)$. This is because adding any vector to a basis will render it linearly dependent and removing any vector from the basis will prevent it from spanning V. In this case V is *finite dimensional*. If no such integer n exists then V is *infinite dimensional*. For an infinite dimensional vector space, there is a one-to-one correspondence between the vectors in any two bases. A subspace of a (finite) n-dimensional vector space has dimension no more than n; it is a proper subspace if its dimension is strictly less than n.

The real vector space $\mathbb{R}^n$ has dimension n. The complex vector space C^n has dimension n over the field $F = \mathbb{C}$ but dimension $2n$ over the field $F = \mathbb{R}$. A *basis* of a vector space F^n is a set of vectors $\{v_1, \cdots, v_n\}$ such that any vector $x \in F^n$ can be expressed as a linear combination of vectors in the basis, i.e., $x = B\alpha$ for some $\alpha \in F^n$ where the columns of B are the vectors $\{v_1, \cdots, v_n\}$. If the basis vectors are orthogonal, i.e., $v_i^{\mathsf{H}} v_j = 0$ for $i \neq j$, then the basis is called an *orthogonal basis*. If the basis vectors are both orthogonal and of unit Euclidean norm ($\|v_i\|_2 = 1$ for all i), then the basis is called an *orthonormal basis*. The basis $\{e_1, \cdots, e_n\}$ of F^n in which the n-vector e_i has a 1 in its ith entry and 0s elsewhere is called the *standard basis*, the *unit basis* or the *unit vector*. It is an orthonormal basis. Two vector spaces U and V over the same filed F is called *isomorphic* if there is an invertible function $f : U \to V$ such that $f(ax + by) = af(x) + bf(y)$ for all $x, y \in U$ and $a, b \in F$. Then f is called an *isomorphism*. Any n-dimensional real vector space is isomorphic to $\mathbb{R}^n$ and any n-dimensional complex vector space is isomorphic to $\mathbb{C}^n$.

Let V be a finite-dimensional vector space and let S_1, S_2 be two given subspaces of

V. Then

$$\dim(S_1 \cap S_2) + \dim(S_1 + S_2) = \dim(S_1) + \dim(S_2)$$

Hence

$$\dim(S_1 \cap S_2) \geq \dim(S_1) + \dim(S_2) - \dim(V)$$

since $S_1 + S_2 := \text{span}\{S_1 \cup S_2\} \subseteq V$. By induction we have $\dim(S_1 \cap \cdots \cap S_k) \geq \dim(S_1) + \cdots + \dim(S_k) - (k-1)\dim(V)$. If $\delta := \dim(S_1) + \cdots + \dim(S_k) - (k-1)\dim(V) \geq 1$ then $S_1 \cap \cdots \cap S_k$ contains at least $\delta \geq 1$ linearly independent vectors. For example, for the vector space $V := \mathbb{R}^3$ and subspaces S_1, S_2 defined by two non-parallel planes, their intersection $S_1 \cap S_2$ is a line in V and has a dimension at least $2+2-3 = 1$. In fact its dimension is exactly 1 because $S_1 + S_2 = V$. If S_3 is a plane that is not parallel to S_1 or S_2, $\dim(S_1 \cap S_2 \cap S_3) \geq 2+2+2-(2)(3) = 0$. It is exactly 0 (their intersection is a point) because $S_1 + S_2 + S_3 = V$.

We can view a matrix $M_{m,n}(F)$ as a vector in the vector space F^{mn}, or an array of scalars F in the vector space $F^{m \times n}$. A third perspective is to view a matrix $A \in M_{m,n}(F)$ as a *linear transformation* $A : F^n \to F^m$ mapping x to Ax. Then

- The *domain* of A is F^n.
- The *range* of A is the subspace $\text{range}(A) := \{Ax \in F^m : x \in F^n\} \subseteq F^m$. The dimension of $\text{range}(A)$ is called the *rank* of A, denoted by $\text{rank}(A)$.
- The *null space* of A is the subspace $\text{null}(A) := \{x \in \mathbb{F}^n : Ax = 0\} \subseteq F^n$. The dimension of $\text{null}(A)$ is called the *nullity* of A, denoted by $\text{nullity}(A)$.

The span $\text{range}(A)$ is also called the *column space* of A. Similarly $\{y^\mathsf{T} A : y \in F^m\}$ is called the *row space* of A. The *rank-nullity theorem* states that

$$\text{rank}(A) + \text{nullity}(A) = n = \text{rank}\left(A^\mathsf{H}\right) + \text{nullity}(A) \tag{A.1}$$

where the last equality holds if $F = \mathbb{C}$ or $\mathbb{R}$ and follows since $\text{rank}(A) = \text{rank}\left(A^\mathsf{H}\right)$. Note that $\text{range}\left(A^\mathsf{H}\right) \subseteq F^n$ whereas $\text{range}(A) \subseteq F^m$.

Henceforth we use $M_{m,n} := M_{m,n}(\mathbb{C})$ to denote the set of $m \times n$ matrices whose elements are in $\mathbb{C}$. We abbreviate them to $M_n := M_n(\mathbb{C})$ if $m = n$ and use $M := M(\mathbb{C})$ when m and n are arbitrary. Similarly for $M_{m,n}(\mathbb{R})$, $M_n(\mathbb{R})$ and $M(\mathbb{R})$ for matrices whose elements are in $\mathbb{R}$. We often write $A \in \mathbb{C}^{m \times n}$ (or $A \in \mathbb{R}^{m \times n}$) and call A a complex (or real) matrix to mean a matrix A in M (or $M(\mathbb{R})$) of size $m \times n$.

A.2 Polyhedral set and extreme point

We follow [54, Chapter 2] and define a *polyhedral set* $X \subseteq \mathbb{R}^n$ as a nonempty set specified by a finite number of affine inequalities:

$$X := \{x \in \mathbb{R}^n : Ax \leq b\}$$

for a given $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. Hence a polyhedral set is nonempty closed and convex. An important characterization of a polyhedral set is the following result e.g. [54, Proposition 2.3.3, p.106].

Theorem A.1 (Minkowski-Weyl representation)**.** A set $X \subseteq \mathbb{R}^n$ is polyhedral if and only if there is a finite set $\{v_1, \ldots, v_m\}$ and a finitely generated cone $K := \text{cone}(a_1, \ldots, a_k)$ such that

$$X = \text{conv}(v_1, \ldots, v_m) \; + \; \text{cone}(a_1, \ldots, a_k)$$

i.e.

$$X = \left\{ x \in \mathbb{R}^n : x = \sum_{i=1}^{m} \alpha_i v_i \; + \; y, \; \alpha_i \geq 0, \sum_i \alpha_i = 1, \; y \in K \right\}$$

□

Given a nonempty convex set $X \subseteq \mathbb{R}^n$ a vector $x \in X$ is an *extreme point* if there does not exist $y \neq x$, $z \neq x$, and $\alpha \in (0,1)$ such that $x = \alpha z + (1-\alpha)y$, i.e., if x is not a convex combination of other vectors in X that are distinct from x. Several facts are useful. An interior point cannot be an extreme point and an open set has no extreme points. A cone may have at most one extreme point, the origin. A polyhedral set has at most finitely many extreme points, and the minimum of a linear program is attained at an extreme point of its polyhedral feasible set. A polyhedral set may not possess any extreme points e.g. $X = \{(x_1, x_2) : x_1 = x_2\}$. The following result from [54, Propositions 2.1.5, p.98] provides an exact characterization of the existence of extreme points for polyhedral sets.

Lemma A.2. Let $X := \{x \in \mathbb{R}^n : Ax \leq b\}$ be a polyhedral set for some $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. Then X has an extreme point if and only if A has n linearly independent rows, i.e., $\text{rank}(A) = n$. □

A convex set that is compact is the convex hull of its extreme points; see e.g. [141, Theorem 2.3.4, p.111]. Carathéodory theorem then implies that every vector is a convex combination of at most $n+1$ extreme points. These constituent extreme points, however, may be different for different vectors.

Lemma A.3. Let $X \subseteq \mathbb{R}^n$ be convex and compact. Then

1 $X = \text{conv}\{\text{extreme points of } X\}$.

2 If $x \in X$ then $x = \sum_{i=1}^{n+1} \alpha_i v_i$ for some extreme points v_i of X (that may depend on x), and some $\alpha_i \in [0,1]$ with $\sum_i \alpha_i = 1$. □

A.3 Schur complement and matrix inversion formula

A.3.1 Schur complement

Let $M \in \mathbb{C}^{n\times n}$ and partition it into blocks:

$$M = \begin{bmatrix} A & B \\ D & C \end{bmatrix}$$

such that $C \in \mathbb{C}^{k\times k}$, $k < n$, is invertible and the other submatrices are of appropriate dimensions. The $(n-k)\times(n-k)$ matrix $M/C := A - BC^{-1}D$ is called the *Schur complement of block C* of matrix M. If A is invertible then the $k \times k$ matrix $M/A := C - DA^{-1}B$ is called the *Schur complement of block A* of matrix M.

Example A.2 (Gaussian elimination)**.** Schur complement arises from applying Gaussian elimination to a system of linear equations such as:

$$\begin{bmatrix} A & B \\ D & C \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \Leftrightarrow \begin{bmatrix} Ax + By \\ Dx + Cy \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

When C is invertible, Gaussian elimination expresses y in terms of x by multiplying the second equation by BC^{-1} and subtracting the result from the first equation. This corresponds to multiplying the equations on the left by a block lower-triangular matrix:

$$\begin{bmatrix} \mathbb{I}_{n-k} & -BC^{-1} \\ 0 & C^{-1} \end{bmatrix} \begin{bmatrix} A & B \\ D & C \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} A - BC^{-1}D & 0 \\ C^{-1}D & \mathbb{I}_k \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} \tag{A.2a}$$

where

$$\begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} := \begin{bmatrix} b_1 - BC^{-1}b_2 \\ C^{-1}b_2 \end{bmatrix}$$

If the Schur complement of C is invertible then the solutions for (x, y) can be read off equation (A.2a) as

$$\begin{aligned} x = \left(A - BC^{-1}D\right)^{-1} \hat{b}_1 &= (M/C)^{-1}\, \hat{b}_1 \\ y = -C^{-1}Dx + \hat{b}_2 &= -C^{-1}D\,(M/C)^{-1}\, \hat{b}_1 + \hat{b}_2 \end{aligned}$$

This means that

$$\begin{bmatrix} A - BC^{-1}D & 0 \\ C^{-1}D & \mathbb{I}_k \end{bmatrix}^{-1} = \begin{bmatrix} (M/C)^{-1} & 0 \\ -C^{-1}D\,(M/C)^{-1} & \mathbb{I}_k \end{bmatrix} \tag{A.2b}$$

□

Gaussian elimination can be represented as

$$\begin{bmatrix} \mathbb{I}_{n-k} & -BC^{-1} \\ 0 & \mathbb{I}_k \end{bmatrix} \begin{bmatrix} A & B \\ D & C \end{bmatrix} \begin{bmatrix} \mathbb{I}_{n-k} & 0 \\ -C^{-1}D & \mathbb{I}_k \end{bmatrix} = \begin{bmatrix} A - BC^{-1}D & 0 \\ 0 & C \end{bmatrix} \tag{A.3}$$

This equation implies (since $\det(M_1 M_2) = \det(M_1)\det(M_2)$)

$$\det(M) = \det(C)\,\det(M/C)$$
$$\operatorname{rank}(M) = \operatorname{rank}(C) + \operatorname{rank}(M/C)$$

Theorem A.4 (Schur complement)**.** Let $M \in \mathbb{C}^{n\times n}$ be partitioned as above with nonsingular C. Let $M/C := A - BC^{-1}D$ be the Schur complement of C of matrix M.

1. M is nonsingular if and only if M/C is nonsingular (given C is nonsingular).
2. $\det(M) = \det(C)\,\det(M/C)$.
3. $\operatorname{rank}(M) = \operatorname{rank}(C) + \operatorname{rank}(M/C)$.
4. Suppose M is symmetric. Then
   1. M is positive definite if and only if C and M/C are positive definite.
   2. Suppose C is positive semidefinite (not just nonsingular). M is positive semidefinite if and only if M/C is positive semidefinite.
5. If M and C are invertible, then M/C is invertible and

$$M^{-1} = \begin{bmatrix} (M/C)^{-1} & -(M/C)^{-1}BC^{-1} \\ -C^{-1}D\,(M/C)^{-1} & C^{-1} + C^{-1}D\,(M/C)^{-1}BC^{-1} \end{bmatrix}$$

6. If M and A are invertible, then $M/A := C - DA^{-1}B$ is invertible and

$$M^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B(M/A)^{-1}DA^{-1} & -A^{-1}B(M/A)^{-1} \\ -(M/A)^{-1}DA^{-1} & (M/A)^{-1} \end{bmatrix}$$

Proof Assertions 1, 2, 3 follow from (A.3). Example A.2 shows that (from (A.2a)):

$$\begin{bmatrix} \mathbb{I}_{n-k} & -BC^{-1} \\ 0 & C^{-1} \end{bmatrix} \begin{bmatrix} A & B \\ D & C \end{bmatrix} = \begin{bmatrix} A - BC^{-1}D & 0 \\ C^{-1}D & \mathbb{I}_k \end{bmatrix} \tag{A.4}$$

M is singular if and only if there exists a nonzero vector (x, y) in $\operatorname{null}(M)$, i.e.,

$$\begin{bmatrix} A - BC^{-1}D & 0 \\ C^{-1}D & \mathbb{I}_k \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0 \quad \Leftrightarrow \quad (A - BC^{-1}D)x = 0, \quad y = -C^{-1}Dx$$

Hence M is singular if and only if $A - BC^{-1}D$ is singular. Applying $\det(M_1 M_2) = \det(M_1)\det(M_2)$ to (A.4) we have $\det(M) = \det(C)\det(A - BC^{-1}D) = \det(C)\det(M/C)$.

For 4, A, C are symmetric and $D^{\mathsf{T}} = B$. Hence (A.3) becomes $FMF^{\mathsf{T}} = \operatorname{diag}(M/C, C)$ where F is nonsingular with

$$F := \begin{bmatrix} \mathbb{I}_{n-k} & -BC^{-1} \\ 0 & \mathbb{I}_k \end{bmatrix}, \qquad F^{-1} = \begin{bmatrix} \mathbb{I}_{n-k} & BC^{-1} \\ 0 & \mathbb{I}_k \end{bmatrix}$$

Then

$$x^{\mathsf{T}}Mx \;=\; \left(F^{-\mathsf{T}}x\right)^{\mathsf{T}} \operatorname{diag}(M/C, C)\left(F^{-\mathsf{T}}x\right) \;=\; y_1^{\mathsf{T}}(M/C)y_1 + y_2^{\mathsf{T}}Cy_2 \tag{A.5a}$$

where

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \;:=\; F^{-\mathsf{T}}x \;=\; \begin{bmatrix} \mathbb{I}_{n-k} & 0 \\ C^{-1}B^{\mathsf{T}} & \mathbb{I}_k \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \;=\; \begin{bmatrix} x_1 \\ C^{-1}B^{\mathsf{T}}x_1 + x_2 \end{bmatrix} \tag{A.5b}$$

If C and M/C are positive definite, then for any $x := (x_1, x_2) \neq 0$, $x^{\mathsf{T}}Mx = y_1^{\mathsf{T}}(M/C)y_1 + y_2^{\mathsf{T}}Cy_2 > 0$, i.e., M is positive definite. Conversely suppose M is positive definite, so that $y_1^{\mathsf{T}}(M/C)y_1 + y_2^{\mathsf{T}}Cy_2 > 0$ for any $(y_1, y_2) \neq 0$. If $y_1^{\mathsf{T}}(M/C)y_1 \leq 0$ for any $y_1 \neq 0$, then choose $x_1 = y_1$ and $x_2 = -C^{-1}B^{\mathsf{T}}x_1$ so that $y_1 \neq 0$ but $y_2 = 0$. We have from (A.5) that $x^{\mathsf{T}}Mx = y_1^{\mathsf{T}}(M/C)y_1 \leq 0$, contradicting that M is positive definite. Similarly if $y_2^{\mathsf{T}}Cy_2 \leq 0$ for any $y_2 \neq 0$, then choose $x_1 = 0$ and $x_2 = y_2$, yielding $x^{\mathsf{T}}Mx = y_2^{\mathsf{T}}Cy_2 \leq 0$, a contradiction. Therefore both M/C and C are positive definite.

If C is nonsingular and positive semidefinite, then C must be positive definite. Then (A.5) implies that M is psd (and not pd) if and only if M/C is psd (and not pd, setting $y_2 = 0$).

To prove 5, we have from (A.2)

$$\begin{bmatrix} A & B \\ D & C \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{I}_{n-k} & -BC^{-1} \\ 0 & C^{-1} \end{bmatrix}^{-1} = \begin{bmatrix} A - BC^{-1}D & 0 \\ C^{-1}D & \mathbb{I}_k \end{bmatrix}^{-1} = \begin{bmatrix} (M/C)^{-1} & 0 \\ -C^{-1}D\,(M/C)^{-1} & \mathbb{I}_k \end{bmatrix}$$

Hence

$$\begin{aligned} \begin{bmatrix} A & B \\ D & C \end{bmatrix}^{-1} &= \begin{bmatrix} (M/C)^{-1} & 0 \\ -C^{-1}D\,(M/C)^{-1} & \mathbb{I}_k \end{bmatrix} \begin{bmatrix} \mathbb{I}_{n-k} & -BC^{-1} \\ 0 & C^{-1} \end{bmatrix} \\ &= \begin{bmatrix} (M/C)^{-1} & -(M/C)^{-1}BC^{-1} \\ -C^{-1}D\,(M/C)^{-1} & C^{-1}D\,(M/C)^{-1}BC^{-1} + C^{-1} \end{bmatrix} \end{aligned}$$

The last assertion can be proved in the same way by eliminating x instead of y in Example A.2; see Exercise A.3. □

Let $I := \{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$, $J := \{j_1, \ldots, j_l\} \subseteq \{1, \ldots, n\}$, and A_{IJ} denote the submatrix obtained from deleting rows not in I and columns not in J.

- If $k = l$, i.e., A_{IJ} is square, then the *minor* M_{IJ} of A is the determinant of the submatrix A_{IJ}.
- If $I = J$, then A_{IJ} is called a *principal submatrix* and M_{IJ} a *principal minor* of A.
- If $I = J = \{1, \ldots, k\}$ with $k \leq n$, then A_{IJ} is called a *leading principal submatrix* of order k and M_{IJ} a *leading principal minor* of order k.

Theorem A.5 (Slyvester's criterion)**.** Suppose A is Hermitian. Then

1 A is positive definite if and only if all its leading principal minors are positive.

This involves n determinants: those of the upper left 1×1 matrix, upper left 2×2 matrix, ..., $\det(A)$.

2 A is positive semidefinite if and only if all its principal minors are nonnegative. This involves $\binom{n}{1}+\binom{n}{2}+\cdots+\binom{n}{n}$ determinants.

□

A.3.2 Matrix inversion lemma

A useful identity is the matrix inversion lemma or Sherman-Morrison-Woodbury formula. Let $A\in\mathbb{C}^{n\times n}$, $B\in\mathbb{C}^{n\times k}$, $C\in\mathbb{C}^{k\times k}$ and $D\in\mathbb{C}^{k\times n}$. Suppose A, C and the $k\times k$ matrix

$$\hat{C} := C^{-1}+DA^{-1}B \tag{A.6a}$$

are invertible. Then

$$(A+BCD)^{-1} = A^{-1} - A^{-1}\left(B\hat{C}^{-1}D\right)A^{-1} \tag{A.6b}$$

An important case is when $k \ll n$. Then the $k\times k$ matrix C is much smaller than A and the multiplication of C by B and D on the left and right respectively produces an $n\times n$ matrix BCD of the right size for addition with A. Similarly reversing the order of multiplication produces a much smaller $k\times k$ matrix $DA^{-1}B$ for addition with C^{-1} to produce the matrix $\hat{C}$ in (A.6a). We can thus view the role of (B,D) as transforming between sizes n and k to simplify the inversion of large matrices. In many applications BCD represents a low-rank update of A in a dynamical system or an additive noise to a transmitted signal A so that $A+BCD$ is the received signal. Suppose A^{-1} has been precomputed. Then $\hat{C}$ is much smaller and easier to invert than $A+BCD$. The matrix inversion formula allows us to compute the inverse of the updated or noisy matrix $A+BCD$ in terms of A^{-1} and $\hat{C}^{-1}$ when they exist.

Many special cases are useful. For instance when $A=\mathbb{I}_n$ and $C=\mathbb{I}_k$ we have:

$$(\mathbb{I}_n+BD)^{-1} = \mathbb{I}_n - B(\mathbb{I}_k+DB)^{-1}D$$

Note that BD is $n\times n$ while DB is $k\times k$ and hence the inverse on the right-hand side can be much easier to compute than that on the left-hand side. Using the push-through identity (see Exercise A.4) this is equivalent to:

$$(\mathbb{I}_n+BD)^{-1} = \mathbb{I}_n - (\mathbb{I}_n+BD)^{-1}BD = \mathbb{I}_n - BD(\mathbb{I}_n+BD)^{-1}$$

When $k=n$ and $B=D=\mathbb{I}_n$ we have the inversion formula for sum of two matrices:

$$(A+C)^{-1} = A^{-1} - A^{-1}\left(C^{-1}+A^{-1}\right)^{-1}A^{-1}$$

Merging $A^{-1}\left(C^{-1}+A^{-1}\right)^{-1}A^{-1}$ we have Hua's identity:

$$(A+C)^{-1} = A^{-1} - \left(A+AC^{-1}A\right)^{-1}$$

A.4 Change of basis, diagonalizability, Jordan form

Recall that we can interpret any $m\times n$ complex matrix A as a linear transformation that maps a vector $x\in\mathbb{C}^n$ to a vector $y=Ax\in\mathbb{C}^m$, where the basis in the domain $\mathbb{C}^n$ is the standard basis consisting of the columns of the $n\times n$ identity matrix $\mathbb{I}_n$ and the basis in the range $\mathbb{C}^m$ is the standard basis consisting of the columns of $\mathbb{I}_m$. Suppose we want to change the basis of the domain to (the columns of) an $n\times n$ nonsingular matrix V and the basis of the range to (the columns of) an $m\times m$ nonsingular matrix U. What is the new matrix $\tilde{A}$ that represents the same linear map with respect to the new bases?

A.4.1 Similarity transformation

Since V and U are bases of $\mathbb{C}^n$ and $\mathbb{C}^m$ respectively we can express any $x\in\mathbb{C}^n$ in terms of V and any vector $y\in\mathbb{C}^m$ in terms of U as

$$x = V\tilde{x} \quad \text{and} \quad y = U\tilde{y}$$

Hence a linear transformation A that maps any vector $x\in\mathbb{C}^n$ to a vector $y=Ax\in\mathbb{C}^m$ with respect to the standard bases implies

$$U\tilde{y} = y = Ax = AV\tilde{x}$$

Hence

$$\tilde{y} = \underbrace{U^{-1}AV}_{\tilde{A}}\tilde{x}$$

This means that any vector $\tilde{x}$ in the domain $\mathbb{C}^n$ with respect to the new basis V is mapped to the (same) vector $\tilde{y}$ in the range $\mathbb{C}^m$ with respect to the new basis U by the matrix (see Figure A.1)

$$\tilde{A} := U^{-1}AV \quad \text{or} \quad A = U\tilde{A}V^{-1}$$

For the special case where $n=m$ and the new bases for the domain and the range are the same, $U=V$,

$$\tilde{A} = V^{-1}AV \tag{A.7}$$

i.e., the new matrix $\tilde{A}$ represents the linear transformation under the new basis V. The mapping of A to $V^{-1}AV$ is called a *similarity transformation* of A by the nonsingular *similarity matrix* V.

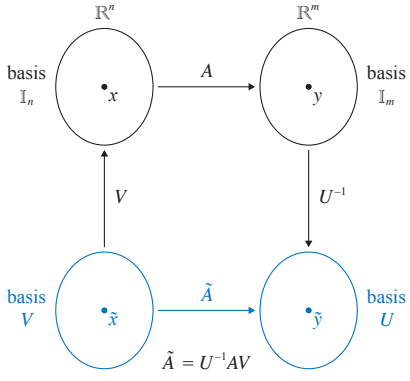


Figure A.1 Change of bases. The new matrix $\tilde{A}$ is similar to A when $n = m$ and $U = V$.

A.4.2 Diagonalizabilty and Jordan form

For the case where $n = m$ and $U = V$, if the basis V in (A.7) is such that $\tilde{A} = \Lambda$ is diagonal then the diagonal entries λ_i of Λ are the eigenvalues of A with the ith columns v_i of V as the corresponding eigenvectors, since

$$AV = V\Lambda \quad \text{or} \quad Av_i = \lambda_i v_i, \quad i = 1, \ldots, n$$

A is said to be *diagonalizable* in this case, i.e., by definition, A is diagonalizable if it is similar to a diagonal matrix Λ.

Not all $n \times n$ matrix A over the complex field is diagonalizable through a similarity transformation. We see above that A is diagonalizable if A has n linearly independent eigenvectors. Indeed having n linearly independent eigenvectors is also necessary for A's diagonalizability.[1] When A has fewer than n linearly independent eigenvectors, A is not similar to a diagonal matrix, but to a *Jordan form*, i.e., there exists an invertible matrix V such that

$$V^{-1}AV = J := \begin{bmatrix} J_1 & & \\ & \ddots & \\ & & J_m \end{bmatrix}$$

where J_i, $i = 1, \ldots, m$, are *Jordan blocks* of A:

$$J_i := \begin{bmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{bmatrix}$$

To compute the columns of V, consider Jordan block J_i and suppose without loss of

[1] A square matrix $A \in \mathbb{C}^{n \times n}$ is said to be *unitarily* diagonalizable if $V^{-1} = V^{\mathsf{H}}$ in (A.7). A matrix A is unitarily diagonalizable if and only if it is normal ($AA^{\mathsf{H}} = A^{\mathsf{H}}A$); see Chapter A.6.

generality that it corresponds to columns $1, 2, \ldots, k_i$. Equate these k_i columns on both sides of $AV = VJ$ to get

$$A\begin{bmatrix} | & | & & | \\ v_1 & v_2 & \cdots & v_{k_i} \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ v_1 & v_2 & \cdots & v_{k_i} \\ | & | & & | \end{bmatrix}\begin{bmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{bmatrix}$$

Therefore v_1 is the eigenvector corresponding to the eigenvalue λ_i and can be computed from

$$(A - \lambda_i I_n)v_1 = 0 \tag{A.8a}$$

The other columns $v_2, \ldots, v_{k_i}$ are not eigenvectors. They satisfy $Av_j = v_{j-1} + \lambda_i v_j$, $j = 2, \ldots, k_i$, and can be computed from

$$(A - \lambda_i I_n)v_j = v_{j-1}, \qquad j = 2, \ldots, k_i \tag{A.8b}$$

Multiplying both sides by $A - \lambda_i I_n$ yields $(A - \lambda_i I_n)^2 v_j = v_{j-2}$. Repeated multiplications then imply that the columns $v_1, \ldots, v_{k_i}$ satisfy:

$$\begin{aligned} (A - \lambda_i I_n)v_1 &= 0 \qquad && (v_1 \text{is eigenvector}) \\ (A - \lambda_i I_n)^2 v_2 &= 0 \qquad && (v_j \text{ are } \textit{generalized eigenvectors}, j = 2, \ldots, k_i) \\ &\vdots \\ (A - \lambda_i I_n)^{k_i} v_{k_i} &= 0 \end{aligned}$$

The characteristic polynomial $p(x) := \det(x\mathbb{I}_n - A)$ of A can be expressed in terms of the eigenvalues λ_i:

$$p(x) := \det(x\mathbb{I}_n - VJV^{-1}) = \det\left(V(x\mathbb{I}_n - J)V^{-1}\right) = \det(x\mathbb{I}_n - J) = \prod_{i=1}^{m} \det(x\mathbb{I}_{k_i} - J_i)$$

where J_i is the ith Jordan block of size $k_i \times k_i$, and $\mathbb{I}_{k_i}$ is the identity matrix of the same size. Since a Jordan block is upper triangular we have

$$\det(x\mathbb{I}_{k_i} - J_i) = (x - \lambda_i)^{k_i}$$

and hence

$$p(x) = \prod_{i=1}^{m} (x - \lambda_i)^{k_i}$$

There can be more than one Jordan block whose diagonal entries are the repeated eigenvalue λ_i. Let q be the number of *distinct* eigenvalues λ_j, $j = 1, \ldots, q$, and let m_j be the number of Jordan blocks corresponding to the distinct eigenvalue λ_j, so that $m = \sum_{j=1}^{q} m_j$. Then the characteristic polynomial can also be expressed in terms of

distinct eigenvalues as:

$$p(x) = \prod_{i=1}^{m} (x - \lambda_i)^{k_i} \quad = \quad \prod_{j=1}^{q} \prod_{i=1}^{m_j} (x - \lambda_j)^{k_i}$$

For each *distinct* eigenvalue λ_j, there are two quantities of interest:

1 *geometric multiplicity* m_j of λ_j: This is the number of Jordan blocks corresponding to λ_j. It is the dimension of the null space of $A - \lambda_j \mathbb{I}_n$ since each such block yields a single eigenvector of A.
2 *algebraic multiplicity* $\sum_{i=1}^{m_j} k_i$ of λ_j: This is the sum of the sizes k_i of all these Jordan blocks. It is the maximum degree of the factor $x - \lambda_j$ in the characteristic polynomial $p(x)$ of M.

Hence for each distinct eigenvalue λ_j

$$\text{algebraic multiplicity } \sum_{i=1}^{m_j} k_i \geq \text{ geometric multiplicity } m_j$$

We summarize implications of algebraic and geometric multiplicities on the diagonalizability of A in the following theorem.

Theorem A.6. With the notations above,

1 For each distinct eigenvalue λ_j, algebraic multiplicity = geometric multiplicity = m_j if and only if all Jordan blocks corresponding to λ_j have sizes $k_i = 1$. In this case, there are m_j eigenvectors corresponding to λ_j, they are linearly independent, and the null space of $A - \lambda_j I_n$ has dimension m_j.
2 A is diagonalizable if and only if algebraic multiplicity = geometric multiplicity for all eigenvalues, if and only if all Jordan blocks have sizes 1 and hence all superdiagonal entries are zero, if and only if A has n linearly independent eigenvectors.
3 As a special case, A is diagonalizable if A has n distinct eigenvalues (and hence all Jordan blocks are of size 1, $m_j = k_i = 1$ = algebraic multiplicity = geometric multiplicity).

A.5 Special matrices

Definition A.2 (Square matrices)**.** 1 A real or complex matrix $A \in \mathbb{F}^{n \times n}$, with $F = \mathbb{R}$ or $\mathbb{C}$, is *symmetric* if $A^{\mathsf{T}} = A$, *skew-symmetric* if $A^{\mathsf{T}} = -A$, and *orthogonal* if $A^{\mathsf{T}} = A^{-1}$.
2 A complex matrix $A \in \mathbb{C}^{n \times n}$ is *Hermitian* if $A^{\mathsf{H}} = A$, *skew-Hermitian* if $A^{\mathsf{H}} = -A$, and *unitary* if $A^{\mathsf{H}} = A^{-1}$.
3 A complex matrix $A \in \mathbb{C}^{n \times n}$ is *normal* if $AA^{\mathsf{H}} = A^{\mathsf{H}}A$. If A is real, this reduces to $AA^{\mathsf{T}} = A^{\mathsf{T}}A$.

4 *Positive semidefiniteness.*
- A complex matrix $A \in \mathbb{C}^{n\times n}$ is *positive semidefinite (psd)* (or *positive definite (pd)*) if $x^{\mathsf{H}}Ax$ is real and nonnegative (or real and positive) for all $x \in \mathbb{C}^n$.
- A real *symmetric* matrix $A \in \mathbb{R}^{n\times n}$ is *positive semidefinite (psd)* (or *positive definite (pd)*) if $x^{\mathsf{T}}Ax \geq 0$ (or $x^{\mathsf{T}}Ax > 0$) for all $x \in \mathbb{R}^n$.
- A complex or real matrix A is *negative semidefinite (nsd)* (or *negative definite (nd)*) if $-A$ is psd (or pd). It is *indefinite* if there are vectors $y, z \in \mathbb{F} \in \{\mathbb{C}, \mathbb{R}\}$ such that $y^{\mathsf{H}}Ay < 0 < z^{\mathsf{H}}Az$.

□

Remark A.1. 1 A *real* orthogonal matrix or a unitary matrix has columns (or rows) that are orthonormal basis of $\mathbb{R}^n$ or $\mathbb{C}^n$. A complex orthogonal matrix however is generally not unitary and their columns (or rows) are generally not orthonormal.

2 All Hermitian (symmetric), skew-Hermitian (skew-symmetric), or unitary complex matrices are normal, but the converse is not generally true. A *real* symmetric matrix is normal, but a complex symmetric matrix may or may not be normal (see Chapter A.6.4). If A is both triangular and normal, then A is diagonal.

3 A complex Hermitian (skew-Hermitian) matrix behaves like a real symmetric (skew-symmetric) matrix, e.g., they have real eigenvalues and are normal matrices. It therefore has a spectral decomposition according to Theorem A.13. A complex Hermitian matrix has real diagonal entries.

4 A complex symmetric matrix may or may not be normal. It therefore may or may not have a spectral decomposition. It always has a singular value decomposition (Theorem A.11) and a Takagi decomposition (Theorem A.17), and these are generally different decompositions.

5 Our definition of psd (or pd) requires symmetry for real matrices, but does not require Hermitian for complex matrices. This is because, for a complex matrix $A \in \mathbb{C}^{n\times n}$, A is psd (or pd) if and only if A is Hermitian and its eigenvalues are nonnegative (or positive), so our Definition A.2 for complex matrices implies Hermitian. For a real matrix $A \in \mathbb{R}^{n\times n}$, on the other hand, A can satisfy $x^{\mathsf{T}}Ax \geq 0$ for all $x \in \mathbb{R}^n$ but not be symmetric (as long as its symmetric component $(A+A^{\mathsf{T}})/2$ is psd or pd). Following [15, Definition 4.1.11, p. 231], we therefore restrict our definition to real symmetric matrices. Then A is psd (or pd) if and only if all its eigenvalues are nonnegative (or positive) [15, Theorem 4.1.10, p.231].

□

Theorem A.7 (Eigenvalues)**.** 1 A matrix A, real or complex, is invertible if and only if all its eigenvalues are nonzero.

2 If a matrix A is real symmetric or complex Hermitian, then all its eigenvalues are real.

3 A matrix A, real or complex, is psd (pd) if and only if $A^{\mathsf{H}} = A$ and all its eigenvalues are real and nonnegative (positive).

□

Definition A.3 (Diagonal dominance)**.** A matrix $A \in \mathbb{C}^{n\times n}$ is *diagonally dominant* if

$$|A_{ii}| \geq \sum_{j:j\neq i} |A_{ij}| \qquad \text{for all rows } i$$

A is *strictly diagonally dominant* if the inequalities are strict for all rows i.

The Geršgorin disc theorem states that all eigenvalues of a matrix $A \in \mathbb{C}^{n\times n}$ lie in the union of n discs

$$\cup_{i=1}^{n} \left\{ z \in \mathbb{C}^n \ : \ |z - A_{ii}| \leq \sum_{j:j\neq i} |A_{ij}| \right\}$$

If A is strictly diagonally dominant then the origin is outside Geršgorin discs, i.e., all eigenvalues of A are nonzero. The geometry of the Geršgorin discs also implies the following property.

Theorem A.8. 1 A strictly diagonally dominant matrix is invertible (but not necessarily positive definite).

2 Suppose $A \in \mathbb{C}^{n\times n}$ is Hermitian with (real) nonnegative diagonal entries $A_{ii} \geq 0$.

- If A is diagonally dominant then it is positive semidefinite.
- If A is strictly diagonally dominant then it is positive definite and invertible.

Proof Part 1 follows from the Geršgorin disc theorem. For part 2, for any $x \in \mathbb{C}^n$ we have

$$x^{\mathsf{H}} A x = \sum_{i,j} A_{ij} x_i^{\mathsf{H}} x_j \ = \ \sum_i \left(A_{ii} |x_i|^2 \ + \ \sum_{j:j\neq i} A_{ij} x_i^{\mathsf{H}} x_j \right)$$

Substitute $A_{ii} \geq \sum_{j:j\neq i} |A_{ij}|$ (diagonal dominance) to get

$$\begin{aligned} x^{\mathsf{H}} A x \ &\geq \ \sum_i \sum_{j:j\neq i} \left(|A_{ij}||x_i|^2 \ + \ A_{ij} x_i^{\mathsf{H}} x_j \right) \\ &= \ \sum_{(i,j):i\neq j} \left(|A_{ij}||x_i|^2 + |A_{ji}||x_j|^2 \ + \ A_{ij} x_i^{\mathsf{H}} x_j + A_{ji} x_j^{\mathsf{H}} x_i \right) \end{aligned}$$

Since $A_{ji} = A_{ij}^{\mathsf{H}}$ (A is Hermitian) we have

$$x^{\mathsf{H}} A x \geq \sum_{(i,j):i\neq j} |A_{ij}| \left(|x_i|^2 + |x_j|^2 - |x_i^{\mathsf{H}}||x_j| - |x_j^{\mathsf{H}}||x_i| \right) = \sum_{(i,j):i\neq j} |A_{ij}| \left(|x_i| - |x_j| \right)^2 \geq 0$$

If A is strictly diagonally dominant then the inequality is strict and therefore A is positive definite. □

Unitary matrices have the following properties (e.g. [15, Theorem 2.1.4, p.84]).

Lemma A.9. Consider a complex matrix $U \in M_n := M_n(\mathbb{C})$. The following are equivalent:

- U is unitary.
- $U^{\mathsf{H}}U = \mathbb{I}$.
- The columns of U are orthornormal.
- U^{H} is unitary.
- $UU^{\mathsf{H}} = \mathbb{I}$.
- The rows of U are orthornormal.
- $\|Ux\|_2 = \|x\|_2$ for all $x \in \mathbb{C}^n$ where $\|\cdot\|_2$ is the Euclidean norm.

A unitary matrix can be interpreted as a rotation operator, i.e., the product Ux rotates the vector x without expanding its Euclidean norm, $\|Ux\|_2^2 = x^{\mathsf{H}}(U^{\mathsf{H}}U)x = x^{\mathsf{H}}x = \|x\|_2^2$. In fact, the Euclidean norm is the only vector norm that is unitarily invariant, i.e., $\|Ux\| = \|x\|$ for all $x \in \mathbb{C}^n$ and all unitary matrices U with $\|e_i\| = 1$; see Chapter A.8.1.

Recall that a unitary matrix is normal because $UU^{\mathsf{H}} = U^{\mathsf{H}}U = \mathbb{I}$, and hence unitarily diagonalizable (Theorem A.13). If it is also symmetric then the unitary matrix is real orthogonal according to the following result [15, Corollary 2.5.18, p.139].

Lemma A.10. Suppose $U \in M_n := M_n(\mathbb{C})$ is unitary and symmetric. Then

1 If $U = \operatorname{diag}(a_1, \cdots, a_n)$ then $a_j = e^{\mathbf{i}\theta_j}$ for some $\theta_j \in \mathbb{R}^n$.
2 *Spectral decomposition*. There exist real orthogonal matrix $Q \in \mathbb{R}^{n\times n}$ and real $\theta_1, \cdots, \theta_n$ in $[0, 2\pi)$ such that

$$U = Q \underbrace{\operatorname{diag}\left(e^{\mathbf{i}\theta_1}, \cdots, e^{\mathbf{i}\theta_n}\right)}_{\Lambda} Q^{\mathsf{T}} =: Q\Lambda Q^{\mathsf{T}}$$

where $\lambda_j := e^{\mathbf{i}\theta_j}$ are the eigenvalues of U and the columns of Q are an orthonormal set of corresponding (real) eigenvectors of U.

A.6 SVD, spectral decompositions, complex symmetric matrices

In this subsection we review the various matrix decompositions and their relationship, as shown in Figure A.2.

A.6.1 Singular value decomposition for any matrix

Consider a complex matrix $A \in \mathbb{C}^{m\times n}$. Suppose there exists a real value $\sigma \geq 0$ and nonzero vectors $v \in \mathbb{C}^m$, $w \in \mathbb{C}^n$ such that

$$Aw = \sigma v \tag{A.9}$$

In this case, (σ, v, w) are called respectively a *singular value*, associated *left singular vector* and *right singular vector* of A. The next result says that every matrix A has

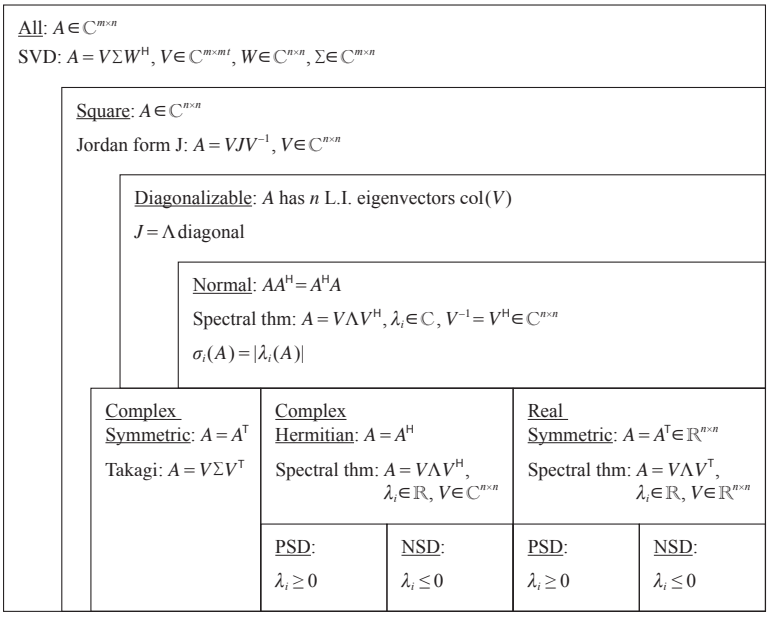


Figure A.2 Matrix decompositions. Singular value decomposition (Thm A.11), Diagonalizability (Thm A.6), Spectral theorems (Thms A.13, A.15, A.16), Takagi's decomposition (Thm A.17).

m orthonormal left singular vectors $v_1, \ldots, v_m \in \mathbb{C}^m$, n orthonormal right singular vectors $w_1, \ldots, w_n \in \mathbb{C}^n$, and at most $q := \min\{m, n\}$ strictly positive singular values $\sigma_1, \ldots, \sigma_q$. Like eigenvalues the singular values σ_i are unique. Like eigenvectors, left and right singular vectors (v_i, w_i) are generally not unique. As we will see below, they are eigenvectors of AA^{H} and $A^{\mathsf{H}}A$ respectively; but the converse may not hold, i.e., not every eigenvector of AA^{H} and that of $A^{\mathsf{H}}A$ may satisfy (A.9). For example, if (v_i, w_i) are singular vectors of unit Euclidean norm, so are $(e^{\mathbf{i}\theta}v_i, e^{\mathbf{i}\theta}w_i)$ for any $\theta \in \mathbb{R}$. Moreover the matrix A can be factorized as follows [15, Theorem 2.6.3, p.150].

Consider an $m \times n$ matrix Σ and a diagonal matrix $\Sigma_q = \text{diag}(\sigma_1, \cdots, \sigma_q)$ of size $q := \min\{m, n\}$. We will abuse notation and call Σ *diagonal*, even if $m \neq n$, if Σ is of the form:

$$\Sigma \;=\; \begin{cases} \Sigma_q & \text{if } m = n \\ \begin{bmatrix} \Sigma_q & 0 \end{bmatrix} & \text{if } n > m = q \\ \begin{bmatrix} \Sigma_q \\ 0 \end{bmatrix} & \text{if } m > n = q \end{cases} \tag{A.10}$$

Theorem A.11 (Singular value decomposition)**.** For any matrix $A \in \mathbb{C}^{m\times n}$, there exists unitary matrices $V \in \mathbb{C}^{m\times m}$ and $W \in \mathbb{C}^{n\times n}$, and a real diagonal matrix $\Sigma \in \mathbb{R}^{m\times n}$ of the form in (A.10) with

$$\sigma_1 \;\geq\; \sigma_2 \;\cdots\; \geq \;\sigma_q \;\geq 0$$

such that

$$AW = V\Sigma \quad \text{or} \quad A = V\Sigma W^{\mathsf{H}} \tag{A.11}$$

with $V^{-1} = V^{\mathsf{H}}$ and $W^{-1} = W^{\mathsf{H}}$. Moreover

1 The nonzero singular values of A are the positive square roots of the eigenvalues of AA^{H} (or equivalently of $A^{\mathsf{H}}A$):

$$\sigma_i = +\sqrt{\lambda_i(AA^{\mathsf{H}})} = +\sqrt{\lambda_i(A^{\mathsf{H}}A)}, \qquad i = 1, \ldots, q$$

2 If $r \leq q$ of the q singular values σ_i are positive, then A is of rank r and

$$A = \sum_{i=1}^{r} \sigma_i \, v_i w_i^{\mathsf{H}}$$

3 If V and W are unitary matrices such that $A = V\Sigma W^{\mathsf{H}}$ then
- the columns of V are an orthonormal set of eigenvectors of AA^{H} because $AA^{\mathsf{H}} = V\Sigma^2 V^{\mathsf{H}}$, and
- the columns of W are an orthonormal set of eigenvectors of $A^{\mathsf{H}}A$ because $A^{\mathsf{H}}A = W\Sigma^2 W^{\mathsf{H}}$;

but the converse does not necessarily hold.

If A is real then V and W can be taken as real orthogonal matrices. □

The rank of A is the number its positive singular values, which is no less than (and can be greater than) the number of its nonzero eigenvalues of A. As we will see below (Theorem A.13) $\text{rank}(A)$ is equal to the number of nonzero (generally complex) eigenvalues if A is normal.

Theorem A.11 does not provide a method to compute the unitary factors (V, W) in the singular value decomposition (A.11). This is because not every pair of orthonormal sets of eigenvectors of AA^{H} and $A^{\mathsf{H}}A$ respectively may be the unitary factors (V, W) in (A.11) when the eigenvalues associated with AA^{H} or with $A^{\mathsf{H}}A$ are not distinct. We describe how to compute unitary factors (V, W) in (A.11) when A is square $(m = n)$ (see [15, Theorem 2.6.3, p.150] for details). When A is not normal, AA^{H} and $A^{\mathsf{H}}A$ are not equal, but they are unitarily similar since they have the same eigenvalues, i.e., there exists a unitary matrix Y such that $A^{\mathsf{H}}A = Y(AA^{\mathsf{H}})Y^{\mathsf{H}}$. Moreover YA is normal and hence it has a spectral decomposition according to Theorem A.13, $YA = X\Lambda X^{\mathsf{H}}$ where $\Lambda := \text{diag}\,(\lambda_1, \cdots, \lambda_n)$ consists of the eigenvalues of YA and the columns of X are an *arbitrary* orthonormal set of corresponding eigenvectors of YA. Let $\lambda_i = |\lambda_i| e^{\mathbf{i}\theta_i}$, $\Sigma_q := \text{diag}\,(|\lambda_1|, \cdots, |\lambda_n|)$, $D := \text{diag}\left(e^{\mathbf{i}\theta_1}, \cdots, e^{\mathbf{i}\theta_n}\right)$ so that $\Lambda = \Sigma_q D$. Then, since $YA = X\Sigma_q D X^{\mathsf{H}}$, we have

$$A = \underbrace{\left(Y^{\mathsf{H}}X\right)}_{V} \Sigma_q \underbrace{\left(DX^{\mathsf{H}}\right)}_{W^{\mathsf{H}}} \tag{A.12}$$

i.e., $V := Y^{\mathsf{H}}X$ and $W := XD^{\mathsf{H}}$. We illustrate this in the next example.

Example A.3. Consider $A := \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. Show that

1 Not arbitrary orthonormal sets of eigenvectors of AA^{H} and $A^{\mathsf{H}}A$ can be the unitary matrices (V, W) in the SVD (A.11).
2 Compute (V, W) according to the prescription (A.12). (Since A is real symmetric and hence normal, an alternative way to compute a (possibly different) pair (V, W) is given in Theorem A.16; see Example A.4.)

Solution. The matrices AA^{H} and $A^{\mathsf{H}}A$ are

$$AA^{\mathsf{H}} \;=\; A^{\mathsf{H}}A \;=\; A^2 \;=\; \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \;=\; I$$

Therefore the eigenvalues of AA^{H} and those of $A^{\mathsf{H}}A$ are 1 and $\Sigma = I$. Moreover every vector x is an eigenvector of AA^{H} and of $A^{\mathsf{H}}A$, but not arbitrary orthonormal sets of eigenvectors can be (V, W) in SVD (A.11). For instance, if Q is any unitary matrix (and hence its columns are an orthonormal set of eigenvectors of AA^{H} and of $A^{\mathsf{H}}A$), $V = W = Q$ does not satisfy (A.11):

$$Q\Sigma Q^{\mathsf{H}} \;=\; QQ^{\mathsf{H}} \;=\; I \;\neq\; A$$

It is therefore necessary that V and W are different matrices in (A.11).

To compute (V, W) using (A.12), we choose $Y = I$ to be the identity matrix that relates AA^{H} and $A^{\mathsf{H}}A$ through unitary similarity, i.e., $A^{\mathsf{H}}A = I = Y(AA^{\mathsf{H}})Y^{\mathsf{H}}$. Next we compute the spectral decomposition of YA: the eigenvalues of $YA = A$ are $\lambda_1 := 1$, $\lambda_2 := -1$ with corresponding orthonormal set of eigenvectors (unique up to a rotation)

$$x_1 \;:=\; \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad\qquad x_2 \;:=\; \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Hence

$$YA \;=\; A \;=\; X\Lambda X^{\mathsf{H}} \;=\; \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \;=\; \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Then $D := \operatorname{diag}\left(e^{\mathbf{i}\theta_1}, e^{\mathbf{i}\theta_2}\right) = \operatorname{diag}(1, -1)$ and hence

$$\Sigma_q \;:=\; \operatorname{diag}\left(|\lambda_1|, |\lambda_2|\right) \;=\; I, \quad V \;:=\; Y^{\mathsf{H}}X \;=\; \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad W \;:=\; XD^{\mathsf{H}} \;=\; \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$$

It can be verified that indeed $A = V\Sigma_q W^{\mathsf{H}}$. □

Suppose $m \leq n$ but $\operatorname{rank}(A) =: r < m$. For a given V in the theorem, even though $A = V\Sigma W^{\mathsf{H}}$, W defined by $W^{\mathsf{H}} := \Sigma^{\dagger}V^{\mathsf{H}}A$ generally does not satisfy the singular value decomposition (A.11) because in that case $V\Sigma W^{\mathsf{H}} = V\Sigma\left(\Sigma^{\dagger}V^{\mathsf{H}}A\right) \neq A$ because

$V\Sigma\Sigma^\dagger V^{\mathsf{H}} \neq I_m$; see Exercise A.7. Here $\Sigma^\dagger$ is obtained from Σ by replacing its positive singular values σ_i by $1/\sigma_i$ and taking the transpose.

The set of singular values making up Σ is unique. The unitary factors (V, W) is non-unique, but given a pair, all possible pairs can be related, according to the following result from [15, Theorem 2.6.5, p.152].

Theorem A.12 (Uniqueness of (V, W))**.** Let $A \in \mathbb{C}^{m\times n}$ have a singular value decomposition $A = V\Sigma W^{\mathsf{H}}$ as in Theorem A.11. Then

1 $A = \hat{V}\Sigma\hat{W}$ for some unitary matrices $(\hat{V}, \hat{W})$ if and only if there are unitary block-diagonal matrices $\tilde{V}$ and $\tilde{W}$ such that

$$\hat{V} = V\tilde{V}, \qquad \hat{W} = W\tilde{W}$$

2 If A is square $(m = n)$ and nonsingular then $\tilde{V} = \tilde{W}$.

□

Properties of singular values.

1 Matrix transpose and conjugate: $\sigma_i(A) = \sigma_i(A^{\mathsf{T}}) = \sigma_i(A^{\mathsf{H}}) = \sigma_i(\overline{A})$.
2 Unitary transformation: for any unitary matrices U and V, $\sigma_i(A) = \sigma_i(UAV)$. In particular $\sigma_i(A) = \sigma_i(UA) = \sigma_i(AV)$ (setting $V = I$ or $U = I$).
3 Interlacing properties:
   - If B denote A with one of its rows *or* columns deleted, then
   $$\sigma_{i+1}(A) \leq \sigma_i(B) \leq \sigma_i(A)$$
   - If B denote A with one of its rows *and* columns deleted, then
   $$\sigma_{i+2}(A) \leq \sigma_i(B) \leq \sigma_i(A)$$
   - If B denote any $(m-k)\times(n-l)$ submatrix of A, then
   $$\sigma_{i+k+l}(A) \leq \sigma_i(B) \leq \sigma_i(A)$$
4 Singular values of $A+B$: for any $A, B \in \mathbb{C}^{m\times n}$
   - $\sum_{i=1}^{k} \sigma_i(A+B) \leq \sum_{i=1}^{k} (\sigma_i(A)+\sigma_i(B))$, $k = \min\{m,n\}$.
   - $\sigma_{i+j-1}(A+B) \leq \sigma_i(A)+\sigma_j(B)$, $i+j-1 \leq \min\{m,n\}$.
5 Singular values of AB: for any $A, B \in \mathbb{C}^{m\times n}$
   - $\sigma_n(A)\sigma_i(B) \leq \sigma_i(AB) \leq \sigma_1(A)\sigma_i(B)$.
   - $\prod_{i=1}^{k} \sigma_i(AB) \leq \prod_{i=1}^{k} \sigma_i(A)\,\sigma_i(B)$.
6 Singular value and eigenvalues: For any matrix $A \in \mathbb{C}^{n\times n}$
   - If A is normal, then $\sigma_i(A) = |\lambda_i(A)|$, $i = 1, \ldots, n$. (Note that $\lambda_i(A) \in \mathbb{C}$.)
   Proof: Spectral theorem gives $A = U\Lambda U^{\mathsf{H}}$; hence $AA^{\mathsf{H}} = U\Lambda\overline{\Lambda}U^{\mathsf{H}} = U|\Lambda|^2U^{\mathsf{H}}$. Hence $|\lambda_i(A)|^2$ are eigenvalues of AA^{H}, implying $\sigma_i(A) = \sqrt{\lambda_i(AA^{\mathsf{H}})} = |\lambda_i(A)|$.

- Weyl's theorem: Assume eigenvalues satisfy $|\lambda_1(A)| \geq \cdots \geq |\lambda_n(A)|$. Then

$$\prod_{i=1}^{k} |\lambda_i(A)| \leq \prod_{i=1}^{k} \sigma_i(A), \qquad k = 1, \ldots, n$$

Consider the set of complex square matrices, i.e., $m = n$. Every square matrix $A \in \mathbb{C}^{n\times n}$ is similar to a Jordan form J, i.e., there exists an invertible matrix $P \in \mathbb{C}^{n\times n}$ such that

$$A = PJP^{-1}$$

A is said to be *diagonalizable* if its Jordan form $J =: \Lambda$ is diagonal. Therefore A is diagonalizable if and only if A has n linearly independent eigenvectors; see Theorem A.6. In that case the columns of P are these eigenvectors, Λ has the corresponding eigenvalues on its diagonal, and $AP = P\Lambda$.

A.6.2 Spectral decomposition for normal matrices

Recall that A is *normal* if $AA^{\mathsf{H}} = A^{\mathsf{H}}A$ and that all unitary, Hermitian, or skew-Hermitian matrices are normal (the converse is not generally true). For any matrices $A, B \in \mathbb{C}^{n\times n}$, if $BA = I$ then B is unique and $B = A^{-1}$. This is because A being nonsingular means that $Ax = b$ and $x^{\mathsf{T}}A = b^{\mathsf{T}}$ has a unique solution x for any $b \in \mathbb{C}^n$; take b to be each column of I.

Normal matrices are exactly those that are *unitarily* diagonalizable to which the *spectral theorem* applies [15, Theorem 2.5.3, p.133].

Theorem A.13 (Spectral theorem for normal matrices)**.** A complex square matrix $A \in \mathbb{C}^{n\times n}$ is normal if and only if it is unitarily diagonalizable, i.e., there exists a unitary matrix $U \in \mathbb{C}^{n\times n}$ and a complex diagonal matrix $\Lambda \in \mathbb{C}^{n\times n}$ with

$$A = U\Lambda U^{\mathsf{H}} \;=\; \sum_{i=1}^{n} \lambda_i u_i u_i^{\mathsf{H}} \tag{A.13}$$

where

1 the diagonal entries of $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ are eigenvalues of A (generally complex);
2 the columns of U are an *arbitrary* orthonormal set of corresponding eigenvectors of A.

Hence if A is normal, then rank A = number of nonzero eigenvalues and the sum in (A.13) becomes

$$A = U\Lambda U^{\mathsf{H}} \;=\; \sum_{i=1}^{\operatorname{rank} A} \lambda_i u_i u_i^{\mathsf{H}}$$

□

Hence while A is diagonalizable if and only if it has n linearly independent eigenvectors, A is unitarily diagonalizable (or equivalently normal) if and only if it has an orthonormal set of n eigenvectors.

The eigenvalues Λ of A in Theorem A.13 are unique, but the eigenspace of A always has more than one orthonromal basis. Since two basis U and V can always be related by a unitary matrix, we have the following uniqueness result from [15, Theorem 2.5.4, p.134].

Theorem A.14 (Uniqueness of unitary U)**.** Let $A \in \mathbb{C}^{n\times n}$ be normal with spectral decomposition $A = U\Lambda U^{\mathsf{H}}$ where U is unitary and Λ is diagonal matrix consisting of the eigenvalues of A. Then

1 $A = V\Lambda V^{\mathsf{H}}$ for a unitary matrix V if and only if there is a block-diagonal unitary matrix W such that $U = VW$.
2 In particular, if A has n distinct eigenvalues then W is a diagonal unitary matrix of the form $W = \text{diag}\left(e^{\mathbf{i}\theta_1}, \cdots, e^{\mathbf{i}\theta_n}\right)$.
3 Two normal matrices A and B are unitarily similar, i.e., $A = WBW^{\mathsf{H}}$ for some unitary matrix W, if and only if they have the same eigenvalues.

□

For a normal matrix A the eigenvalues λ_i are complex in general. A normal matrix A is Hermitian if and only if all its eigenvalues are real. If A is Hermitian then the eigenvalues are real [176, Theorem 4.1.5, p.171].

Theorem A.15 (Spectral theorem for Hermitian matrices)**.** A complex square matrix $A \in \mathbb{C}^{n\times n}$ is Hermitian if and only if it is unitarily diagonalizable with real eigenvalues, i.e., there exist a unitary matrices $U \in \mathbb{C}^{n\times n}$ and a real diagonal matrix $\Lambda \in \mathbb{R}^{n\times n}$ with

$$A = U\Lambda U^{\mathsf{H}} \;=\; \sum_{i=1}^{n} \lambda_i u_i u_i^{\mathsf{H}} \tag{A.14}$$

where

1 $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ is real and consists of the eigenvalues of A;
2 the columns of U are an *arbitrary* orthonormal set of corresponding eigenvectors of A.

Hence if A is Hermitian, then rank A = number of nonzero eigenvalues and the sum in (A.14) becomes

$$A = U\Lambda U^{\mathsf{H}} \;=\; \sum_{i=1}^{\text{rank } A} \lambda_i u_i u_i^{\mathsf{H}}$$

Moreover, if A is real and symmetric then U above can be taken as real and orthogonal. □

To explain the last statement let A be a real symmetric matrix. First a Hermitian matrix A has real eigvenvalues λ because if v are the corresponding eivenvectors, then $Av = \lambda v$ and hence $v^{\mathsf{H}}Av = \lambda\|v\|^2$. Taking Hermitian transpose shows $v^{\mathsf{H}}A^{\mathsf{H}}v = v^{\mathsf{H}}Av = \bar{\lambda}\|v\|^2$ where $\bar{\lambda}$ denotes the complex conjugate of λ. Therefore $\bar{\lambda} = \lambda$, i.e., λ is real. Next for eigenvector v, take the Hermitian transpose of $Av = \lambda v$ we have $v^{\mathsf{H}}A^{\mathsf{H}} = v^{\mathsf{H}}A = \lambda v^{\mathsf{H}}$ since λ is real. If A is real symmetric then taking the transpose we have $A\bar{v} = \lambda\bar{v}$ where $\bar{v}$ is the componentwise complex conjugate of v. Therefore if v is an eigenvector of a real symmetric matrix A corresponding to λ, then so is its complex conjugate $\bar{v}$ as well as the real vector $v + \bar{v}$, i.e., the eigenvector of A can be taken to be real.

For general matrices, about the only characterization of its eigenvalues is that they are roots of the characteristic polynomial (see the discussion leading up to Theorem A.6). For Hermitian matrices, however, the spectral theorem leads to a variational characterization of eigenvalues [176, Theorem 4.2.2, p.176]. If $A \in \mathbb{C}^{n\times n}$ is Hermitian then

$$\lambda_{\min} \leq \frac{x^{\mathsf{H}}Ax}{x^{\mathsf{H}}x} \leq \lambda_{\max}, \qquad \forall x \in \mathbb{C}^n \tag{A.15a}$$

and

$$\lambda_{\min} = \min_{x\neq 0} \frac{x^{\mathsf{H}}Ax}{x^{\mathsf{H}}x}, \qquad \lambda_{\max} = \max_{x\neq 0} \frac{x^{\mathsf{H}}Ax}{x^{\mathsf{H}}x} \tag{A.15b}$$

Theorem A.15 implies that A is positive semidefinite if and only if A is Hermitian and all its eigenvalues are (real and) nonnegative, and that A is positive definite if and only if A is Hermitian and all its eigenvalues are (real and) positive.

A.6.3 SVD and unitary diagonalization

Consider a normal matrix $A \in \mathbb{C}^{n\times n}$. Since $AA^{\mathsf{H}} = A^{\mathsf{H}}A$, they have the same eigenvectors. This does *not* mean, in general, that $W = V$ in a singular value decomposition $A = V\Sigma W^{\mathsf{H}}$. Indeed, if $W = V$ then it is necessary that $A = V\Sigma V^{\mathsf{H}}$ is positive semidefinite, but a normal A may not be positive semidefinite. The eigenvalues of a normal matrix are complex, those of a Hermitian matrix are real, and those of a positive semedefinite matrix are real and nonnegative. The following relationship between singular value decomposition of a normal matrix A and its unitary diagonalization is proved in Exercise A.9.

Theorem A.16 (SVD and unitary diagonalization)**.** Consider a normal matrix $A \in \mathbb{C}^{n\times n}$ and let $A = U\Lambda U^{\mathsf{H}}$ be a unitary diagonalization of A described in Theorem A.13 where $\Lambda := \operatorname{diag}(\lambda_i)$ has the eigenvalues $\lambda_i \in \mathbb{C}$ of A on its diagonal and the columns of

U are an *arbitrary* orthonormal set of corresponding eigenvectors. Write $\lambda_i = |\lambda_i| e^{i\theta_i}$ for some $\theta_i \in \mathbb{R}$; set $\theta_i = 0$ if $\lambda_i = 0$. Let $D := \text{diag}\,(e^{\mathbf{i}\theta_1}, \ldots, e^{\mathbf{i}\theta_m})$. Then

1 $V := U, \Sigma := |\Lambda|, W := UD^{\mathsf{H}}$ form a singular value decomposition $A = V\Sigma W^{\mathsf{H}}$ of A.
2 The pseudo-inverse of A is $A^{\dagger} := U\Lambda^{\dagger}U^{\mathsf{H}}$ where the diagonal matrix $\Lambda^{\dagger}$ is obtained from Λ by replacing nonzero $\lambda_i \in \mathbb{C}$ by $1/\lambda_i$.
3 A is Hermitian if and only if D in W is a real matrix, i.e., $e^{\mathbf{i}\theta_i} = 1$ or -1.
4 A is positive semidefinite if and only if $V = W := U$ and $\Sigma := \Lambda$ forms a singular value decomposition $A = V\Sigma W^{\mathsf{H}} = U\Lambda U^{\mathsf{H}}$, i.e., SVD and unitary diagonalization of A coincide.

The theorem also prescribes a way to compute a singular value decomposition $A = V\Sigma W$ when A is normal. In this case we can take the columns of V to be an arbitrary orthonormal set of eigenvectors of A (which will also be eigenvectors of AA^*). This may not be the case if A is not normal and the more general method prescribed by (A.12) is needed to compute SVD (see Example A.3). The theorem is illustrated in the following example.

Example A.4. Use Theorem A.16 to compute the SVD of the normal matrix A in Example A.3.

Solution. Clearly $A = A^{\mathsf{H}} = A^{\mathsf{T}} = \bar{A}$ and A is real symmetric and hence normal. Its eigenvalues are $\lambda_i = \pm 1$ with corresponding eigenvectors in the columns of U in the unitary diagonalization:

$$A = U\Lambda U^{\mathsf{H}} \;:=\; \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 1 & \\ & -1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\frac{1}{\sqrt{2}}$$

Note that A is not positive semidefinite and therefore $W \neq U$ in the singular value decomposition of A. According to Theorem A.16, the angle matrix $D = \text{diag}(1, -1)$ and the unitary factors (V, W) in the SVD $A = V\Sigma W^{\mathsf{H}}$ are given by

$$\Sigma := |\Lambda| = I, \quad V := U = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad W := UD^{\mathsf{H}} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$$

which agrees with those computed in Example A.3. (The decomposition in these two examples agree because the matrix Y in Example A.3 has been chosen to be $Y = I$ so that $YA = A$.) □

A.6.4 Complex symmetric matrices

Consider a complex symmetric matrix $A \in \mathbb{C}^{n\times n}$ with $A = A^{\mathsf{T}}$. Then $A^{\mathsf{H}} = \bar{A}$ where $\bar{A}$ is the matrix obtained from A by taking its complex conjugate componentwise. A is not Hermitian unless A is a real matrix. The following result, from [15, Corollary 2.6.6, p.153], is called the Takagi's factorization for complex symmetric matrices.

Theorem A.17 (Takagi's decomposition)**.** A complex matrix $A \in \mathbb{C}^{n\times n}$ is symmetric $A = A^{\mathsf{T}}$ if and only if there is a unitary matrix $U \in \mathbb{C}^{n\times n}$ and a real nonnegative diagonal matrix $\Sigma := \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$ such that

$$A = U\Sigma U^{\mathsf{T}} \tag{A.16}$$

where Σ consists of the nonnegative square roots of the eigenvalues of $A\bar{A}$. □

The columns of the unitary matrix U in (A.16) are generally neither the singular vectors nor the eigenvectors of A; see the proof below. A Takagi decomposition of a complex symmetric matrix A is therefore generally different from its singular value decomposition. A Takagi decomposition of a real symmetric matrix may not have real factors. In contrast, its spectral decomposition in terms of its eigenvalues, rather than singular values, can always use real orthogonal factors according to Theorem A.15.

We provide a sketch of the proof from [15, Corollary 2.6.6, p.153].

Proof sketch of Theorem A.17 Let a singular value decomposition of A be $A = V\Sigma W^{\mathsf{H}}$ according to Theorem A.11. Since $A = A^{\mathsf{T}}$ we have $A = V\Sigma W^{\mathsf{H}} = \bar{W}\Sigma\bar{V}^{\mathsf{H}}$ where $(\bar{V}, \bar{W})$ are componentwise complex conjugate of (V, W). The uniqueness Theorem A.12 then implies the existence of unitary block-diagonal matrices $(\tilde{V}, \tilde{W})$ such that

$$\bar{V} \;=\; W\tilde{V}, \qquad\qquad \bar{W} \;=\; V\tilde{W} \tag{A.17a}$$

Indeed, according to Autonne's uniqueness theorem ([15, Theorem 2.6.5, p.152]), $\tilde{V}$ and $\tilde{W}$ can be taken to have identical blocks except the last block corresponding to the diagonal zero-block in (A.10). Specifically suppose A has rank r and d distinct positive singular values $s_1 > s_2 > \cdots > s_d > 0$ with (algebraic) multiplicities $n_1, \cdots, n_d$. Then $r := \sum_{i=1}^{d} n_i \le n$. We can separate the diagonal of the $n \times n$ matrix Σ into $d+1$ diagonal blocks of diagonal submatrices $s_i I_{n_i}$ and 0_{n-r}:

$$\Sigma \;=\; \mathrm{diag}\left(s_1 I_{n_1}, \cdots, s_d I_{n_d}, 0_{n-r}\right) \tag{A.17b}$$

where I_k denotes the identity matrix of size k and 0_k denotes the $k \times k$ zero matrix. (If A is of full rank $r = n$ then the zero block 0_{n-r} is absent.) Then Autonne's uniqueness theorem ([15, Theorem 2.6.5, p.152]) implies that $A = V\Sigma W^{\mathsf{H}} = \bar{W}\Sigma\bar{V}^{\mathsf{H}}$ if and only if there are unitary matrices V_i of sizes n_i and V_{d+1}, W_{d+1} of size $n - r$ such that

$$\tilde{V} \;=\; \mathrm{diag}\left(V_1, \cdots, V_d, V_{d+1}\right), \qquad \tilde{W} \;=\; \mathrm{diag}\left(V_1, \cdots, V_d, W_{d+1}\right) \tag{A.17c}$$

and $\bar{V} = W\tilde{V}$, $\bar{W} = V\tilde{W}$. But $\tilde{V} = W^{\mathsf{H}}\bar{V} = \left(V^{\mathsf{H}}\bar{W}\right)^{\mathsf{T}} = \tilde{W}^{\mathsf{T}}$ and hence $V_i = V_i^{\mathsf{T}}$ are symmetric matrices for $i = 1, \ldots, d$.

Lemma A.10 then implies that there exist unitary symmetric matrices $R_i \in \mathbb{C}^{n_i \times n_i}$ such that $V_i = R_i^2$ for $i = 1, \cdots, d$. Substitute this and (A.17) into $A = \bar{W}\Sigma\bar{V}^{\mathsf{H}}$, we have $A = \bar{W}\Sigma V^{\mathsf{T}} = V\tilde{W}\Sigma V^{\mathsf{T}}$. But (taking $W_{d+1} := I_{n-r}$)

$$\tilde{W}\Sigma \;=\; \mathrm{diag}\left(R_1^2, \cdots, R_d^2, I_{n-r}\right) \cdot \mathrm{diag}\left(s_1 I_{n_1}, \ldots, s_d I_{n_d}, 0_{n-r}\right) \;=:\; R\Sigma R$$

where $R := \text{diag}\,(R_1, \cdots, R_d, I_{n-r})$. Hence

$$A \;=\; V(\tilde{W}\Sigma)V^{\mathsf{T}} \;=\; V\,(R\Sigma R)\,V^{\mathsf{T}} \;=\; \underbrace{(VR)}_{U}\;\Sigma\;\underbrace{(VR)^{\mathsf{T}}}_{U^{\mathsf{T}}}$$

where the last equality uses the symmetry of R. This completes the proof. □

A complex symmetric matrix $A \in \mathbb{C}^{n\times n}$ may or may not be normal. Complex symmetric matrices are useful for power systems because the admittance matrix Y (see Chapter 4.2) are complex symmetric, and generally not Hermitian. See Exercise 16.23 for a complex symmetric matrix that is not diagonalizable (and hence not normal). See Exercise 16.24 for a complex symmetric matrix that is normal and hence unitarily diagonalizable, and Exercise 4.3 for characterizations of symmetric and normal matrices.

A.7 Pseudo-inverse

Consider a matrix $A \in \mathbb{C}^{m\times n}$. Let $\text{null}(A)$ denote the *null space* (also called kernel) of A, i.e., $\text{null}(A) := \{x \in \mathbb{C}^n \,:\, Ax = 0\}$. Let $\text{range}(A)$ denote the *range space* (also called column space) of A, i.e., $\text{range}(A) := \{y \in \mathbb{C}^m \,:\, y = Ax \text{ for some } x \in \mathbb{C}^n\}$. In this subsection we treat A as a mapping from $\mathbb{C}^n$ to $\mathbb{C}^m$ and A^{H} a mapping from $\mathbb{C}^m$ to $\mathbb{C}^n$. Then $\text{null}(A)$ and $\text{range}(A^{\mathsf{H}})$ are linear spaces and they are orthogonal complements of each other because, if $x_1 \in \text{null}(A)$ and $x_2 \in \text{range}(A^{\mathsf{H}})$ so that $x_2 = A^{\mathsf{H}}y$ for some y, then

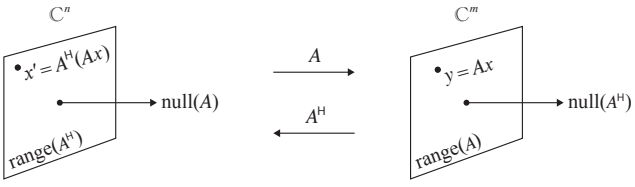
$$x_2^{\mathsf{H}} x_1 = y^{\mathsf{H}} A\, x_1 \;=\; 0$$

We denote this fact by the notation $\mathbb{C}^n = \text{range}(A^{\mathsf{H}}) \oplus \text{null}(A)$, as shown in the upper panel of Figure A.3(a). This implies

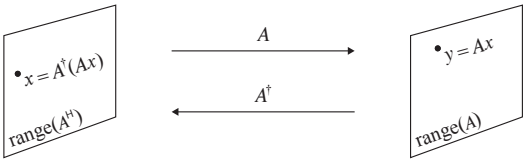
$$\dim(\text{range}(A^{\mathsf{H}})) \;+\; \dim(\text{null}(A)) = n \tag{A.18}$$

The rank of a matrix $A \in \mathbb{C}^{m\times n}$, denoted rank A, is the largest number of linearly independent columns of A, or equivalently the largest number of linearly independent rows of A. By definition rank $A = \dim(\text{range}(A))$. A square matrix $A \in \mathbb{C}^{n\times n}$ is called *nonsingular* if rank $A = n$; it is called *singular* if rank $A < n$. Some simple facts are collected in the following.

Theorem A.18. 1 For any $A \in \mathbb{C}^{m\times n}$, rank A = rank A^{H} = rank A^{T} = rank $\bar{A}$.
2 For any $A \in \mathbb{C}^{m\times n}$, rank $A \le \min\{m, n\}$.
3 If $A \in \mathbb{C}^{m\times m}$ and $C \in \mathbb{C}^{n\times n}$ are nonsingular, then for any $B \in \mathbb{C}^{m\times n}$, rank B = rank ABC, i.e., left or/and right multiplication by a nonsingular matrix does not change rank.



(a) Orthogonal decomposition of $\mathbb{C}^n$ and $\mathbb{C}^m$



(b) A and $A^{\dagger}$ are inverses between range(A^{H}) and range(A)

Figure A.3 Orthogonal decomposition of $\mathbb{C}^n$ and $\mathbb{C}^m$ and pseudo-inverse $A^{\dagger}$. For any $x \in \text{range}(A^H)$, $x = A^{\dagger}(Ax)$ which is generally different from $x' = A^{\mathsf{H}}(Ax)$.

4 For any $A \in \mathbb{C}^{m\times n}$, rank A + dim(null(A)) $= n$. This follows from substituting rank A^{H} = rank A into (A.18).

If we consider the matrix $A \in \mathbb{C}^{m\times n}$ as a mapping from $\mathbb{C}^n$ to $\mathbb{C}^m$ and restrict it to $A : \text{range}(A^{\mathsf{H}}) \to \text{range}(A)$, then A is surjective and injective (see Exercise A.10). Hence an inverse always exists from range$(A) \to$ range(A^{H}). We will denote this inverse by $A^{\dagger}$; see Figure A.3(b). Let $A = V\Sigma W^{\mathsf{H}}$ be its singular value decomposition and let rank $A = r \le \min\{m,n\}$. We will show that

$$A^{\dagger} = W\Sigma^{\dagger}V^{\mathsf{H}} \tag{A.19}$$

where $\Sigma^{\dagger}$ is a real diagonal $n\times m$ matrix of rank r obtained from the $m\times n$ diagonal matrix Σ by replacing the (positive) singular values σ_i by $1/\sigma_i$ and taking the transpose. When $r = m = n$, $\Sigma^{\dagger}$ =diag$\left(\frac{1}{\sigma_1},\ldots,\frac{1}{\sigma_n}\right) = \Sigma^{-1}$ so that $A^{\dagger} = A^{-1}$ since

$$A^{\dagger}A = \left(W\Sigma^{-1}V^{\mathsf{H}}\right)\left(V\Sigma W^{\mathsf{H}}\right) \;=\; \mathbb{I}_n$$

If $x \in \text{range}(A^{\mathsf{H}})$ then $A^{\dagger}(Ax) = W(\Sigma^{\dagger}\Sigma)W^{\mathsf{H}}x = x$ since $A^{\dagger}$ is the inverse of A between range(A^{H}) and range(A). In contrast $A^{\mathsf{H}}(Ax) = W(\Sigma^{\mathsf{T}}\Sigma)W^{\mathsf{H}}x = x'$ which is also in range(A^{H}) but generally different from x; see Figure A.3(b).

For a general $x \in \mathbb{R}^n$, $A^{\dagger}A \neq \mathbb{I}_n$ but the next result shows that $A^{\dagger}A$ equals $\mathbb{I}_n$ plus null(A). Specifically, let $A \in \mathbb{C}^{m\times n}$ with rank $A = r \le \min\{m,n\}$. Let $A = V\Sigma W^{\mathsf{H}}$ be its

singular value decomposition. Decompose the various matrices such that

$$\Sigma = \begin{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix} & 0 \\ 0 & 0 \end{bmatrix} =: \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix}, \qquad V =: \begin{bmatrix} V_r & V_{m-r} \end{bmatrix}, \qquad W =: \begin{bmatrix} W_r & W_{n-r} \end{bmatrix}$$

where Σ_r is $r \times r$ diagonal matrix, the matrices $V_r \in \mathbb{C}^{m\times r}$ and $W_r \in \mathbb{C}^{n\times r}$ consist of the first r columns of V and W respectively, and the matrices $V_{m-r} \in \mathbb{C}^{m\times(m-r)}$ and $W_{n-r} \in \mathbb{C}^{n\times(n-r)}$ consist of the remaining columns of V and W respectively. Then

$$\begin{aligned} A &= \begin{bmatrix} V_r & V_{m-r} \end{bmatrix} \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} W_r^{\mathsf{H}} \\ W_{n-r}^{\mathsf{H}} \end{bmatrix} = V_r \Sigma_r W_r^{\mathsf{H}} \\ A^\dagger &= \begin{bmatrix} W_r & W_{n-r} \end{bmatrix} \begin{bmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_r^{\mathsf{H}} \\ V_{m-r}^{\mathsf{H}} \end{bmatrix} = W_r \Sigma_r^{-1} V_r^{\mathsf{H}} \end{aligned}$$

and $A^{\mathsf{H}} = W_r \Sigma_r V_r^{\mathsf{H}}$. Hence the range spaces of $A, A^\dagger, A^{\mathsf{H}}$ depend only on the nonzero singular values and the first r columns of V and W. The remaining columns V_{m-r}, W_{n-r} span their null spaces and can be interpreted as a measure of how different the pseudo-inverse $A^\dagger$ is from an inverse, as the following theorem shows. The theorem is illustrated in Figures A.4.

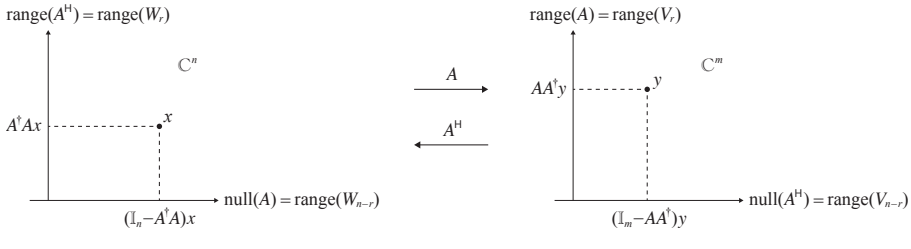


Figure A.4 Orthogonal decomposition of $\mathbb{C}^n$ and $\mathbb{C}^m$ using singular value decomposition of A.

Theorem A.19. With the notations above,

1 $A^\dagger := W\Sigma^\dagger V^{\mathsf{H}}$ satisfies ($\mathbb{I}_n$ denotes the $n \times n$ identity matrix)

$$\begin{aligned} A^\dagger A &= \mathbb{I}_n - W_{n-r} W_{n-r}^{\mathsf{H}} = W_r W_r^{\mathsf{H}} \\ AA^\dagger &= \mathbb{I}_m - V_{m-r} V_{m-r}^{\mathsf{H}} = V_r V_r^{\mathsf{H}} \end{aligned}$$

2 $\mathrm{null}(A) = \mathrm{range}(W_{n-r})$ and $\mathrm{range}(A^{\mathsf{H}}) = \mathrm{range}(W_r)$.
3 $\mathrm{null}(A^{\mathsf{H}}) = \mathrm{range}(V_{m-r}) = \mathrm{null}(A^\dagger)$ and $\mathrm{range}(A) = \mathrm{range}(V_r)$.
4 $A^\dagger A$ is the orthogonal projection of $x \in \mathbb{C}^n$ onto $\mathrm{range}(A^{\mathsf{H}})$. $\mathbb{I}_n - A^\dagger A$ is the orthogonal projection of $x \in \mathbb{C}^n$ onto $\mathrm{null}(A)$.
5 Similarly $AA^\dagger$ is the orthogonal projection of $y \in \mathbb{C}^m$ on to $\mathrm{range}(A)$ and $\mathbb{I}_m - AA^\dagger$ is the orthogonal projection of $y \in \mathbb{C}^m$ onto $\mathrm{null}(A^{\mathsf{H}})$.
6 $AA^\dagger A = A$, $A^\dagger AA^\dagger = A^\dagger$, and $A^{\mathsf{H}} AA^\dagger = A^{\mathsf{H}}$.

Proof We have

$$A^\dagger A = \begin{bmatrix} W_r & W_{n-r} \end{bmatrix} \begin{bmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_r^{\mathsf{H}} \\ V_{n-r}^{\mathsf{H}} \end{bmatrix} \cdot \begin{bmatrix} V_r & V_{n-r} \end{bmatrix} \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} W_r^{\mathsf{H}} \\ W_{n-r}^{\mathsf{H}} \end{bmatrix}$$
$$= \begin{bmatrix} W_r & W_{n-r} \end{bmatrix} \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} W_r^{\mathsf{H}} \\ W_{n-r}^{\mathsf{H}} \end{bmatrix} \;=\; W_r W_r^{\mathsf{H}}$$

Even though $W^{\mathsf{H}} = W^{-1}$, W_r^{H} is not the inverse of W_r (unless $r = n \neq m$) since W_r is not even square. Since

$$WW^{\mathsf{H}} = \begin{bmatrix} W_r & W_{n-r} \end{bmatrix} \begin{bmatrix} W_r^{\mathsf{H}} \\ W_{n-r}^{\mathsf{H}} \end{bmatrix} \;=\; W_r W_r^{\mathsf{H}} + W_{n-r} W_{n-r}^{\mathsf{H}} \;=\; \mathbb{I}_n$$

we have

$$A^\dagger A = \mathbb{I}_n \;-\; W_{n-r} W_{n-r}^{\mathsf{H}}$$

Similarly $AA^\dagger = \mathbb{I}_m \;-\; V_{m-r} V_{m-r}^{\mathsf{H}}$.

To show that $\mathrm{null}(A) = \mathrm{range}(W_{n-r})$ consider any $x \in \mathbb{C}^n$. Since columns of W are an orthonormal basis of $\mathbb{C}^n$ we can write $x = \sum_j b_j w_j$ for some $b_j \in \mathbb{C}$ where w_j are columns of W. Then

$$Ax = V\Sigma W^{\mathsf{H}} \sum_j b_j w_j \;=\; V\Sigma \sum_j b_j \begin{bmatrix} w_1^{\mathsf{H}} w_j \\ \vdots \\ w_n^{\mathsf{H}} w_j \end{bmatrix} \;=\; V\Sigma \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} \;=\; V \begin{bmatrix} \sigma_1 b_1 \\ \vdots \\ \sigma_r b_r \\ 0_{n-r} \end{bmatrix}$$

where 0_{n-r} is the zero vector of size $n-r$. Since V is nonsingular and $\sigma_j > 0$, $Ax = 0$ if and only if $b_1 = \cdots = b_r = 0$. Hence $\mathrm{null}(A) = \mathrm{range}(W_{n-r})$ if and only if $x \in \mathrm{range}(W_{n-r})$. That $\mathrm{range}(A^{\mathsf{H}}) = \mathrm{range}(W_r)$ follows from $A^{\mathsf{H}} = W\Sigma^{\mathsf{T}} V^{\mathsf{H}} = W_r \Sigma_r V_r^{\mathsf{H}}$.

The proof of $\mathrm{null}(A^\dagger) = \mathrm{range}(V_{m-r})$ follows the same argument and is presented in the matrix notation as follows. Any $y \in \mathbb{C}^m$ can be written in terms of the columns of V, i.e., $y = V_r b_r + V_{m-r} b_{m-r}$ for some b_r, b_{m-r}. Then

$$A^\dagger y = W_r \Sigma_r^{-1} V_r^{\mathsf{H}} (V_r b_r + V_{m-r} b_{m-r}) \;=\; W_r \Sigma_r^{-1} V_r^{\mathsf{H}} V_r b_r$$

since $V_r^{\mathsf{H}} V_{m-r} = 0_{r\times(m-r)}$. Hence $A^\dagger y = 0$ if and only if $b_r = 0$ and $y = V_{m-r} b_{m-r}$. This means $\mathrm{null}(A^\dagger) = \mathrm{range}(V_{m-r})$. Since $A^{\mathsf{H}} = W_r \Sigma_r V_r^{\mathsf{H}}$ the same argument shows that $\mathrm{null}(A^{\mathsf{H}}) = \mathrm{range}(V_{m-r})$.

The remaining assertions follow from parts 1, 2, 3. For example

$$AA^\dagger A = A\left(\mathbb{I}_n \;-\; W_{n-r} W_{n-r}^{\mathsf{H}}\right) \;=\; A \;-\; AW_{n-r} W_{n-r}^{\mathsf{H}} \;=\; A$$

Similarly $A^\dagger AA^\dagger = A^\dagger$, and $A^{\mathsf{H}} AA^\dagger = A^{\mathsf{H}}$. □

We remark on some implications of Theorem A.19.

Remark A.2 ($Ax = b$)**.** 1 Theorem A.19.1 says that $\mathbb{I}_n = A^\dagger A + W_{n-r} W_{n-r}^{\mathsf{H}}$ and hence for any $x \in \mathbb{R}^n$,

$$x = \underbrace{A^\dagger A x}_{\text{projection onto range}(A^{\mathsf{H}})} + \underbrace{W_{n-r}\left(W_{n-r}^{\mathsf{H}} x\right)}_{\text{projection onto range}(W_{n-r})} \tag{A.20}$$

See Figure A.4. Similarly for $y = Ax \in \mathbb{R}^m$. Therefore

$$\mathbb{R}^n = A^\dagger (A\mathbb{R}^n) + \text{range}(W_{n-r}) \tag{A.21a}$$

$$\mathbb{R}^m = A\left(A^\dagger \mathbb{R}^m\right) + \text{range}(V_{m-r}) \tag{A.21b}$$

2 The theorem implies that $A^\dagger$ in (A.19) and A are inverses of each other when restricted to range(A^{H}) and range(A) (see Exercise A.11). Therefore, even though (V, W) in the singular value decomposition are generally not unique, $A^\dagger$ is uniquely defined. Treated as a mapping from $\mathbb{C}^m$ to $\mathbb{C}^n$, $A^\dagger$ is called a *pseudo-inverse* of A.

3 There is a solution x for $Ax = b$ if and only if b is in range(A) or equivalently b is orthogonal to null (A^{H}), in which case the set of solutions is given by

$$x = A^\dagger b + w, \qquad w \in \text{null}(A) = \text{range}(W_{n-r})$$

Moreover $A^\dagger b$ is the solution to $Ax = b$ with the smallest Euclidean norm $\|x\|_2 = \|A^\dagger b\|_2 + \|w\|_2$.

4 Consider $Ax = b$ when b is not in range(A) and therefore there is no x that satisfies this equation. The theorem says that $\hat{x} = A^\dagger b$ is a 'best estimate' of x from b in that $A\hat{x}$ equals the projection of b onto range(A) and the estimation error $b - A\hat{x} = (\mathbb{I}_m - AA^\dagger)b$ is the projection of b onto null(A^{H}). This achieves the minimum estimation error under the Euclidan norm; see Exercise A.14.

5 Theorem A.19.6 is easy to understand given Theorems A.19.4 and A.19.5. Consider any vector $y \in \mathbb{C}^m$. The operation $AA^\dagger$ removes y's component in the null space of A^{H}, i.e., $AA^\dagger y$ projects y to range(A). It is then mapped under $A^\dagger$ into $\mathbb{C}^n$ to $A^\dagger(AA^\dagger y)$. Since $AA^\dagger y$ is already in range(A) over which $A^\dagger$ is an inverse of A, this operation should be the same as $A^\dagger$, i.e., $A^\dagger AA^\dagger y = A^\dagger y$ for all y. Similarly the projection operation $A^\dagger A$ to range(A^{H}) followed by the mapping A is the same operation as the mapping A.

For general matrix $A \in \mathbb{C}^{m\times n}$, its pseudo-inverse is given in terms of its singular value decomposition by (A.19). For special matrices the next result provide some explicit formulae.

Corollary A.20. Consider a matrix $A \in \mathbb{C}^{m\times n}$ with rank $A = r \le \min\{m, n\}$. Let $A = V\Sigma W^{\mathsf{H}}$ be its singular value decomposition and $A^\dagger = W\Sigma^\dagger V^{\mathsf{H}}$ be its pseudo-inverse.

1 If $m = n$ and A is positive semidefinite then $A + V_{n-r} V_{n-r}^{\mathsf{H}}$ is invertible and

$$A^\dagger = \left(A + V_{n-r} V_{n-r}^{\mathsf{H}}\right)^{-1} - V_{n-r} V_{n-r}^{\mathsf{H}}$$

2 If $r = m \leq n$ then $A^\dagger = A^{\mathsf{H}} \left(AA^{\mathsf{H}}\right)^{-1}$.
3 If $r = n \leq m$ then $A^\dagger = \left(A^{\mathsf{H}}A\right)^{-1} A^{\mathsf{H}}$.
4 If $r = m = n$ then $A^\dagger = A^{-1}$.

Proof Since A is positive semidefinite its singular value decomposition coincides with its spectral decomposition according to Theorem A.16.3, so

$$A = V\Sigma W^{\mathsf{H}} \;=\; V\Lambda V^{\mathsf{H}} \;=\; V_r \Lambda_r V_r^{\mathsf{H}}$$

where V is a unitary matrix whose columns are orthonormal eigenvectors of A, $\Lambda := \mathrm{diag}(\lambda_i)$ is the diagonal matrix of eigenvalues

$$\lambda_1 \;\geq\; \cdots \;\geq\; \lambda_r \;>\; 0 \;=\; \lambda_{r+1} \;=\; \cdots \;=\; \lambda_n$$

and matrices are decomposed as before:

$$\Lambda \;=:\; \begin{bmatrix} \Lambda_r & 0 \\ 0 & 0 \end{bmatrix}, \qquad V \;=:\; \begin{bmatrix} V_r & V_{n-r} \end{bmatrix}, \qquad x \;=:\; \begin{bmatrix} x_r \\ x_{n-r} \end{bmatrix} \in \mathbb{C}^n$$

To show that $A + V_{n-r}V_{n-r}^{\mathsf{H}}$ is invertible consider any $x \in \mathbb{C}^n$ in the null space of A expressed in terms of the basis V as $x = Va =: V_r a_r + V_{n-r} a_{n-r}$. We have

$$(A + V_{n-r}V_{n-r}^{\mathsf{H}})\, x = \left(V_r \Lambda_r V_r^{\mathsf{H}} + V_{n-r}V_{n-r}^{\mathsf{H}}\right)(V_r a_r + V_{n-r} a_{n-r}) \;=\; V_r \Lambda_r a_r \;+\; V_{n-r} a_{n-r}$$

where we have used $V_r^{\mathsf{H}} V_{n-r} = 0$. Hence

$$(A + V_{n-r}V_{n-r}^{\mathsf{H}})\, x = \begin{bmatrix} V_r & V_{n-r} \end{bmatrix} \begin{bmatrix} \Lambda_r a_r \\ a_{n-r} \end{bmatrix} \;=\; V \begin{bmatrix} \Lambda_r a_r \\ a_{n-r} \end{bmatrix}$$

Since V and Λ_r are nonsingular, $(A + V_{n-r}V_{n-r}^{\mathsf{H}})x = 0$ if and only if $a = 0$, proving the nonsingularity of $A + V_{n-r}V_{n-r}^{\mathsf{H}}$.

To show that $A^\dagger = \left(A + V_{n-r}V_{n-r}^{\mathsf{H}}\right)^{-1} - V_{n-r}V_{n-r}^{\mathsf{H}}$ we will prove that $A^\dagger + V_{n-r}V_{n-r}^{\mathsf{H}}$ is the inverse of $A + V_{n-r}V_{n-r}^{\mathsf{H}}$. We have (using again $V_r V_{n-r}^{\mathsf{H}} = 0$)

$$\begin{aligned} \left(A^\dagger + V_{n-r}V_{n-r}^{\mathsf{H}}\right)\left(A + V_{n-r}V_{n-r}^{\mathsf{H}}\right) &= \left(V_r \Lambda_r^{-1} V_r^{\mathsf{H}} + V_{n-r}V_{n-r}^{\mathsf{H}}\right)\left(V_r \Lambda_r V_r^{\mathsf{H}} + V_{n-r}V_{n-r}^{\mathsf{H}}\right) \\ &= V_r V_r^{\mathsf{H}} \;+\; V_{n-r}V_{n-r}^{\mathsf{H}} \;=\; VV^{\mathsf{H}} \;=\; \mathbb{I}_n \end{aligned}$$

as desired.

If If $r = m \leq n$ then $V_r = V$ and

$$\Sigma \;=:\; \begin{bmatrix} \Sigma_r & 0 \end{bmatrix}, \qquad W \;=:\; \begin{bmatrix} W_r & W_{n-r} \end{bmatrix}$$

Then $A = V\Sigma W^{\mathsf{H}} = V\Sigma_r W_r^{\mathsf{H}}$ and hence $AA^{\mathsf{H}} = \left(V\Sigma_r W_r^{\mathsf{H}}\right)\left(W_r \Sigma_r V^{\mathsf{H}}\right) = V\Sigma_r^2 V^{\mathsf{H}}$ is invertible since $W_r^{\mathsf{H}} W_r = I_r$. Since V is unitary we have $\left(AA^{\mathsf{H}}\right)^{-1} = V\Sigma_r^{-2}V^{\mathsf{H}}$. Hence

$$A^{\mathsf{H}}\left(AA^{\mathsf{H}}\right)^{-1} = \left(W_r \Sigma_r V^{\mathsf{H}}\right)\left(V\Sigma_r^{-2}V^{\mathsf{H}}\right) \;=\; W_r \Sigma_r^{-1} V^{\mathsf{H}} \;=\; W\Sigma^\dagger V^{\mathsf{H}} \;=\; A^\dagger$$

The case of $r = n \leq m$ is similarly proved in Exercise A.12. If $r = m = n$ then $\Sigma^\dagger = \Sigma^{-1}$ so that $A^\dagger = A^{-1}$ since $A^\dagger A = \left(W\Sigma^{-1}V^{\mathsf{H}}\right)\left(V\Sigma W^{\mathsf{H}}\right) = \mathbb{I}_n$.

□

Consider a partitioned matrix $A = [B\ C]$. In general $A^\dagger \neq \begin{bmatrix} B^\dagger \\ C^\dagger \end{bmatrix}$. [2] Several expressions for $A^\dagger$ in terms of $B^\dagger$ and $C^\dagger$ are derived in [177] under various necessary and sufficient conditions. The particularly simple case is the following result from [177, Corollary 1.4].

Lemma A.21. Suppose $A = [B\ C]$. Then

$$A^\dagger = \begin{bmatrix} B^\dagger \\ C^\dagger \end{bmatrix}$$

if and only if $(\mathbb{I} - BB^\dagger)C = C$ (i.e., if and only if C is in $\text{null}(B^\mathsf{H})$).

A.8 Norms and inequalities

A.8.1 Vector norms

This subsection mostly follows [15, Chapter 5].

Definition A.4 (Normed linear space)**.** Let V be a vector space over the field F with $F = \mathbb{R}$ or $\mathbb{C}$. A function $\|\cdot\| : V \to \mathbb{R}$ is a *norm*, or *vector norm*, on V if, for all $x, y \in V$ and all $c \in F$,

1. *Positivity*: $\|x\| \geq 0$ and $\|x| = 0$ if and only if $x = 0$.
2. *Homogeneity*: $\|cx\| = |c|\,\|x\|$.
3. *Triangular inequality*: $\|x + y\| \leq \|x\| + \|y\|$.

The real or complex vector space together with a norm $(V, \|\cdot\|)$ is called a *normed linear space* or *normed vector space*. □

Examples of vector norms on $V = \mathbb{C}^n$ include: for any $x \in \mathbb{C}^n$,

- *Sum norm (l_1 norm)*: $\|x\|_1 := \sum_i |x_i|$.
- *Euclidean norm (l_2 norm)*: $\|x\|_2 := \sqrt{\sum_i |x_i|^2}$.
- *Max norm (l_∞ norm)*:$\|x\|_\infty := \max_i |x_i|$.
- *l_p norm*: $\|x\|_p := (\sum_i |x_i|^p)^{1/p}$, $p \geq 1$.

[2] Let the singular value decompositions of B and C be $B = V_1\Sigma_1 W_1^\mathsf{H}$ and $C = V_2\Sigma W_2^\mathsf{H}$. We can write

$$A = \begin{bmatrix} V_1 & V_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} W_1^\mathsf{H} & 0 \\ 0 & W_2^\mathsf{H} \end{bmatrix}$$

However $(VMW^\mathsf{H})^\dagger = WM^\dagger V^\mathsf{H}$ only if V and W are unitary [177, Lemma 1]. The matrix $[V_1\ V_2]$ is not unitary.

It can be shown that $\|x\|_\infty = \lim_{p\to\infty} \|x\|_p$ for all $x \in \mathbb{C}^n$. We therefore often define l_p norms for $p \in [1,\infty]$. The Euclidean norm, and positive scalar multiples of the Euclidean norm, are the only norms on $\mathbb{C}^n$ that are *unitarily invariant*: $\|Ux\|_2 = \|x\|_2$ for any $x \in \mathbb{C}^n$ and any unitary matrix $U \in \mathbb{C}^{n\times n}$ (Exercise A.17). The unit balls $B := \{x \in \mathbb{R}^2 : \|x\| \le 1\}$ for l_1, l_2 and l_∞ norms are shown in Figure A.5.

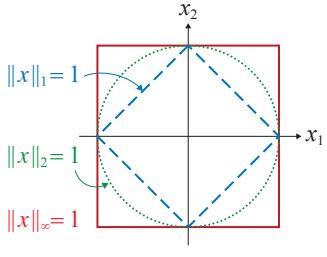


Figure A.5 The boundaries of unit balls for l_1, l_2 and l_∞ norms.

An example of infinite dimensional normed vector spaces is the set $C[a,b]$ of all continuous real or complex-valued functions $f : [a,b] \to \mathbb{R}$ or $f : [a,b] \to \mathbb{C}$ on the real interval $[a,b]$. The L_p norms on $C[a,b]$ are

- L_1 *norm*: $\|f\|_1 := \int_a^b |f(t)|dt$.
- L_2 *norm*: $\|f\|_2 := \sqrt{\int_a^b |f(t)|^2 dt}$.
- L_p *norm*: $\|f\|_p := \left(\int_a^b |f(t)|^p dt\right)^{1/p}$, $p \ge 1$.
- L_∞ *norm*: $\|f\|_\infty := \max\{|f(x)| : x \in [a,b]\}$.

There are two important properties of finite dimensional real or complex vector spaces V (i.e., $F = \mathbb{R}$ or $\mathbb{C}$) that do not necessarily hold for infinite dimensional vector spaces. First all norms are equivalent in the sense that, given two norms $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ on a finite dimensional vector space V, there exist c_m, c_M such that (e.g., [15, Corollary 5.4.5, p.327])

$$c_m \|x\|_\alpha \le \|x\|_\beta \le c_M \|x\|_\alpha, \qquad x \in V \tag{A.22}$$

This means that if a sequence $\{x_i\} \subseteq V$ converges in some norm, it converges in all norms. For l_p norms the best bounds are [15, Problem 5.4.P3, p.333]: for $1 \le p_1 < p_2 < \infty$,

$$\|x\|_{p_2} \le \|x\|_{p_1} \le n^{\left(\frac{1}{p_1}-\frac{1}{p_2}\right)} \|x\|_{p_2}$$

For example $\|x\|_2 \le \|x\|_1 \le \sqrt{n}\|x\|_2$, $\|x\|_\infty \le \|x\|_1 \le n\|x\|_\infty$, $\|x\|_\infty \le \|x\|_2 \le \sqrt{n}\|x\|_\infty$ (see Figure A.5). In contrast, for an infinite dimensional vector space such as $C[a,b]$, a sequence $\{f_k\}$ of functions in $C[a,b]$ may converge under the L_1 norm, remains bounded under L_2 norm, but diverge under the L_∞ norm (unbounded $\|f_k\|_\infty$).

Second a sequence $\{x_i\} \subseteq V$ converges to a vector in a finite dimensional vector space V if and only if it is a *Cauchy sequence*, i.e., for any $\epsilon > 0$ there exists a positive integer $N(\epsilon)$ such that $\|x_i - x_j\| \leq \epsilon$ for any $i, j \geq N(\epsilon)$. A normed linear space V is said to be *complete* with respect to its norm $\|\cdot\|$ if every sequence in V that is a Cauchy sequence with respect to $\|\cdot\|$ converges to a point in V. Therefore all finite dimensional real or complex vector spaces are complete with respect to any norm, but infinite dimensional normed vector spaces, such as $C[a,b]$ with the L_1 norm, may not be complete.

Definition A.5 (Inner product space)**.** Let V be a (finite or infinite dimensional) vector space over the field F with $F = \mathbb{R}$ or $\mathbb{C}$. A function $\langle\cdot,\cdot\rangle : V \times V \to \mathbb{F}$ is an *inner product* if, for all $x, y, z \in V$ and all $c \in F$,

1. *Positivity*: $\langle x,x\rangle \geq 0$ and $\langle x,x\rangle = 0$ if and only if $x = 0$.
2. *Additivity*: $\langle x+y,z\rangle = \langle x,z\rangle + \langle y,z\rangle$.
3. *Homogeneity*: $\langle cx,y\rangle = c\langle x,y\rangle$.
4. *Hermitian property*: $\langle x,y\rangle = \overline{\langle x,y\rangle}$.

where $\overline{a}$ denotes the complex conjugate of $a \in F$. The real or complex vector space together with an inner product $(V, \langle\cdot,\cdot\rangle)$ is called an *inner product space*. □

Note that regardless of $F = \mathbb{R}$ or $\mathbb{C}$, a norm in Definition A.4 takes value in $\mathbb{R}$ whereas an inner product in Definition A.5 takes value in F. Implicit in the nonnegativity property is that, while $\langle x,y\rangle \in F$, $\langle x,x\rangle \in \mathbb{R}$. The function defined on $\mathbb{C}^n$ by $\langle x,y\rangle := x^{\mathsf{H}}y \in F := \mathbb{C}$ is an inner product called the Euclidean inner product. Let $M \in \mathbb{F}^{n\times n}$ be a positive definite matrix and define the function $\langle x,y\rangle_M := y^{\mathsf{H}}Mx$. Then $\langle\cdot,\cdot\rangle_M$ is also an inner product.

If $\langle\cdot,\cdot\rangle$ is an inner product on a real or complex vector space V, then the function $\|\cdot\| : V \to [0,\infty)$ defined by $\|x\| := \langle x,x\rangle^{1/2}$ is a norm on V. Such a norm is said to be *derived from an inner product*. The Euclidean norm $\|\cdot\|_2$ is a norm derived from the Euclidean inner product. An inner product space is therefore also a normed linear space with its derived norm. Not all norms are derived from an inner product, e.g., $\|\cdot\|_1$, $\|\cdot\|_\infty$ are not derived norms.

Inner products are defined for infinite dimensional vector spaces as well. For example an inner product on the vector space $C[a,b]$ of all continuous real or complex-valued functions on the real interval $[a,b]$ is

$$\langle f,g\rangle := \int_a^b f(t)\overline{g(t)}dt, \qquad f,g \in C[a,b]$$

The L_2 norm $\|f\|_2 := \sqrt{\int_a^b |f(t)|^2 dt}$ defined above is derived from the inner product $\langle f,f\rangle$.

A.8.2 Cauchy-Schwarz inequality, Hölder's inequality, dual norm

We now present an extremely useful inequality, the Cauchy-Schwarz inequality, and two generalizations.

Cauchy-Schwarz and Hölder's inequalities.

The Cauchy-Schwarz inequality is an important property of all inner products on any finite or infinite dimensional vector space. The inequality holds regardless of whether the norm on the vector space is derived from the inner product. Hence $\langle x,x\rangle$, $\langle y,y\rangle$ on the right-hand side of (A.23) may not be the squared norms on V.

Theorem A.22 (Cauchy-Schwarz inequality)**.** Let $(V,\langle\cdot,\cdot\rangle)$ be an inner product space over a field F with $F=\mathbb{R}$ or $\mathbb{C}$. Then

$$|\langle x,y\rangle|^2 \;\le\; \langle x,x\rangle\,\langle y,y\rangle, \qquad\qquad x,y\in V \tag{A.23}$$

with equality if and only if $x=ay$ for some $a\in F$ (i.e., x and y are linearly dependent).

Proof To prove the Cauchy-Schwarz inequality suppose without loss of generality $y\neq 0$ (the inequality holds if $x=y=0$). Let $z:=\langle y,y\rangle x-\langle x,y\rangle y$. Then, since $\langle a_1u_1+a_2u_2,b_1v_1+b_2v_2\rangle = a_1\overline{b}_1\langle u_1,v_1\rangle+a_1\overline{b}_2\langle u_1,v_2\rangle+a_2\overline{b}_1\langle u_2,v_1\rangle+a_2\overline{b}_2\langle u_2,v_2\rangle$,

$$\begin{aligned} 0 \;&\le\; \langle z,z\rangle \;=\; \langle\langle y,y\rangle x-\langle x,y\rangle y,\; \langle y,y\rangle x-\langle x,y\rangle y\rangle \\ &=\; \langle y,y\rangle^2\langle x,x\rangle-\langle x,y\rangle\overline{\langle x,y\rangle}\langle y,x\rangle \;=\; \langle y,y\rangle\Big(\langle x,x\rangle\langle y,y\rangle-|\langle x,y\rangle|^2\Big) \end{aligned}$$

which implies the inequality since $\langle y,y\rangle>0$. □

Cauchy-Schwarz inequality has numerous applications. One example is the following bounds on samples in terms of their sample mean and standard deviation. Let $x_1,\ldots,x_n$ be n given real numbers with sample mean μ and sample standard deviation σ defined by:

$$\mu \;:=\; \frac{1}{n}\sum_i x_i, \qquad\qquad \sigma \;:=\; \left(\frac{1}{n}\sum_i (x_i-\mu)^2\right)^{1/2}$$

It can then be shown that (Exercise A.18)

$$\mu-\sigma\sqrt{n-1} \;\le\; x_i \;\le\; \mu+\sigma\sqrt{n-1}, \qquad\qquad i=1,\ldots,n$$

with equality for some i if and only if $x_p=x_q$ for all $p,q\neq i$.

Hölder's inequalities.

A generalization of the Cauchy-Schwarz inequality is Hölder's inequality. Hölder's inequality holds for general L^p spaces (the vector space of measurable functions f for which its L_p norm is finite), but we will restrict ourselves to $V=\mathbb{R}^n$ or $\mathbb{C}^n$ with l_p norms.

Theorem A.23 (Hölder's inequality)**.** Consider the vector space $V = F^n$ with $F = \mathbb{R}$ or $\mathbb{C}$ with l_p norms, $p \in [1, \infty]$. Then for any $p, q \geq 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$ (with the interpretation that if $p = 1$ then $q = \infty$)

$$\sum_{i=1}^{n} |x_i y_i| \;\leq\; \|x\|_p \, \|y\|_q \,, \qquad\qquad x,\; y \in V \tag{A.24}$$

with equality if and only if $x^p := (x_i^p, i = 1, \ldots, n)$ and $y^q := (y_i^q, i = 1, \ldots, n)$ are linearly dependent, i.e., $x^p = a y^q$ for some scalar $a \in F$.

The theorem can be proved by applying the following property to the convex function $f(x) = x^p$ for $p > 1$: for all $\alpha_i \geq 0$, $\sum_{i=1}^n \alpha_i = 1$, for all x_i,

$$f\left(\sum_{i=1}^{n} \alpha_i x_i \right) \;\leq\; \sum_{i=1}^{n} \alpha_i f(x_i)$$

Setting $p = q = 2$ leads to the Cauchy-Schwarz inequality

$$\left| x^{\mathsf{H}} y \right| \;\leq\; \sum_{i=1}^{n} |x_i y_i| \;\leq\; \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2} \left(\sum_{i=1}^{n} y_i^2 \right)^{1/2} \;=\; \|x\|_2 \, \|y\|_2, \qquad x,\; y \in V$$

with equality if and only if the vectors x and y are linearly dependent ($x^p = a y^q \Leftrightarrow x = a^{1/p} y^{q/p}$). Note that this inequality is weaker than Hölder's inequality, though the Cauchy-Schwarz inequality holds for general inner products on arbitrary vector spaces with arbitrary norms.

Dual norm.

Another generalization of the Cauchy-Schwarz inequality holds with dual norm, as we define now. Consider any norm $\|\cdot\|$ on the vector space $V = F^n$ with $F = \mathbb{R}$ or $\mathbb{C}$. Define its *dual norm* $\|\cdot\|_*$ by: for any $x \in F^n$

$$\|x\|_* := \max_{y : \|y\| = 1} \operatorname{Re} x^{\mathsf{H}} y \;=\; \max_{y : \|y\| = 1} \left| x^{\mathsf{H}} y \right| \tag{A.25}$$

The maximization is attained since inner product is continuous and the feasible set is compact. (If we think of x^{H} as an $1 \times n$ matrix then $\|x\|_*$ is the matrix norm induced by the general vector norm $\|\cdot\|$ on $\mathbb{F}^n$; see below.)

A very useful inequality is

$$\operatorname{Re} x^{\mathsf{H}} y \;\leq\; \left| x^{\mathsf{H}} y \right| \;\leq\; \|x\| \, \|y\|_* \qquad\qquad \forall x, y \in \mathbb{F}^n \tag{A.26}$$

which follows directly from the definition of the dual norm. It says that the absolute inner product of any two vectors are upper bounded by the product of the norm of one of the vectors and its dual norm of the other vector. For the Euclidean norm $\|\cdot\|_2$ this is the Cauchy-Schwarz inequality, but (A.26) holds for *any* norm. Comparing this with Hölder's inequality (A.24), the left-hand side of (A.26) is smaller than that of (A.24), $\left| x^{\mathsf{H}} y \right| \leq \sum_i |x_i y_i|$. The norms on the right-hand side of (A.26) are not restricted to l_p

norms as those in (A.24) are. Indeed we now use Hölder's inequality to show that l_p and l_q norms are the dual of each other if $1/p + 1/q = 1$, and hence $\|x\|\,\|y\|_*$ reduces to the norms in Hölder's inequality if $\|\cdot\|$ is an l_p norm.

To simplify exposition we allow p, q with $1/p + 1/q = 1$ to take values in $[1, \infty]$ with the interpretation that if $p = 1$ then $q := \infty$.

Lemma A.24. Let $p, q \in [1, \infty]$ and $1/p + 1/q = 1$. The l_p norm and the l_q norm are dual of each other.

Proof We prove the case of $1 < p < \infty$; the case of $p = 1$ or $p = \infty$ follows a similar idea. Fix a pair $1 < p, q < \infty$ with $1/p + 1/q = 1$. Hölder's inequality implies, for all $x \in F^n$,

$$\|x\|_q \geq \max_{y:\|y\|_p=1} \sum_i |x_i y_i| \quad \geq \quad \max_{y:\|y\|_p=1} \left|x^{\mathsf{H}} y\right| \quad = \quad \|x\|_*$$

Therefore $\|x\|_q \geq \|x\|_*$, the dual norm of $\|\cdot\|_p$. To prove the reverse inequality we have from (A.26)

$$\|x\|_* \geq \left(\|y\|_p\right)^{-1} \left|x^{\mathsf{H}} y\right| \quad = \quad \left(\sum_i |y_i|^p\right)^{-1/p} \left|\sum_i \overline{x}_i y_i\right|, \qquad \forall y \in \mathbb{F}^n$$

Choose

$$y_i := |x_i|^{q/p} \, \frac{x_i}{|x_i|}$$

so that the inequality becomes (using $q = 1 + \frac{q}{p}$)

$$\|x\|_* \geq \left(\sum_i |x_i|^q\right)^{-1/p} \sum_i |x_i|^{1+q/p} \quad = \quad \left(\sum_i |x_i|^q\right)^{\frac{1}{q}} \quad = \quad \|x\|_q$$

Hence $\|x\|_* = \|x\|_q$ when $\|\cdot\| = \|\cdot\|_p$. □

In light of Lemma A.24, examples of (A.26) include:

$$\begin{aligned}
&\left|x^{\mathsf{H}} y\right| \leq \|x\|_p \, \|y\|_q \qquad \left(p^{-1} + q^{-1} = 1\right) \\
&\left|x^{\mathsf{H}} y\right| \leq \|x\|_2 \, \|y\|_2 \qquad (p = q = 2, \text{Cauchy-Schwarz inequality}) \\
&\|x\|_2^2 \leq \|x\|_1 \, \|x\|_\infty \qquad (y := x, p = 1, q = \infty)
\end{aligned}$$

A crucial fact for the vector space $V = \mathbb{R}^n$ or $\mathbb{C}^n$ is that the dual of a dual norm is the original norm, i.e., $\|\cdot\|_{**} = \|\cdot\|$ for an arbitrary norm $\|\cdot\|$ on V (see [15, Theorem 5.5.9, p.338]). For the special case of l_p norms, this is implied by Lemma A.24. Moreover the only l_p norm that is its own dual is the Euclidean norm $\|\cdot\|_2$ ([15, Theorem 5.4.17, p.331]). This fact and a remarkable property of dual norm specialized to $\mathbb{R}^n$ are used in Chapter A.10 to prove a mean value theorem for vector-valued functions (Lemma A.34). Specifically, for the vector space $V = \mathbb{R}^n$, it is shown in Chapter A.10 that, given

any $x \in \mathbb{R}^n$, there is a normalized $y_*(x) \in \mathbb{R}^n$ with $\|y_*(x)\|_* = 1$ such that the norm $\|x\|$ is attained by their inner product, $\|x\| = x^\mathsf{T} y_*(x)$. Similarly, there exists an $y(x)$ with $\|y(x)\| = 1$ such that $\|x\|_* = x^\mathsf{T} y(x)$. This is remarkable because it says that any norm $\|\cdot\|$ and its dual norm are always attained by the Euclidean inner product even if $\|\cdot\|$ may not be a derived norm, e.g., $\|\cdot\|_1$, $\|\cdot\|_\infty$.

A.8.3 Matrix norms

This subsection mostly follows [15, Chapter 5.6]. The set $M_{m,n} := M_{m,n}(\mathbb{C})$ of all $m \times n$ complex matrices is a vector space whether we view an element $A \in M_{mn}$ as a vector in $V = \mathbb{C}^{mn}$ over field $F = \mathbb{C}$ or $\mathbb{R}$ or an array of numbers in $V = \mathbb{C}^{m\times n}$ over $F = \mathbb{C}$ or $\mathbb{R}$. A matrix norm on M_{mn} therefore follows the same definition as in Definition A.4.

Definition A.6 (Matrix norm)**.** A function $\|\cdot\| : M_{m,n} \to \mathbb{R}$ is a *matrix norm*, or simply a *norm*, if, for all complex matrices $A, B \in M_{m,n}$, $c \in \mathbb{C}$,

1. *Positivity*: $\|A\| \geq 0$ and $\|A\| = 0$ if and only if $A = 0$.
2. *Homogeneity*: $\|cA\| = |c|\,\|A\|$.
3. *Triangular inequality*: $\|A + B\| \leq \|A\| + \|B\|$.

□

A key difference between the vector spaces $\mathbb{C}^{mn}$ and $\mathbb{C}^{m\times n}$ is that matrix multiplication is defined for elements A, B of $\mathbb{C}^{m\times n}$. We would therefore like to estimate the 'size' of a matrix product AB in terms of the 'sizes' of A and B. This is done by matrix norms $\|\cdot\|$ that also satisfies a fourth property:

4. . *Submultiplicativity*: $\|AB\| \leq \|A\|\,\|B\|$ when A and B have compatible sizes (e.g., $m = n$) and the norms are properly defined for AB, A and B.

Not all matrix norms are submultiplicative. Some authors include submultiplicativity in the definition of matrix norm when restricted to square matrices ($m = n$), e.g., [15, Chapter 5.6]. In the following we first discuss a special class of matrix norms, called induced norms, that are not only submultiplicative, but also have a certain minimality property. Then we discuss vector norms that are l_p norms on the vector space C^{n^2}. They may or may not be submultiplicative.

Induced norms.

A widely used matrix norm $\|\cdot\|_{m,n}$ on $M_{m,n}(\mathbb{C})$ is an *induced norm*, induced by any vector norms $\|\cdot\|_n$ and $\|\cdot\|_m$ on C^n and C^m respectively, defined by: for $A \in M_{m,n}$,

$$\|A\|_{m,n} := \max_{x:\|x\|_n=1} \|Ax\|_m = \max_{x:x\neq 0} \frac{\|Ax\|_m}{\|x\|_n} \tag{A.27}$$

It is sometimes called an *operator norm*. Every induced norm is submultiplicative: for $A \in \mathbb{C}^{m\times n}$, $B \in \mathbb{C}^{n\times k}$ with arbitrary norms $\|\cdot\|_m$, $\|\cdot\|_n$, $\|\cdot\|_k$ on $\mathbb{C}^m$, $\mathbb{C}^n$, $\mathbb{C}^k$ respectively,

$$\|AB\|_{m,k} = \max_{\substack{x:x\neq 0\\ \|Bx\|\neq 0}} \frac{\|ABx\|_m}{\|x\|_k} = \max_{\substack{x:x\neq 0\\ \|Bx\|\neq 0}} \frac{\|ABx\|_m}{\|Bx\|_n}\frac{\|Bx\|_n}{\|x\|_k} \leq \max_{y:y\neq 0}\frac{\|Ay\|_m}{\|y\|_n}\max_{x:x\neq 0}\frac{\|Bx\|_n}{\|x\|_k} = \|A$$

It also satisfies the additional properties:

1 $\|I\|_{m,n} = 1$ for the identity matrix I.
2 $\|Ax\|_m \leq \|A\|_{m,n}\|x\|_n$ for any $A \in \mathbb{C}^{m\times n}$ and any $x \in \mathbb{C}^n$ (follows from submultiplicativity).
3 $\|A\|_{m,n} = \max\{|y^{\mathsf{H}}Ax| : \|x\| = \|y\|_* = 1,\ x \in \mathbb{C}^n, y \in \mathbb{C}^m\}$.

Examples of induced norms on $M_{m,n}$ are norms induced by the l_p norm on both $\mathbb{C}^n$ and $\mathbb{C}^m$:

$$\|A\|_p := \max_{x:\|x\|_p=1}\|Ax\|_p = \max_{x:x\neq 0}\frac{\|Ax\|_p}{\|x\|_p}$$

Theorem A.25. Let $A \in M_{m,n}$ a $m\times n$ complex matrix. Then the induced norms $\|\cdot\|_1$, $\|\cdot\|_2$ and $\|\cdot\|_\infty$ satisfy:

1 *Max column sum* (induced by l_1 norm): $\|A\|_1 = \max_j \sum_i |A_{ij}|$.
2 *Max row sum* (induced by l_∞ norm):$\|A\|_\infty = \max_i \sum_j |A_{ij}|$.
3 *Spectral norm* (induced by l_2 norm): $\|A\|_2 = \sigma_{\max}(A) = \sqrt{\lambda_{\max}\left(A^{\mathsf{H}}A\right)}$ where $\sigma_{\max}(A)$ is the largest singular value of A and $\lambda_{\max}\left(A^{\mathsf{H}}A\right) \geq 0$ is the largest eigenvalue of the positive semidefinite matrix $A^{\mathsf{H}}A$.
4 If A is square and nonsingular then $\|A^{-1}\|_2 = 1/\sigma_{\min}(A)$, the reciprocal of the smallest singular value of A.
5 $\|A^{\mathsf{H}}A\|_2 = \|AA^{\mathsf{H}}\|_2 = \|A\|_2^2$.
6 $\|A\|_2 = \max\{|y^{\mathsf{H}}Ax| : \|x\|_2 = \|y\|_2 = 1,\ x \in \mathbb{C}^n, y \in \mathbb{C}^m\}$.

A norm $\|\cdot\|$ is *unitarily invariant* if $\|A\| = \|UAV\|$ for all $A \in M_n$ and for all unitary matrices $U, V \in M_n$. It is *self-adjoint* if $\|A\| = \|A^{\mathsf{H}}\|$ for all $A \in M_n$. The following result shows that the spectral norm is the only induced norm that is unitarily invariant and self-adjoint [15, Theorems 5.6.34, 5.6.35].

Lemma A.26. Let $\|\cdot\|$ be a submultiplicative matrix norm on M_n. The following are equivalent:

1 $\|\cdot\|$ is the spectral norm.
2 $\|\cdot\|$ is an induced norm that is unitarily invariant, i.e., $\|A\| = \|UAV\|$ for all $A \in M_n$ and for all unitary matrices $U, V \in M_n$.
3 $\|\cdot\|$ is an induced norm that is self-adjoint, i.e., $\|A\| = \|A^{\mathsf{H}}\|$ for all $A \in M_n$.

Other matrix norms.

We can also view a complex matrix $A \in M_{m,n}$ as a vector in $\mathbb{C}^{mn}$ and treat the l_p norms on $\mathbb{C}^{mn}$ as matrix norms on $M_{m,n}$. We sometimes refer these norms as *vector norms* on $M_{m,n}$. Examples include

- l_1 *norm*: $\|A\|_{\text{sum}} := \sum_{i,j} |A_{ij}|$.
- l_2 *or Frobenius norm*: $\|A\|_F := \left(\sum_{i,j} |A_{ij}|^2\right)^{1/2}$.
- l_∞ *norm*: $\|A\|_{\max} := \max_{i,j} |A_{ij}|$.

The *Frobenius inner product* on complex matrices in $M_{m,n}$ is defined to be

$$\langle A, B\rangle_F \;:=\; \operatorname{tr} B^{\mathsf{H}} A \;=\; \sum_{i=1}^{m}\sum_{j=1}^{n} \overline{B}_{ij} A_{ij}$$

It is simply the Euclidean inner product when we view a matrix $A \in M_{m,n}$ as a vector in $\mathbb{C}^{mn}$. The Forbenius norm is then derived from the Frobenius inner product, $\|A\|_F := \sqrt{\langle A, A\rangle_F}$.

They satisfy the following properties

Theorem A.27. Let $A \in M_n$ be a $n \times n$ complex matrix.

1 $\|\cdot\|_{\text{sum}}$ and $\|\cdot\|_F$ are submultiplicative matrix norms, but $\|\cdot\|_{\max}$ is a matrix norm that is not submultiplicative.
2 The Frobenius norm is given by

$$\|A\|_F \;=\; \left|\operatorname{tr}\left(AA^{\mathsf{H}}\right)\right|^{1/2} \;=\; \sqrt{\sum_i \sigma_i^2(A)} \;=\; \sqrt{\sum_i \lambda_i\left(AA^{\mathsf{H}}\right)}$$

where $\sigma_i(A)$ denote the singular values of A and $\lambda_i(AA^{\mathsf{H}})$ denote the eigenvalues of the positive semidefinite matrix AA^{H}.
3 $\|A\|_F = \|A^{\mathsf{H}}\|_F = \|UAV\|_F$ for any unitary matrices $U, V \in M_n$ (unitarily invariant).

Hence while the spectral norm $\|\cdot\|_2$ is the only unitarily invariant and the only self-adjoint induced norm (Lemma A.26), the Frobenius norm $\|\cdot\|_F$ is a unitarily invariant and self-adjoint norm that is not induced by a vector norm on $\mathbb{C}^n$.

Since M_n is a finite dimensional vector space over field $F = \mathbb{C}$ or $\mathbb{R}$, all matrix norms, whether or not they are submultiplicative, are equivalent in the sense of (A.22) and therefore have the same convergence sequences. In particular a matrix norm that is not submultiplicative is equivalent to every submultiplicative matrix norm, and vice versa. Moreover any vector norm on M_n becomes a submultiplicative matrix norm when scaled up sufficiently [15, Theorems 5.7.8, 5.7.11, pp. 372].

Lemma A.28. 1 Given any matrix norm $N(\cdot)$ (e.g., a vector norm) on M_n and any submultiplicative matrix norm $\|\cdot\|$ on M_n, there exists finite positive constants c_m, c_M such that

$$c_m\|A\| \;\le\; N(A) \;\le\; c_M\|A\|, \qquad A \in M_n \tag{A.28}$$

2 Let $N(\cdot)$ be a vector norm on M_n and $c(N) := \max_{N(A)=1=N(B)} N(AB)$. Then $\gamma N(\cdot)$ is a submultiplicative matrix norm on M_n if and only if $\gamma \ge c(N)$

Spectral radius, matrix norm and convergence.

Induced norms have a certain minimality property among matrix norms. This can be useful, e.g., in analyzing iterative algorithms of the form $x(t+1) = Ax(t)$. We now describe the relationship between the spectral radius $\rho(A)$ of a matrix A, its matrix $\|A\|$, and convergence properties of A^k and $\sum_{j\le k} A^j$.

Theorem A.29 (Spectral radius, singular values, norms)**.** Let $\|\cdot\|$ be a submultiplicative matrix norm on M_n and $A \in M_n$. Let λ_i and σ_i be the eigenvalues and singular values of A respectively with

$$|\lambda_1| \;\ge\; \cdots \;\ge\; |\lambda_n|, \qquad \sigma_1 \;\ge\; \cdots \;\ge\; \sigma_n$$

Let $\rho(A) := |\lambda_1|$ denote the spectral radius of A.

1 $|\lambda_1| \le \sigma_1$ and $|\lambda_n| \ge \sigma_n > 0$, i.e., $|\lambda_i| \in [\sigma_n, \sigma_1]$.
2 For all i, $1/\|A^{-1}\| \le |\lambda_i| \le \rho(A) \le \|A\|$ if A is nonsingular.
3 Given any $\epsilon > 0$ there is a submultiplicative matrix norm $\|\cdot\|$ such that $\rho(A) \le \|A\| \le \rho(A) + \epsilon$. Moreover

$$\rho(A) \;=\; \inf\{\|A\| : \|\cdot\| \text{ is an induced norm}\}$$

In Theorem A.29, 1 is proved in [15, Theorem 5.6.9], 2 follows from 1 by taking $\|\cdot\|$ to be the spectral norm, and 3 is proved in [15, Lemma 5.6.10, p.347]. See Exercise A.22 for details.

As mentioned above M_n is a finite dimensional vector space over field $F = \mathbb{C}$ or $\mathbb{R}$, convergence of matrices is defined in the same way as the convergence of elements in any normed vector space $(V, \|\cdot\|)$, i.e., a sequence $\{x_k\} \subseteq V$ converges to a limit $x \in V$ if $\|x_k - x\| \to 0$ as $k \to \infty$.

Definition A.7 (Matrix convergence)**.** We say a sequence $\{A^k\} \subseteq M_n$ (or a power series $\{\sum_{j\le k} A^j\} \subseteq M_n$) *converges* if there exists a matrix $A \in M_n$ such that $A^k \to A$ (or $\sum_{j\le k} A^j \to A$) as $k \to \infty$ with respect to the underlying matrix norm $\|\cdot\|$, i.e., if $\lim_{k\to\infty} \|A^k - A\| = 0$ (or $\lim_{k\to\infty} \|\sum_{j\le k} A^j - A\| = 0$).

All matrix norms, whether or not they are submultiplicative, are norms on M_n and therefore equivalent in the sense of (A.22). Hence if A^k converges under a norm, it converges under all norms.

Theorem A.30 (Sequence convergence)**.** Let $\|\cdot\|$ be a submultiplicative matrix norm on M_n and $A \in M_n$. Let $\rho(A)$ denote the spectral radius of A.

1 If $\|A\| < 1$ then $\lim_{k\to\infty} A^k = 0$, i.e., $\left|[A^k]_{ij}\right| \to 0$ as $k \to \infty$ for all i, j.
2 $\rho(A) < 1$ if and only if $\lim_{k\to\infty} A^k = 0$.
3 *Gelfand formula*: $\rho(A) = \lim_{k\to\infty} \|A^k\|^{1/k}$.

In Theorem A.30, 1 is proved in [15, Lemma 5.6.11] and uses the fact that if A^k converges then it converges under the vector norm $\|A\|_{\max} := \max_{i,j} |A_{ij}|$, and 2 is proved in [15, Lemma 5.6.12] and says that, unlike $\|A\| < 1$, $\rho(A) < 1$ is both necessary and sufficient for the convergence of $\lim_{k\to\infty} A^k$. Theorem A.30.3 holds not only for multiplicative matrix norms, but also for any matrix norm, including vector norms [15, Corollary 5.6.14, Theorem 5.7.10]. It follows from the fact that, under a submultiplicative matrix norm, $\tilde{A} := (\rho(A)+\epsilon)^{-1}A$ has spectral radius strictly less than 1 and converges for any $\epsilon > 0$, implying that $\|A^k\|^{1/k} \le \rho(A)+\epsilon$ for sufficiently large k. On the other hand $\rho(A) \le \|A^k\|^{1/k}$ and hence $\rho(A) = \lim_{k\to\infty} \|A^k\|^{1/k}$. Extension to norms that are no submultiplicative makes use of (A.28).

Remark A.3. We often want to establish $\|A\| < 1$ for some matrix norm in order to prove convergence of sequences or power series of A. We are therefore interested in a *minimal matrix norm* $\|\cdot\|$, i.e., a submultiplicative norm on M_n such that the only submultiplicative norm $N(\cdot)$ on M_n with $N(A) \le \|A\|$ for all $A \in M_n$ is $N(\cdot) = \|\cdot\|$. It can be shown that a submultiplicative matrix norm on M_n is minimal if and only if it is an induced norm [15, Theorem 5.6.32, p.356]. □

The sum $S_k := \sum_{j=0}^{k} a_j$ of a finitely many complex numbers $a_j \in \mathbb{C}$ does not depend on the order in which a_j are summed. An infinite series $S := \lim_{k\to\infty} S_k = \sum_{j=0}^{\infty} a_j$ may, e.g., $S := 1-1+1-1+\cdots$ where the partial sums S_k oscillate between 1 and -1. This motivates a stronger notion of convergence. Specifically an infinite sum $\sum_{j=0}^{\infty} a_j$ of complex numbers $a_j \in \mathbb{C}$ is said to *converge absolutely* if $\lim_{k\to\infty} \sum_{j=0}^{k} |a_j| = a$ for some real number $a \in \mathbb{R}$.

Definition A.8 (Series convergence)**.** Considered a norm vector space $(M_n, \|\cdot\|)$. We say a power series $\{\sum_{j\le k} A^j\} \subseteq M_n$

1 *converges* if there exists a matrix $A \in M_n$ such that $\sum_{j\le k} A^j \to A$ as $k \to \infty$, i.e., if $\lim_{k\to\infty} \|\sum_{j\le k} A^j - A\| = 0$.
2 *converges absolutely* if there exists a matrix $A \in M_n$ such that $\sum_{j\le k} A^j \to A$) as $k \to \infty$ with respect to the underlying matrix norm $\|\cdot\|$, i.e., if $\lim_{k\to\infty} \|\sum_{j\le k} A^j - A\| = 0$.

For a complex power series $S(z) := \lim_{k\to\infty} \sum_{j=0}^{k} a_j z^j$, it is known that there is a *radius of convergence* $R \ge 0$, possibly ∞, such that the power series converges

absolutely for $|z| < R$, diverges if $|z| > R$, and may converge or diverge if $|z| = R$. For any complex $n \times n$ matrix $A \in M_n$ and any submultiplicative matrix norm $\|\cdot\|$ we have

$$\left\| \sum_k a_k A^k \right\| \;\leq\; \sum_k |a_k| \, \|A^k\| \;\leq\; \sum_k |a_k| \, \|A\|^k$$

where the first inequality is due to the triangular inequality and the second due to submultiplicativity. This means that a matrix power series $\sum_{k=0}^{\infty} a_k A^k$ converges absolutely if there exists a matrix norm $\|\cdot\|$ such that $\|A\| < R$, the radius of convergence for $\sum_k a_k z^k$, i.e., see Exercise A.24. Such a norm exists if and only if $\rho(A) < R$ because, given any $\epsilon > 0$, there exists a (submultiplicative) matrix norm $\|\cdot\|$ with $\rho(A) \leq \|A\| \leq \rho(A) + \epsilon$ [15, Lemma 5.6.10, p.347]. This fact and some corollaries are summarized in the next result [15, pp.350-351].

Theorem A.31 (Series convergence)**.** Let $A \in M_n$.

1 Let R be the radius of convergence of a scalar power series $\sum_{k=0}^{\infty} a_k z^k$. The matrix power series $\sum_{k=0}^{\infty} a_k A^k$ converges if $\rho(A) < R$, which holds if there exists a multiplicative matrix norm $\|\cdot\|$ on M_n such that $\|A\| < R$.

Let $\|\cdot\|$ be a submultiplicative matrix norm on M_n.

2. If $\|I - A\| < 1$ then A is nonsingular and

$$A^{-1} \;=\; \sum_{k=0}^{\infty} (I - A)^k$$

3. If $\|A\| < 1$ then $I - A$ is nonsingular and

$$(I - A)^{-1} \;=\; \sum_{k=0}^{\infty} A^k$$

4. If $\|I\| = 1$ (e.g., if $\|\cdot\|$ is an induced norm) and $\|A\| < 1$ then

$$\frac{1}{1 + \|A\|} \;\leq\; \|(I - A)^{-1}\| \;\leq\; \frac{1}{1 - \|A\|}$$

The theorem is proved in Exercise **??**.

A.9 Differentiability, complex differentiability, analyticity

Differentiability of real-valued functions.

A real-valued function $f : \mathbb{R} \to \mathbb{R}$ is said to be *differentiable at* $x \in \mathbb{R}$ if the limit

$$f'(x) \;:=\; \lim_{\substack{h \in \mathbb{R} \\ h \to 0}} \frac{f(x+h) - f(x)}{h} \tag{A.29}$$

exists. If $f'(x)$ exists, it is called the *gradient* or *derivative of f at $x \in \mathbb{R}$*. If f is differentiable at every $x \in X \subseteq \mathbb{R}$ then f is called *differentiable* on X. The straight line $\{h \in \mathbb{R} : f(x) + f'(x)h\}$ can be interpreted as a linear approximation of f at x in the sense that the error $\epsilon(h)$ is smaller than linear, i.e.,

$$\lim_{h \to 0} \frac{\epsilon(h)}{h} := \lim_{h \to 0} \frac{f(x+h) - (f(x) + f'(x)h)}{h} = 0$$

We use this to generalize differentiability to $\mathbb{R}^n$: a real-valued function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *differentiable at $x \in \mathbb{R}^n$* if there exists a vector $m \in \mathbb{R}^n$ such that

$$\lim_{\substack{h \in \mathbb{R}^n \\ h \to 0}} \frac{f(x+h) - f(x) - m^\mathsf{T} h}{\|h\|} = 0$$

When this holds, m is called the *gradient* or *derivative of f at $x \in \mathbb{R}^n$* and denoted $\nabla f(x)$. If f is differentiable at every $x \in X \subseteq \mathbb{R}^n$ then f is called *differentiable* on X. If f is differentiable with respect to $x_j \in \mathbb{R}$, when all other $x_k, k \neq j$ are held fixed, then it is called *partially differentiable at $x \in \mathbb{R}^n$ with respect to x_j*. The derivative is called the *partial derivative of f at x with respect to x_j* and denoted $\frac{\partial f}{\partial x_j}(x)$:

$$\frac{\partial f}{\partial x_j}(x) := \lim_{\substack{t \in \mathbb{R} \\ t \to 0}} \frac{f(x + te_j) - f(x)}{t}$$

where $e_j \in \mathbb{R}^n$ is the unit vector with 1 in the j position and 0 elsewhere. The row vector of partial derivatives of f at $x \in \mathbb{R}^n$ is

$$\frac{\partial f}{\partial x}(x) := \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) & \cdots & \frac{\partial f}{\partial x_n}(x) \end{bmatrix}$$

The partial derivative $\frac{\partial f}{\partial x}(x)$ describes the behavior of f at x only along the coordinate axes whereas the derivative $\nabla f(x)$ describes its behavior in all directions. If f is differentiable then it is partially differentiable, but the converse does not generally hold.

Theorem A.32. If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at $x \in \mathbb{R}^n$ then it is partially differentiable at x (i.e., $\frac{\partial f}{\partial x}(x)$ exists). Moreover its gradient $\nabla f(x)$ is given by

$$\nabla f(x) = \left[\frac{\partial f}{\partial x}(x) \right]^\mathsf{T}$$

The following example shows that the converse may not hold.

Example A.5. Consider $f : \mathbb{R}^2 \to \mathbb{R}$ defined by:

$$f(x, y) := \begin{cases} 0 & \text{if} \quad xy = 0 \\ 1 & \text{if} \quad x \neq 0 \text{ or } y \neq 0 \end{cases}$$

i.e., $f = 0$ on the x and y-axes and $f = 1$ everywhere else. It is partially differentiable over $\mathbb{R}^2$. It is discontinuous at every point on the axes and hence cannot be differentiable at those points. □

The partial derivative $\frac{\partial f}{\partial x}(x)$ in Example A.5 exists, but not continuous, on the axes. If $f : X \to \mathbb{R}$ is partially differentiable on an open set $X \subseteq \mathbb{R}^n$ and $\frac{\partial f}{\partial x}(x)$ is continuous on X (i.e., the partial derivative $\frac{\partial f}{\partial x}(x)$ exists and is continuous at every $x \in X$), then f is called *continuously differentiable on* X.

Theorem A.33. If $f : X \to \mathbb{R}$ is continuously differentiable on an open set $X \subseteq \mathbb{R}^n$, then it is differentiable on X.

Complex differentiability of complex-valued functions.

A complex-valued function $f : \mathbb{C} \to \mathbb{C}$ is *complex differentiable* at $z \in \mathbb{C}$ if

$$f'(z) := \lim_{\substack{h \in \mathbb{C} \\ h \to 0}} \frac{f(z+h) - f(z)}{h} \tag{A.30}$$

exists. When $f'(z)$ exists we will call it the *complex derivative* or *derivative of* f *at* $z \in \mathbb{C}$. Note that $f'(z)$ is generally a complex number. If f is complex differentiable at every $z \in Z \subseteq \mathbb{C}$ then f is called *holomorphic* on Z.

Even though complex differentiability in (A.30) looks similar to differentiability in (A.29), (A.30) is a much stronger notion because h must approach 0 from all directions in the complex plane. To see this we can reformulate a complex-valued function and complex differentiability in $\mathbb{R}^2$ where $f : \mathbb{C} \to \mathbb{C}$ is written in terms of its real and imaginary parts, $f(x,y) =: f_r(x,y) + \mathbf{i} f_i(x,y)$ where $x, y \in \mathbb{R}$ and $f_r, f_i \in \mathbb{R}$. Then (A.30) implies, taking $h = t(1 + \mathbf{i}0)$ and $h = t(0 + \mathbf{i})$ respectively,

$$f'(x,y) = \lim_{\substack{t \in \mathbb{R} \\ t \to 0}} \frac{f(x+t,y) - f(x,y)}{t(1+\mathbf{i}0)} = \lim_{\substack{t \in \mathbb{R} \\ t \to 0}} \left(\frac{f_r(x+t,y) - f_r(x,y)}{t} + \mathbf{i} \frac{f_i(x+t,y) - f_i(x,y)}{t} \right)$$

$$f'(x,y) = \lim_{\substack{t \in \mathbb{R} \\ t \to 0}} \frac{f(x,y+t) - f(x,y)}{t(0+\mathbf{i}1)} = \lim_{\substack{t \in \mathbb{R} \\ t \to 0}} \left(\frac{f_r(x,y+t) - f_r(x,y)}{\mathbf{i}t} + \mathbf{i} \frac{f_i(x,y+t) - f_i(x,y)}{\mathbf{i}t} \right)$$

Hence if $f =: f_r + \mathbf{i} f_i$ is holomorphic on Z then it must satisfy

$$\frac{\partial f_r}{\partial x} = \frac{\partial f_i}{\partial y}, \qquad \frac{\partial f_i}{\partial x} = -\frac{\partial f_r}{\partial y}$$

on Z. These equations are called the Cauchy-Riemann equations.

Analyticity.

A real-valued function $f : X \to \mathbb{R}$ on an open set $X \subseteq \mathbb{R}$ is said to be *real analytic on* X if at every point $x_0 \in X$ there is an open neighborhood $B_\delta(x_0) := \{x \in X : |x - x_0| < \delta\}$ around x_0 such that

$$f(x) = \sum_{k=0}^{\infty} a_k (x - x_0)^k, \qquad x \in B_\delta(x_0) \tag{A.31a}$$

Equivalently f is real analytic on X if it is infinitely differentiable so that the Taylor series around every point $x_0 \in X$ converges to $f(x)$ for all $x \in B_\delta(x_0)$, i.e.,

$$f(x) \;=\; \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x-x_0)^k, \qquad x \in B_\delta(x_0) \tag{A.31b}$$

with $a_k := f^{(k)}(x_0)/k!$. The neighborhood $B_\delta(z_0)$ is called the *region of convergence* for (A.31). A function f defined on a subset of $\mathbb{R}$ is said to be *real analytic at* $x \in \mathbb{R}$ if there is a neighborhood $B_\delta(x)$ of x on which f is real analytic.

A complex-valued function $f : Z \to \mathbb{C}$ on an open set $Z \subseteq \mathbb{C}$ is said to be *complex analytic on* Z or *analytic on* Z if at every point $z_0 \in Z$ there is a neighborhood $B_\delta(z_0) := \{z \in Z : |z - z_0| < \delta\}$ around z_0 such that

$$f(z) \;=\; \sum_{k=0}^{\infty} a_k (z-z_0)^k, \qquad z \in B_\delta(z_0) \tag{A.32a}$$

Equivalently f is analytic on Z if it is infinitely complex differentiable so that the Taylor series around every point $z_0 \in Z$ converges to $f(z)$ for all $z \in B_\delta(z_0)$, i.e.,

$$f(z) \;=\; \sum_{k=0}^{\infty} \frac{f^{(k)}(z_0)}{k!}(z-z_0)^k, \qquad z \in B_\delta(z_0) \tag{A.32b}$$

with $a_k := f^{(k)}(x_0)/k!$. A function f defined on a subset of $\mathbb{C}$ is said to be *analytic at* $z \in \mathbb{C}$ if there is a neighborhood $B_\delta(z)$ of z on which f is analytic.

An important property of holomorphic function is: $f : \mathbb{C} \to C$ is holomorphic on an open set $Z \subseteq \mathbb{C}$ if and only if it is complex analytic on Z.

A.10 Mean value theorems

When restricted to the vector space $\mathbb{R}^n$ endowed with any norm $\|\cdot\|$, the definition of dual norm $\|\cdot\|_*$ in (A.25) reduces to: for any $x \in \mathbb{R}^n$,

$$\|x\|_* := \max_{y:\|y\|=1} x^{\mathsf{T}} y \;=\; \max_{y:\|y\|=1} \left|x^{\mathsf{T}} y\right| \tag{A.33}$$

The maximization is attained since inner product is continuous and the feasible set is compact. Hence there is a normalized $y(x) \in \mathbb{R}^n$ that satisfies

$$x^{\mathsf{T}} y(x) \;=\; \|x\|_* \;\text{ and }\; \|y(x)\| \;=\; 1 \tag{A.34a}$$

Recall a crucial fact that, for the vector space $V = \mathbb{R}^n$ or $\mathbb{C}^n$, the dual of a dual norm is the original norm, i.e., $\|\cdot\|_{**} = \|\cdot\|$ for an arbitrary norm $\|\cdot\|$ on V (see [15, Theorem 5.5.9, p.338]). Therefore, given any $x \in \mathbb{R}^n$, there exists an $y_*(x) \in \mathbb{R}^n$ such that

$$x^{\mathsf{T}} y_*(x) \;=\; \|x\| \;\text{ and }\; \|y_*(x)\|_* \;=\; 1 \tag{A.34b}$$

because

$$\|x\| = \|x\|_{**} \;=\; \max_{y:\|y\|_*=1} x^{\mathsf{T}} y \;=\; x^{\mathsf{T}} y_*(x)$$

where $y_*(x)$ is a maximizer (which clearly exists).[3] Remarkably, for $\mathbb{R}^n$, (A.34) says that both the norm and its dual norm of any vector can be attained by the inner product of the vector with another vector, for any norm that may not be derived from an inner product, e.g., $\|\cdot\|_1$, $\|\cdot\|_\infty$.

We now use (A.26) and (A.34b) to prove the mean value theorem for vector-valued functions.

Lemma A.34 (MVT for vector-valued function)**.** Consider a continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}^m$. Given any x, y, w in $\mathbb{R}^n$ we have

$$w^{\mathsf{T}}(f(y) - f(x)) = w^{\mathsf{T}} \frac{\partial f}{\partial x}(z)\,(y - x) \tag{A.35a}$$

$$\|f(y) - f(x)\| \le \left\| \frac{\partial f}{\partial x}(z) \right\| \|y - x\| \tag{A.35b}$$

where $z := \alpha x + (1-\alpha) y$ for some $\alpha \in [0,1]$, $\|\cdot\|$ is any norm, and for matrix, it denotes the induced norm. If we take $w = e_i$ we obtain the usual mean value theorem for a scalar valued function: $f_i(y) - f_i(x) = \frac{\partial f_i}{\partial x}(z)(y - x)$.

Proof of Lemma A.34. Fix any x, y, w in $\mathbb{R}^n$. Let $z(\alpha) := (1-\alpha)x + \alpha y$ for $\alpha \in [0,1]$ so that $z(0) = x$ and $z(1) = y$, and $z(\alpha)$ traces the straight path from x to y. Define the function

$$g(\alpha) \;:=\; g_w(\alpha) := w^{\mathsf{T}} f(z(\alpha))$$

as a function of $\alpha \in [0,1]$. Since g is from $\mathbb{R}$ to $\mathbb{R}$ the standard mean value theorem implies that

$$g(1) - g(0) = g'(\beta)$$

for some $\beta \in [0,1]$ that depends on w. Since $g(0) = w^{\mathsf{T}} f(x)$ and $g(1) = w^{\mathsf{T}} f(y)$ this becomes (using chain rule)

$$w^{\mathsf{T}}(f(y) - f(x)) = w^{\mathsf{T}} \frac{\partial f}{\partial x}(z(\beta))\,(y - x)$$

proving (A.35a).

[3] For the p-norm the dual is the q-norm with $p^{-1} + q^{-1} = 1$ (see Lemma A.24) and

$$(y(x))_i := \frac{x_i^{p-1}}{\|x\|_p^{p-1}} \operatorname{sign}\left((x_i)^p\right)$$

so that $x^{\mathsf{T}} y(x) = \|x\|_p$ and $\|y(x)\|_q = 1$.

To prove (A.35b), use (A.34b) to choose $w \in \mathbb{R}^n$ such that[4]

$$w^{\mathsf{T}}(f(y)-f(x)) \;=\; \|f(y)-f(x)\| \;\text{ and }\; \|w\|_* = 1$$

Substituting this w into (A.35a) yields

$$\begin{aligned}\|f(y)-f(x)\| = w^{\mathsf{T}}(f(y)-f(x)) \;&=\; w^{\mathsf{T}}\frac{\partial f}{\partial x}(z(\beta))\,(y-x)\\ &\le \|w\|_* \cdot \left\|\frac{\partial f}{\partial x}(z(\beta))(x-y)\right\|\\ &\le \left\|\frac{\partial f}{\partial x}(z(\beta))\right\| \cdot \|x-y\|\end{aligned}$$

proving (A.35b). In the above, the first inequality follows from (A.26) and the second inequality follows from the definition of the induced norm of $\frac{\partial f}{\partial x}$. This completes the proof of Lemma A.34. □

A.11 Algebraic graph theory

Consider a graph $G = (N, E)$ with $N := \{1, \ldots, n\}$. G can either be undirected or directed with an arbitrary orientation. Two nodes j and k are *adjacent* if $(j,k) \in E$. A *complete* graph is one where every pair of nodes is adjacent. A subgraph of G is a graph $F = (N', E')$ with $N' \subseteq N$ and $E' \subseteq E$. A *clique* of G is a complete subgraph of G. A *maximal clique* of G is a clique that is not a subgraph of another clique of G.

By a *path* connecting nodes j and k we mean either a set of *distinct* nodes $(j, n_1, \ldots, n_i, k)$ such that $(j,n_1), (n_1,n_2), \ldots, (n_i,k)$ are edges in E or this set of edges, depending on the context. A *cycle* $(n_1, \ldots, n_i)$ is a path such that $(n_1,n_2), \ldots, (n_i,n_1)$ are edges in E. By convention we exclude a pair of adjacent nodes (j,k) as a cycle. G is *connected* if there is a path between every pair of nodes. G is *k-vertex connected* or *k-connected*, $k = 1, \ldots, n$, if it remains connected after removing fewer than k nodes. G is *k-edge-connected*, $k = 1, \ldots, n$, if it remains connected after removing fewer than k edges. Hence if G is k-connected (k-edge-connected) then it is j-connected (j-edge-connected), $j \le k$. A *connected component* of G is a subgraph of G that is connected.

A cycle in G that has no chord (an edge connecting two nodes that are non-adjacent in the cycle) is called a *minimal cycle*. G is *chordal* if all its minimal cycles are of

[4] If the norm $\|\cdot\|$ is Euclidean then the argument below simplifies to: setting $w := f(y) - f(x)$ in (A.35a) yields

$$\begin{aligned}\|f(y)-f(x)\|_2^2 &= (f(y)-f(x))^{\mathsf{T}}\frac{\partial f}{\partial x}(z(\beta))\,(y-x)\\ &\le \|f(y)-f(x)\|_2 \cdot \left\|\frac{\partial f}{\partial x}(z(\beta))\right\|_2 \|y-x\|_2\end{aligned}$$

proving (A.35b). This is done in [178].

length 3 (recall that an edge (j,k) is not considered a cycle). A *chordal extension* of G is a chordal graph on the same set of nodes as G that contains G as a subgraph. Every graph has a chordal extension; e.g. the complete graph on the same set of nodes is a trivial chordal extension.

Suppose now the graph $G = (N,E)$ is directed with an arbitrary orientation. Let $n := |N|$ and $m := |E|$. Let C denote the $n \times m$ incidence matrix defined by:

$$C_{jl} = \begin{cases} 1 & \text{if } l = j \to k \text{ for some bus } k \\ -1 & \text{if } l = i \to j \text{ for some bus } i \\ 0 & \text{otherwise} \end{cases}$$

Let the $(n-1) \times m$ matrix $\hat{C}$ denote the reduced incidence matrix of G obtained from C by removing its first row. If G has c connected components, then $\text{rank}(C) = n - c$. In particular if G is connected then $\text{rank}(C) = n - 1$. Indeed C can be written as a block diagonal matrix with the kth diagonal block C_k being the incident matrix of the kth connected component that has n_k nodes. It can be proved that $\text{rank}(C_k) = n_k - 1$.

We take $\mathbb{R}^n$ as the *node space* of G and it has a simple structure. The null space $\text{null}(C^\mathsf{T})$ consists of all $\theta \in \mathbb{R}^n$ such that $C^\mathsf{T}\theta = 0$. This implies that $\theta_i = \theta_j$ if $(i,j) \in E$ is a link, i.e., a vector θ is in $\text{null}(C^\mathsf{T})$ if and only if θ_i takes the same value at every node in the same connected component. In particular, if G is connected, then $\text{null}(C^\mathsf{T})$ is $\text{span}(\mathbf{1})$ and therefore its orthogonal complement $\text{range}(C)$ has dimension $n-1$ and consists of all vectors $p \in \mathbb{R}^n$ such that $\mathbf{1}^\mathsf{T} p = 0$. See Figure A.6.

We take $\mathbb{R}^m$ as the *edge space* of G. Since $\text{rank}(C^\mathsf{T}) = \text{rank}(C) = n-1$ for a connected G, $\dim(\text{null}(C)) = m - n + 1$; see Figure A.6. A *cycle* in G is a set of edges in E that forms a cycle subgraph. Given a cycle σ in G, pick an orientation for σ, say, clockwise. Define the indicator function (vector) $z(\sigma)$ as

$$z_l(\sigma) = \begin{cases} +1 & \text{if edge } l \text{ is in } \sigma \text{ and has the same orientation as } \sigma \\ -1 & \text{if edge } l \text{ is in } \sigma \text{ and has the opposite orientation as } \sigma \\ 0 & \text{otherwise} \end{cases}$$

Partition N into two nonempty disjoint subsets N_1 and N_2. A *cut* in G is a set of edges in E each of which has one endpoint in N_1 and the other endpoint in N_2. Given a cut κ in G, pick an orientation, say, from N_1 to N_2. Define the indicator function (vector) $z(\kappa)$ as

$$z_l(\kappa) = \begin{cases} +1 & \text{if edge } l \text{ is in } \kappa \text{ and has the same orientation as } \kappa \\ -1 & \text{if edge } l \text{ is in } \kappa \text{ and has the opposite orientation as } \kappa \\ 0 & \text{otherwise} \end{cases}$$

Both the vectors $z(\sigma)$ and $z(\kappa)$ are in $\{0,1,-1\}^m$. Given a partition of N into N_1 and N_2, the indicator function $z(\kappa)$ of the cut can be expressed as

$$z(\kappa) := \pm\frac{1}{2}\left(\sum_{i \in N_1} c_i - \sum_{i \in N_2} c_i\right)$$

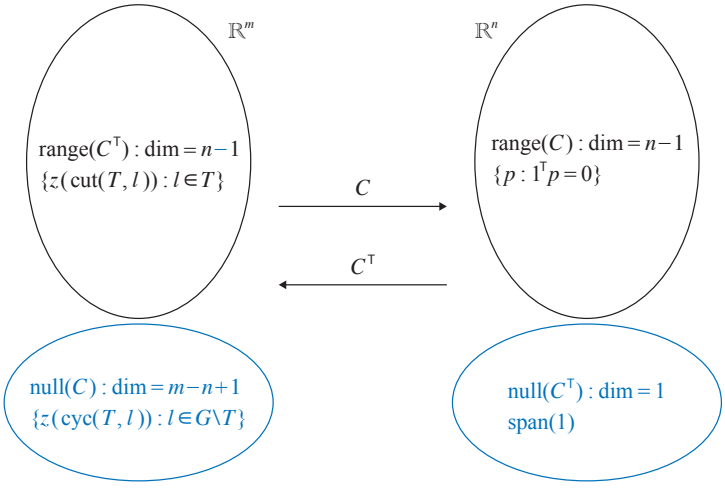
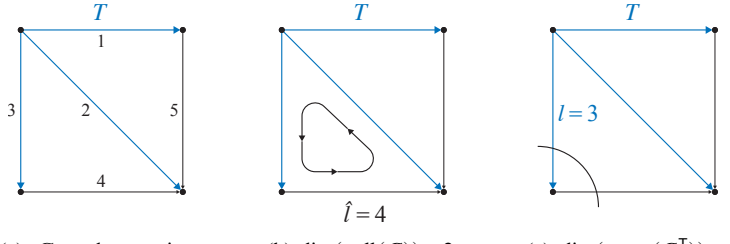


Figure A.6 The edge space $\mathbb{R}^m = \text{null}(C) \oplus \text{range}(C^\mathsf{T})$ and the vertex space $\mathbb{R}^n = \text{null}(C^\mathsf{T}) \oplus \text{range}(C)$.

where c_i are the ith rows of C. This means that $z(\kappa)$ is in the range space of C^T, and hence is orthogonal to the kernal of C, i.e., if $C\tilde{z} = 0$ then $z^\mathsf{T}(\kappa)\tilde{z} = 0$. Call the null space of C the *cycle subspace* of G and its orthogonal complement the *cut subspace* of G; see Figure A.6.

Fix any spanning tree T of G. For each edge l of G not in T, there is a unique cycle consisting of l and only edges in T; denote this cycle by $\text{cyc}(T,l)$. For each edge l of T, there is a unique cut consisting of l and only edges not in T; denotes this cut by $\text{cut}(T,l)$. Give $\text{cyc}(T,l)$ and $\text{cut}(T,l)$ the orientations that coincide with the orientation of l in G. These definitions are illustrated in Figure A.7. The following properties of



(a) G and spanning tree T

(b) $\dim(\text{null}(C)) = 2$

(c) $\dim(\text{range}(C^\mathsf{T})) = 3$

Figure A.7 A connected graph with a spanning tree T, with cycle subspace $\text{null}(C)$ and cut subspace $\text{range}(C^\mathsf{T})$. The cycle subspace has dim 2 with a basis vector for each $\hat{l} \notin T$, e.g., $z(\text{cyc}(T,\hat{l})) = (0,-1,1,1,0)$. The cut subspace has dim 3 with a basis vector for each $l \in T$, e.g., $z(\text{cut}(T,l)) = (0,0,1,-1,0)$.

the edge space of G are illustrated in Figure A.6.

Theorem A.35 (Edge space $\mathbb{R}^m$ of G)**.** 1 The cycle subspace $\text{null}(C)$ is a vector space of dimension $m-n+1$; $z(\sigma) \in \text{null}(C)$ for any cycle σ.
2 Given a spanning tree T, the set $\{z(\text{cyc}(T,l)) : l \in G \setminus T\}$ forms a basis that spans $\text{null}(C)$.
3 The cut subspace $\text{range}\left(C^\mathsf{T}\right)$ is a vector space of dimension $n-1$; $z(\kappa) \in \text{range}\left(C^\mathsf{T}\right)$ for any cut κ.
4 Given a spanning tree T, the set $\{z(\text{cut}(T,l)) : l \in T\}$ forms a basis that spans $\text{range}\left(C^\mathsf{T}\right)$.
5 The edge space of G is the orthogonal direct sum of its cycle subspace and cut subspace, i.e., $\mathbb{R}^m = \text{null}(C) \oplus \text{range}\left(C^\mathsf{T}\right)$ and $z_\sigma^\mathsf{T} z_\kappa = 0$ for any $z_\sigma \in \text{null}(C)$ and $z_\kappa \in \text{range}\left(C^\mathsf{T}\right)$.

Theorem A.36. 1 (Poincaré 1901) Any square submatrix of the incidence matrix C of a graph G has determinant equal to $0, +1$, or -1.
2 Let $F \subseteq E$ with $|F| = n-1$. Let C_F be an $(n-1)\times(n-1)$ submatrix of C, consisting of the intersection of those $n-1$ columns of C corresponding to the $n-1$ edges in F and any $n-1$ rows of C. Then C_F is invertible if and only if the subgraph induced by F is a spanning tree of G.
3 (Inverse of C_T) Let T be a spanning tree of G. Let C_T denote the corresponding $(n-1)\times(n-1)$ submatrix. Then $\left[C_T^{-1}\right]_{li} = \pm 1$ if edge l is in the unique path in T joining node i and the reference node 0 corresponding to the row excluded from C_T. Otherwise $\left[C_T^{-1}\right]_{li} = 0$.

A basis for the cycle subspace $\text{null}(C)$ and that of the cut subspaces $\text{range}\left(C^\mathsf{T}\right)$ can be explicitly determined in terms of the incidence matrix C, as follows. Partition C such that columns $1,\ldots,N$ are the edges of a spanning tree T of G. Partition C as (node 0 is the reference bus):

$$C = \begin{bmatrix} \hat{C}_T & \hat{C}_{-T} \\ d_{0T} & d_{-0T} \end{bmatrix} \tag{A.36a}$$

By Theorem A.36, $\hat{C}_T$ is invertible and its $n-1$ rows form a basis since T is a spanning tree of G. Let Z_σ denote the $m \times (m-n+1)$ matrix whose columns are the basis $\{z(\text{cyc}(T,l)) \,|\, l \in G \setminus T\}$ of the cycle subspace $\text{null}(C)$, written as (possibly after rearranging the columns):

$$Z_\sigma = \begin{bmatrix} Z_T \\ \mathbb{I}_{m-n+1} \end{bmatrix} \tag{A.36b}$$

The lower submatrix of Z_σ is $\mathbb{I}_{m-n+1}$ because these rows correspond to edges not in the spanning tree T and the orientations of the cycles have been chosen so that they coincide with the orientation of these edges. By the definition of Z_σ we have the important topological relation $C\,Z_\sigma = 0$. Using (A.36) we therefore have

$$Z_T = -\hat{C}_T^{-1}\hat{C}_{-T}$$

From Theorem A.36.3, each column of Z_T corresponds to a directed edge $i \to j$ not

in the spanning tree T, and its nonzero entries correspond to edges on the unique path between node i and node j in T. Hence a basis for the cycle subspace is given by the columns of

$$Z_\sigma = \begin{bmatrix} -\hat{C}_T^{-1}\hat{C}_{-T} \\ \mathbb{I}_{m-n+1} \end{bmatrix} \tag{A.37a}$$

Note that Theorem A.36 implies that $\hat{C}_T^{-1}$ has integral entries, so Z also has integral entries. Similarly, we can explicitly determine the cut matrix. Let Z_κ denote the $m \times n-1$ matrix whose columns are the basis $\{z(\text{cut}(T,l)) \,|\, l \in T\}$ of the cut subspace range$\left(C^\mathsf{T}\right)$, written as (possibly after rearranging the columns):

$$Z_\kappa = \begin{bmatrix} I_{n-1} \\ Z_{-T} \end{bmatrix}$$

Since every column of Z_κ belongs to the orthogonal complement of null(C), we have $Z_\sigma^\mathsf{T} Z_\kappa = 0$. Hence

$$Z_{-T} = \hat{C}_{-T}^\mathsf{T}\, \hat{C}_T^{-\mathsf{T}}$$

where $A^{-T} := (A^{-1})^\mathsf{T} = (A^\mathsf{T})^{-1}$ for any invertible matrix A and the basis for the cut space is

$$Z_\kappa = \begin{bmatrix} \mathbb{I}_{n-1} \\ \hat{C}_{-T}^\mathsf{T}\, \hat{C}_T^{-\mathsf{T}} \end{bmatrix} \tag{A.37b}$$

Since $Z_T = -\hat{C}_T^{-1}\hat{C}_{-T}$ in Z_σ and $Z_{-T} = \hat{C}_{-T}^\mathsf{T}\, \hat{C}_T^{-\mathsf{T}}$ in Z_κ, we have $Z_T^\mathsf{T} + Z_{-T} = 0_{(m-n+1)\times n-1}$. This implies for $l \in T$ and $\hat{l} \in G \backslash T$ that

$$l \in \text{cyc}(T, \hat{l}) \Leftrightarrow \hat{l} \in \text{cut}(T, l)$$

Example A.6. For the graph in Figure A.7 we have

$$Z_\sigma = \left[\begin{array}{cc} 0 & 1 \\ -1 & -1 \\ 1 & 0 \\ \hline 1 & 0 \\ 0 & 1 \end{array}\right] \text{ and } Z_\kappa = \left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \hline 0 & 1 & -1 \\ -1 & 1 & 0 \end{array}\right]$$

One can verify that, indeed, $Z_T^\mathsf{T} + Z_{-T} = 0_{2\times 3}$

This structure can be used to understand loop flows in the DC power flow model. We call a line flow vector P a *loop flow* if it satisfies power balance with zero injections, i.e., $CP = 0$. Hence P_σ is a loop flow if and only if it is in the cycle subspace null(C) of G, i.e., $P_\sigma = Z_\sigma \alpha$ for some vector $\alpha \in \mathbb{R}^{m-n+1}$. Given any balanced injection vector p with $\sum_j p_j = 0$, the line flows P that satisfy $p = CP$ are not unique. If P satisfies $p = CP$, so does $P + P_\sigma$ for any loop flow P_σ. See Remark 4.10.

A matrix is called *totally unimodular* if any square submtrix has determinant equal

to $0, +1$, or -1. Hence Theorem A.36.1 implies that the incidence matrix C of any *directed* graph G is totally unimodular.

Theorem A.37. Given any (directed) graph G,

1 The incidence matrix C of any *directed* graph G is totally unimodular.
2 If A is a totally unimodular matrix and b is an integral vector, then, for any c, the solution of the linear program

$$\min_x \ c^T x \ \text{ subject to } Ax \le b$$

has an optimal solution which is integral, provided a finite solution exists.

The significance of the theorem is that many optimization problem on graphs have LP formulations where A is the incidence matrix or its variant, e.g. max flow, shortest path problems.

A.12 Bibliographical notes

There are many excellent texts on linear algebra. Most of the materials in Chapter A.6 can be found in [176, Chapter 7.3] for singular value decomposition and properties of singular values, in [176, Chapters 2.5, 4.1] for spectral theorems for normal and Hermitian matrices, and [176, Chapter 4.4.] for complex symmetric matrices. The basic notions of algebraic graph theory in Chapter A.11 mostly follow [179].

There are many classic texts on nonsmooth convex analysis and optimization (e.g. Rockafellar, Clarke, ...). The materials in Section **??** mostly follow [54, Chapter 5], [141]. Books on nonsmooth analysis include [180, 141, 181] with [180] focuses more on control theory for applications of nonsmooth analysis and [141, 181] more on nonsmooth convex optimization. The emphasis of [141] is on $\mathbb{R}^n$ whereas that of [181] is on infinite dimensional vector spaces.

A.13 Problems

Chapters A.3–A.6.

Exercise A.1 (Matrix sum and product)**.** Let $A, B \in \mathbb{C}^{n\times n}$.

1 Show that if A, B are nonsingular then AB is nonsingular but $A+B$ can be singular.
2 Suppose $A \succ 0$ and $B \succ 0$. Show that $A+B \succ 0$ but AB may not be positive definite. Show that if $AB = BA$ or if A and B have the same set of eigenvectors then $AB \succ 0$. (Hint: $AB = BA$ if and only if A and B are simultaneously diagonalizable.)

3 Give an example of $A \succ 0$ and $B \succ 0$ that share the same set of eigenvectors and hence $AB \succ 0$.
4 Given an example where $A \succ 0$ and $B \succ 0$ but $AB \not\succ 0$.

Exercise A.2 (Invertibility of complex symmetric matrix)**.** Let $M = A + \mathbf{i}B$ where $A, B \in \mathbb{R}^{n \times n}$ and $\alpha = \rho + \mathbf{i}\epsilon$ where $\rho, \epsilon \in \mathbb{R}^n$. Show that, if M is (complex) symmetric, then

$$\alpha^{\mathsf{H}} M \alpha \;=\; (\rho^{\mathsf{T}} A \rho + \epsilon^{\mathsf{T}} A \epsilon) \;+\; \mathbf{i}(\rho^{\mathsf{T}} B \rho + \epsilon^{\mathsf{T}} B \epsilon)$$

Show that, if M is (complex) symmetric, then

1 If $A \succ 0$ then M^{-1} exists and $\mathrm{Re}(M^{-1}) \succ 0$.
2 If $B \succ 0$ then M^{-1} exists and $\mathrm{Im}(M^{-1}) \prec 0$.

Exercise A.3 (Schur complement)**.** Let $M \in \mathbb{C}^{n \times n}$ and partition it into blocks:

$$M = \begin{bmatrix} A & B \\ D & C \end{bmatrix}$$

such that $A \in \mathbb{C}^{(n-k)\times(n-k)}$, $k < n$, and the other submatrices are of appropriate dimensions. If M and A are invertible, show that

$$M^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B(M/A)^{-1}DA^{-1} & -A^{-1}B(M/A)^{-1} \\ -(M/A)^{-1}DA^{-1} & (M/A)^{-1} \end{bmatrix}$$

where $M/A := C - DA^{-1}B$ is the Schur complement of A of matrix M.

Exercise A.4 (Push-through identities)**.** Let $A \in \mathbb{C}^{n \times n}$, $B \in \mathbb{C}^{n \times k}$ and $C \in \mathbb{C}^{k \times n}$. Then

1 $(\mathbb{I}_n + BC)^{-1}B = B(\mathbb{I}_k + CB)^{-1}$ provided the inverses exist.
2 $(A + BC)^{-1}B = B(A + CB)^{-1}$ provided $n = k$, $AB = BA$ and the inverses exist.

Note that when $k \ll n$, $\mathbb{I}_k + CB$ can be much easier to invert than $\mathbb{I}_n + BC$.

Exercise A.5. Find the singular value decomposition, pseudo-inverse $A^{\dagger}$, $\mathrm{null}(A)$, $\mathrm{range}(A)$, $\mathrm{null}(A^T)$ and $\mathrm{range}(A^T)$ of the following:

1 $A = \begin{bmatrix} a \\ b \end{bmatrix}$.
2 $A = [1 \;\; 2]$.

3 $A = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$.

4 $A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$.

Discuss the existence and uniqueness of solutions to $Ax = b$ given b.

Exercise A.6. Consider $A = \begin{bmatrix} 1 & 1 \\ 1 & -2 \end{bmatrix}$. Let $B := \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $C := \begin{bmatrix} 1 \\ -2 \end{bmatrix}$ so that $A = [B\ C]$. Show that $A^\dagger = A^{-1} \neq \begin{bmatrix} B^\dagger \\ C^\dagger \end{bmatrix}$.

Exercise A.7 (Singular value decomposition)**.** On the uniqueness of the unitary matrix W in Theorem A.11, suppose $\operatorname{rank}(A) =: r < m \leq n$. For a given V given in Theorem A.11, show that W defined by $W^{\mathsf{H}} := \Sigma^\dagger V^{\mathsf{H}} A$ generally does not satisfy the singular value decomposition (A.11). Here $\Sigma^\dagger$ is obtained from Σ by replacing its positive singular values σ_i by $1/\sigma_i$ and taking the transpose.

Exercise A.8 (Singular value decomposition)**.** Let $x \in \mathbb{C}^n$ be an $n \times 1$ matrix. Compute a singular value decomposition of x.

Exercise A.9 (SVD and unitary diagonalization)**.** Prove Theorem A.16.

Chapter A.7.

Exercise A.10 (Pseudo-inverse of A)**.** Consider a matrix $A \in \mathbb{C}^{m \times n}$ as a mapping $A : \mathbb{C}^n \to \mathbb{C}^m$ and its Hermitian transpose $A^{\mathsf{H}} : \mathbb{C}^m \to \mathbb{C}^n$. Show that the mapping A restricted from $\operatorname{range}(A^{\mathsf{H}})$ to $\operatorname{range}(A)$ is surjective and injective. This means that an inverse, denoted $A^\dagger : \operatorname{range}(A) \to \operatorname{range}(A^{\mathsf{H}})$, always exists for any matrix A.

Exercise A.11 (Pseudo-inverse of A)**.** For the mapping A in Exercise A.10, show that $A^\dagger = W \Sigma^\dagger V^{\mathsf{H}}$, i.e., A and $A^\dagger$ are inverse of each other when restricted to $\operatorname{range}(A^{\mathsf{H}})$ and $\operatorname{range}(A)$.

Exercise A.12 (Pseudo-inverse of A)**.** Consider a matrix $A \in \mathbb{C}^{m\times n}$ with rank $A = r \leq \min\{m,n\}$. Let $A = V\Sigma W^{\mathsf{H}}$ be its singular value decomposition and $A^{\dagger} = W\Sigma^{\dagger}V^{\mathsf{H}}$ be its pseudo-inverse. Prove (Corollary A.20.3): If $r = n \leq m$ then $A^{\dagger} = \left(A^{\mathsf{H}}A\right)^{-1}A^{\mathsf{H}}$.

Exercise A.13 (Pseudo-inverse of A)**.** Consider a matrix $A \in \mathbb{C}^{m\times n}$ with rank $A = r \leq \min\{m,n\}$. Instead of using the formula $A^{\dagger} = W\Sigma^{\dagger}V^{\mathsf{H}}$, use the fact that $A^{\dagger}$ and A are inverse of each other when restricted to range(A^{H}) and range(A) to prove:

1 If $r = m \leq n$ then $A^{\dagger} = A^{\mathsf{H}}\left(AA^{\mathsf{H}}\right)^{-1}$.
2 If $r = n \leq m$ then $A^{\dagger} = \left(A^{\mathsf{H}}A\right)^{-1}A^{\mathsf{H}}$.

Exercise A.14 (Pseudo-inverse and norm minimization)**.** Consider a matrix $A \in \mathbb{R}^{m\times n}$ with rank $A = m \leq n$. Show that the pseudo-inverse solution $A^{\dagger}b$ of $Ax = b$ is the optimal solution of the quadratic program

$$\min_{x\in\mathbb{R}^n} \ \frac{1}{2}\|x\|_2^2 \quad \text{s.t.} \quad Ax = b$$

Optimization problems often have multiple equivalent formulations that involve different variables and constraints. The next two exercises explore the relationship between these equivalent constraints and their Lagrange multipliers when the constraints are affine. See also Exercise **??** on equivalent formulations of economic dispatch with reduced model.

Exercise A.15 (Equivalent constraints)**.** Consider the equations $A_1x = b_1$ and $A_2x = b_2$ with $x \in \mathbb{R}^n$, $A_1 \in \mathbb{R}^{m\times n}$, $A_2 \in \mathbb{R}^{k\times n}$, $b_1 \in \mathbb{R}^m$, $b_2 \in \mathbb{R}^k$, and m may not be equal to k. Suppose

- *Feasibility*: $b_1 \in$ range(A_1) and $b_2 \in$ range(A_2) so solutions for these equations always exist.
- *Equivalence*: x satisfies $A_1x = b_1$ if and only if it satisfies $A_2x = b_2$.

Remark A.2 implies that the solution set of $A_1x = b_1$ is given by

$$X_1 := \{x : x \ = \ A_1^{\dagger}b_1 + w_1,\ w_1 \in \text{null}(A_1)\}$$

and the solution set of $A_2x = b_2$ is given by

$$X_2 := \{x : x \ = \ A_2^{\dagger}b_2 + w_2,\ w_2 \in \text{null}(A_2)\}$$

Show that $X_1 = X_2$.

Exercise A.16 (Equivalent constraints)**.** Consider the setup in Exercise A.15 and the equivalent problems

$$\min_x \quad f(x) \quad \text{subject to} \quad A_1 x = b_1 \qquad [\lambda_1] \tag{A.38}$$

$$\min_x \quad f(x) \quad \text{subject to} \quad A_2 x = b_2 \qquad [\lambda_2] \tag{A.39}$$

with Lagrange multipliers λ_1, λ_2 respectively. Suppose f is differentiable (not necessarily convex). Let (x^*, λ_1^*) be a primal-dual optimal point with zero duality gap for (A.38) and (x^*, λ_2^*) be a primal-dual optimal point with zero duality gap for (A.39). Show that $A_1^T \lambda_1^* = A_2^T \lambda_2^*$.

Chapter A.8.

Exercise A.17 (Euclidean norm)**.** Show that the Euclidean norm $\|\cdot\|_2$ on $\mathbb{C}^n$ is the only unitarily invariant norm with $\|e_i\| = 1$. Positive scalar multiples of Euclidean norms are also unitarily invariant with $\|e_i\|$ not necessarily 1.

Exercise A.18 (Cauchy-Schwarz inequality)**.** Let $x_1, \ldots, x_n$ be n given real numbers with sample mean μ and sample standard deviation σ defined by:

$$\mu := \frac{1}{n}\sum_i x_i, \qquad \sigma := \left(\frac{1}{n}\sum_i (x_i - \mu)^2\right)^{1/2}$$

It can then be shown that (Exercise A.18)

$$\mu - \sigma\sqrt{n-1} \;\le\; x_i \;\le\; \mu + \sigma\sqrt{n-1}, \qquad i = 1, \ldots, n$$

with equality for some i if and only if $x_p = x_q$ for all $p, q \neq i$.

Exercise A.19 (Hölder's inequality)**.** Prove Theorem A.23 on the vector space $V = C^n$ or $\mathbb{R}^n$ with l_p norms (Hölder's inequality): For any $p, q \geq 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$

$$\sum_{i=1}^{n} |x_i y_i| \;\le\; \|x\|_p \, \|y\|_q, \qquad x,\ y \in V$$

with equality if and only if $x^p := (x_i^p, i = 1, \ldots, n)$ and $y^q := (y_i^q, i = 1, \ldots, n)$ are linearly dependent, i.e., $x^p = a y^q$ for some scalar $a \in F$.

Exercise A.20 (Induced norms)**.** Let $A \in M_{m,n}$ be a $m \times n$ complex matrix. Prove Theorem A.25:

1 *Max column sum* (induced by l_1 norm): $\|A\|_1 = \max_j \sum_i |A_{ij}|$.

2 *Max row sum* (induced by l_∞ norm):$\|A\|_\infty = \max_i \sum_j |A_{ij}|$.
3 *Spectral norm* (induced by l_2 norm): $\|A\|_2 = \sigma_{\max}(A) = \sqrt{\lambda_{\max}\left(A^{\mathsf{H}}A\right)}$ where $\sigma_{\max}(A)$ is the largest singular value of A and $\lambda_{\max}\left(A^{\mathsf{H}}A\right) \geq 0$ is the largest eigenvalue of the positive semidefinite matrix $A^{\mathsf{H}}A$.
4 If A is square and nonsingular then $\|A^{-1}\|_2 = 1/\sigma_{\min}(A)$, the reciprocal of the smallest singular value of A.
5 $\|A^{\mathsf{H}}A\|_2 = \|AA^{\mathsf{H}}\|_2 = \|A\|_2^2$.
6 $\|A\|_2 = \max\{|y^{\mathsf{H}}Ax| : \|x\|_2 = \|y\|_2 = 1,\ x \in \mathbb{C}^n, y \in \mathbb{C}^m\}$.

Exercise A.21 (Vector norms on matrices)**.** Prove Theorem A.27: Let $A \in M_n$ be a $n \times n$ complex matrix.

1 $\|\cdot\|_{\text{sum}}$ and $\|\cdot\|_F$ are submultiplicative matrix norms, but $\|\cdot\|_{\max}$ is a matrix norm that is not submultiplicative.
2 The Frobenius norm is given by

$$\|A\|_F = \left|\text{tr}\left(AA^{\mathsf{H}}\right)\right|^{1/2} = \sqrt{\sum_i \sigma_i^2(A)} = \sqrt{\sum_i \lambda_i\left(AA^{\mathsf{H}}\right)}$$

where $\sigma_i(A)$ denote the singular values of A and $\lambda_i(AA^{\mathsf{H}})$ denote the eigenvalues of the positive semidefinite matrix AA^{H}.
3 $\|A\|_F = \|A^{\mathsf{H}}\|_F = \|UAV\|_F$ for any unitary matrices $U, V \in M_n$ (unitarily invariant).

Exercise A.22 (Spectral radius, singular values, norms)**.** Let $A \in M_n$. Let $\|\cdot\|$ be a submultiplicative matrix norm on M_n and $A \in M_n$. Let λ_i and σ_i be the eigenvalues and singular values of A respectively with

$$|\lambda_1| \geq \cdots \geq |\lambda_n|, \qquad \sigma_1 \geq \cdots \geq \sigma_n$$

Let $\rho(A) := |\lambda_1|$ denote the spectral radius of A. Prove Theorem A.29:

1 $|\lambda_1| \leq \sigma_1$ and $|\lambda_n| \geq \sigma_n > 0$, i.e., $|\lambda_i| \in [\sigma_n, \sigma_1]$.
2 For all i, $1/\|A^{-1}\| \leq |\lambda_i| \leq \rho(A) \leq \|A\|$ if A is nonsingular.
3 Given any $\epsilon > 0$ there is a submultiplicative matrix norm $\|\cdot\|$ such that $\rho(A) \leq \|A\| \leq \rho(A) + \epsilon$. Moreover

$$\rho(A) = \inf\{\|A\| : \|\cdot\| \text{ is an induced norm}\}$$

Exercise A.23 (Sequence convergence)**.** Let $\|\cdot\|$ be a submultiplicative matrix norm on M_n and $A \in M_n$. Let $\rho(A)$ denote the spectral radius of A. Prove Theorem A.30:

1 If $\|A\| < 1$ then $\lim_{k\to\infty} A^k = 0$, i.e., $\left|[A^k]_{ij}\right| \to 0$ as $k \to \infty$ for all i, j.
2 $\rho(A) < 1$ if and only if $\lim_{k\to\infty} A^k = 0$.
3 *Gelfand formula*: $\rho(A) = \lim_{k\to\infty} \|A^k\|^{1/k}$.

Exercise A.24 (Series convergence)**.** Suppose there exists a matrix norm $\|\cdot\|$ such that $\|A\| < R$ where R is the radius of convergence for the power series $\sum_k a_k z^k$. Show that the matrix power series $\sum_k a_k A^k$ converges absolutely, i.e., $\lim_{k\to\infty} |a_k| \|A^k\|$

Chapter **??**.

Bibliography

[1] A. R. Bergen and V. Vittal, *Power Systems Analysis*, 2nd ed. Prentice Hall, 2000.

[2] J. D. Glover, M. S. Sarma, and T. J. Overbye, *Power system analysis and design*, 5th ed. Cengage Learning, 2008.

[3] A. J. Wood, B. F. Wollenberg, and G. B. Sheblé, *Power Generation, Operation, and Control*, 3rd ed. John Wiley & Sons, Inc., 2014.

[4] A. J. Conejo and L. Baringo, *Power system operations*. Springer, 2017.

[5] A. Gómez-Expósito, A. J. Conejo, and C. Cañizares, Eds., *Electric Energy Systems: analysis and operation*, 2nd ed. CRC Press, 2018.

[6] D. Kirschen and G. Strbac, *Fundamentals of Power System Economics*. John Wiley & Sons, 2004.

[7] W. H. Kersting, *Distribution System Modeling and Analysis*, 2nd ed. CRC Press, 2007.

[8] L. O. Chua, C. A. Desoer, and E. S. Kuh, *Linear and nonlinear circuits*. McGraw-Hill Book Company, 1987.

[9] F. Dörfler, J. W. Simpson-porco, and F. Bullo, "Electrical networks and algebraic graph theory: Models, properties, and applications," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 977–1005, May 2018.

[10] D. Venkatramanan and S. Dhople, "Per-unit modeling via similarity transformation," *IEEE Transactions on Energy Conversion*, pp. 1–13, 2022.

[11] M. Bazrafshan and N. Gatsis, "Comprehensive modeling of three-phase distribution systems via the bus admittance matrix," *IEEE Trans. Power Syst.*, vol. 33, no. 2, pp. 2015–2029, March 2018.

[12] Y. Yuan, S. H. Low, O. Ardakanian, and C. Tomlin, "Inverse power flow problem," *IEEE Transactions on Control of Network Systems*, vol. 10, no. 1, pp. 261–273, March 2023.

[13] M. K. Singh, S. Taheri, V. Kekatos, K. P. Schneider, and C.-C. Liu, "Joint grid topology reconfiguration and design of Watt-VAR curves for DER," in *2022 IEEE Power & Energy Society General Meeting (PESGM)*, July 2022.

[14] A. Trias, "The holomorphic embedding load flow method," in *Prc. IEEE Power and Energy Society General Meeting*, July 2012.

[15] R. A. Horn and C. R. Johnson, *Matrix analysis*, 2nd ed. Cambridge University Press, 2013.

[16] M. Bazrafshan and N. Gatsis, "Convergence of the Z-bus method for three-phase distribution load-flow with ZIP loads," *IEEE Transactions on Power Systems*, vol. 33, no. 1, pp. 153–165, January 2018.

[17] C. Wang, A. Bernstein, J. Y. L. Boudec, and M. Paolone, "Explicit conditions on existence and uniqueness of load-flow solutions in distribution networks," *IEEE Trans. Smart Grid*, vol. 9, no. 2, pp. 953–962, March 2018.

[18] B. Stott and O. Alsaç, "Fast decoupled load flow," *IEEE Trans. on Power Apparatus and Systems*, vol. PAS-93, no. 3, pp. 859–869, 1974.

[19] O. Alsaç, J. Bright, M. Prais, and B. Stott, "Further developments in LP-based optimal power flow," *IEEE Trans. on Power Systems*, vol. 5, no. 3, pp. 697–711, 1990.

[20] J. E. V. Ness and J. H. Griffin, "Elimination methods for load-flow studies," *Transactions of the American Institute of Electrical Engineers. Part III: Power Apparatus and Systems*, vol. 80, no. 3, pp. 299–302, April 1961.

[21] W. F. Tinney and C. E. Hart, "Power flow solution by Newton's method," *IEEE Trans. on Power Apparatus and Systems*, vol. PAS-86, no. 11, pp. 1449–1460, November 1967.

[22] W. Tinney, "Compensation methods for network solutions by optimally ordered triangular factorization," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-91, no. 1, pp. 123–127, January/ February 1972.

[23] G. Gross and H. W. Hong, "A two-step compensation method for solving short circuit problems," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-101, no. 6, pp. 1322–1331, June 1982.

[24] M. E. Baran and F. F. Wu, "Optimal Capacitor Placement on radial distribution systems," *IEEE Trans. Power Delivery*, vol. 4, no. 1, pp. 725–734, 1989.

[25] ——, "Optimal Sizing of Capacitors Placed on A Radial Distribution System," *IEEE Trans. Power Delivery*, vol. 4, no. 1, pp. 735–743, 1989.

[26] I. A. Hiskens, "Analysis tools for power systems – contending with nonlinearities," *Proc. IEEE*, vol. 83, no. 11, pp. 1573–1587, November 1995.

[27] I. A. Hiskens and R. Davy, "Exploring the power flow solution space boundary," *IEEE Trans. Power Systems*, vol. 16, no. 3, pp. 389–395, 2001.

[28] B. C. Lesieutre and I. A. Hiskens, "Convexity of the set of feasible injections and reproceedings of the ieee global conference on signal and information processing (globalsip), washington, dc, december 2016 adequacy in FTR markets," *IEEE Trans. Power Systems*, vol. 20, no. 4, pp. 1790–1798, 2005.

[29] Y. V. Makarov, Z. Y. Dong, and D. J. Hill, "On convexity of power flow feasibility boundary," *IEEE Trans. Power Systems*, vol. 23, no. 2, pp. 811–813, May 2008.

[30] D. Shirmohammadi, H. W. Hong, A. Semlyen, and G. X. Luo, "A compensation-based power flow method for weakly meshed distribution and transmission networks," *IEEE Transactions on Power Systems*, vol. 3, no. 2, pp. 753–762, May 1988.

[31] M. Farivar and S. H. Low, "Branch flow model: relaxations and convexification (parts I, II)," *IEEE Trans. on Power Systems*, vol. 28, no. 3, pp. 2554–2572, August 2013.

[32] F. Zhou and S. H. Low, "A note on branch flow model with line shunts," *IEEE Transactions on Power Systems*, vol. 31, no. 1, pp. 537–540, January 2021.

[33] S. Bose, S. H. Low, T. Teeraratkul, and B. Hassibi, "Equivalent relaxations of optimal power flow," *IEEE Trans. Automatic Control*, vol. 60, no. 3, pp. 729–742, March 2015.

[34] L. Gan, N. Li, U. Topcu, and S. H. Low, "Exact convex relaxation of optimal power flow in radial networks," *IEEE Transactions on Automatic Control*, 2014.

[35] F. Geth and B. Liu, "Notes on BIM and BFM optimal power flow with parallel lines and total current limits," in *IEEE Power and Energy Systems General Meeting*, Denver, CO, July 2022.

[36] M. Farivar, L. Chen, and S. H. Low, "Equilibrium and dynamics of local voltage control in distribution systems," in *Proc. IEEE CDC*, December 2013.

[37] H. Zhu and H. J. Liu, "Fast local voltage control under limited reactive power: optimality and stability analysis," *IEEE Transactions on Power Systems*, vol. 31, no. 5, pp. 3794–3803, September 2016.

[38] X. Zhou, M. Farivar, Z. Liu, L. Chen, and S. H. Low, "Reverse and forward engineering of local voltage control in distribution networks," *IEEE Transactions on Automatic Control*, vol. 66, no. 3, pp. 1116–1128, March 2021.

[39] S. H. Low, "Convex relaxation of optimal power flow, II: exactness," *IEEE Trans. on Control of Network Systems*, vol. 1, no. 2, pp. 177–189, June 2014.

[40] D. Deka, S. Backhaus, and M. Chertkov, "Structure learning in power distribution networks," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 1061–1074, September 2018.

[41] J. R. Berg, E. S. Hawkins, and W. W. Pleines, "Mechanized calculation of unbalanced load flow on radial distribution circuits," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS–86, no. 4, pp. 451–421, April 1967.

[42] W. H. Kersting and D. L. Mendive, "An application of ladder network theory to the solution of three-phase radial load-flow problems," in *Presented at the IEEE Winter Power Meeting*, New York, NY, January 1976.

[43] W. H. Kersting, *Distribution systems modeling and analysis*. CRC, 2002.

[44] G.-X. Luo and A. Semlyen, "Efficient load flow for large weakly meshed networks," *IEEE Transactions on Power Systems*, vol. 5, no. 4, pp. 1309–1316, 1990.

[45] C. S. Cheng and D. Shirmohammadi, "A three-phase power flow method for real-time distribution system analysis," *IEEE Transactions on Power Systems*, vol. 10, no. 2, pp. 671–679, May 1995.

[46] M. Srinivas, "Distribution load flows: a brief review," in *Power Engineering Society Winter Meeting, 2000. IEEE*, vol. 2, 2000, pp. 942–945 vol.2.

[47] R. Zimmerman and H.-D. Chiang, "Fast decoupled power flow for unbalanced radial distribution systems," *Power Systems, IEEE Transactions on*, vol. 10, no. 4, pp. 2045–2052, 1995.

[48] H.-D. Chiang and M. E. Baran, "On the existence and uniqueness of load flow solution for radial distribution power networks," *IEEE Trans. Circuits and Systems*, vol. 37, no. 3, pp. 410–416, March 1990.

[49] H.-D. Chiang, "A decoupled load flow method for distribution power networks: algorithms, analysis and convergence study," *International Journal Electrical Power Energy Systems*, vol. 13, no. 3, pp. 130–138, June 1991.

[50] K. N. Miu and H.-D. Chiang, "Existence, uniqueness, and monotonic properties of the feasible power flow solution for radial three-phase distribution networks," *IEEE Transactions on Circuits and Systems – I: Fundamental Theory and Applications*, vol. 47, no. 10, pp. 1502–1514, October 2000.

[51] E. I. Administration, "Electric power annual 2023," U.S. Department of Energy, Tech. Rep., October 2024. [Online]. Available: https://www.eia.gov/electricity/annual/pdf/epa.pdf

[52] J. Machowski, J. Bialek, and J. Bumby, *Power system dynamics: Stability and Control*, 2nd ed. John Wiley & Sons, Inc., 2008.

[53] N. Li, G. Qu, and M. Dahleh, "Real-time decentralized voltage control in distribution networks," in *Allerton Conference on Communication,Control and Computing*, Monticello, IL, 2014.

[54] D. P. Bertsekas, *Convex optimization theory*. Athena Scientific, 2009.

[55] A. Nedić and A. Ozdaglar, "Approximate primal solutions and rate analysis for dual subgradient methods," *SIAM J. Optim.*, vol. 19, no. 4, pp. 1757–1780, 2009.

[56] K. C. Border, "Miscellaneous notes on optimization theory and related topics," August 2020, lecture Notes.

[57] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[58] D. P. Bertsekas, "Projected Newton methods for optimization problems with simple constraints," *SIAM Journal on Control and Optimization*, vol. 20, no. 2, pp. 221–246, March 1982.

[59] A. M. Geoffrion, "Generalized benders decomposition," *Journal of Optimization Theory and Applications*, vol. 10, no. 4, pp. 238–260, 1972.

[60] B. Fang, C. Zhao, and S. H. Low, "Convergence of backward/forward sweep for power flow solution in radial networks," *IEEE Trans. on Control of Network Systems*, 2025, to appear.

[61] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and distributed computation*. Prentice-Hall, 1989.

[62] D. P. Bertsekas, *Nonlinear programming*. Athena Scientific, 1995.

[63] P. Milgrom and I. Segal, "Envelope theorems for arbitrary choice sets," *Econometrica*, vol. 70, no. 2, pp. 583–601, March 2002.

[64] J. F. Benders, "Partitioning procedures for solving mixed-variables programming problems," *Numerische Mathematik*, vol. 4, pp. 238–252, 1962.

[65] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributionally robust learning*, ser. Foundations and Trends in Machine Learning. NOW Publishers, 2010, vol. 3, no. 1.

[66] K. A. Clements, P. W. Davis, and K. D. Frey, "An interior point algorithm for weighted least absolute value power system state estimation," in *Proc. IEEE Power & Energy Society Winter Meeting*, 1991.

[67] F. Capitanescu and L. Wehenkel, "Experiments with the interior-point method for solving large scale optimal power flow problems," *Electric Power Systems Research*, vol. 95, pp. 276–283, 2013.

[68] A. Verma, "Power grid security analysis : An optimization approach," Ph.D. dissertation, Columbia University, 2009.

[69] J. Lavaei and S. H. Low, "Zero duality gap in optimal power flow problem," *IEEE Trans. on Power Systems*, vol. 27, no. 1, pp. 92–107, February 2012.

[70] K. Lehmann, A. Grastien, and P. V. Hentenryck, "AC-feasibility on tree networks is NP-hard," *IEEE Transactions on Power Systems*, vol. 31, no. 1, pp. 798–801, January 2016.

[71] K. Lehmann, "Computational complexity of electrical power system problems," Ph.D. dissertation, Australian National University, October 2017.

[72] M. Khonji, S. C.-K. Chau, and K. Elbassioni, "Optimal power flow with inelastic demands for demand response in radial distribution networks," *IEEE Trans. Control of Network Sys.*, vol. 5, no. 1, pp. 513–524, March 2018.

[73] D. Bienstock and A. Verma, "Strong NP-hardness of AC Power Flows Feasibility," *Operations Research Letters*, vol. 47, pp. 494–501, September 2019.

[74] M. Khonji, S. C.-K. Chau, and K. Elbassioni, "Combinatorial optimization of ac optimal power flow with discrete demands in radial networks," *IEEE Trans. Control of Network Sys.*, vol. 7, no. 2, pp. 887–898, June 2020.

[75] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company, 1979.

[76] ——, ""strong" NP-completeness results: Motivation, examples, and implications," *Journal of ACM*, vol. 25, no. 3, pp. 499–508, July 1978.

[77] S. C.-K. Chau, K. Elbassioni, and M. Khonji, "Combinatorial optimization of alternating current electric power systems," *Foundations and Trends in Electric Energy Systems*, vol. 3, no. 1–2, pp. 1–139, 2018.

[78] S. Gopinath, H. Hijazi, T. Weisser, H. Nagarajan, M. Yetkin, K. Sundar, and R. Bent, "Proving global optimality of ACOPF solutions," *Electric Power Systems Research*, vol. 189, 2020.

[79] F. Zhou and S. H. Low, "Conditions for exact convex relaxation and no spurious local optima," *IEEE Transactions on Control of Network Systems*, vol. 9, no. 3, pp. 1468–1480, 2022.

[80] ——, "A sufficient condition for local optima to be globally optimal," in *Proc. of the 59th IEEE Conference on Decision Control (CDC)*, December 2020.

[81] ARPA-E, "SCOPF problem formulation: Challenge 1," 2019, grid Optimization Competition.

[82] C. Petra and I. Aravena, "Solving realistic security-constrained optimal power flow problems," October 2021, arXiv:2110.01669v1.

[83] F. E. Curtis, D. K. Molzahn, S. Tu, A. Wächter, E. Wei, and E. Wong, "A decomposition algorithm for large-scale security-constrained AC optimal power flow," October 2021, arXiv:2110.01737v1.

[84] A. Gholami, K. Sun, S. Zhang, and A. X. Sun, "Solving large-scale security constrained AC optimal power flow problems," *Submitted to Operations Research*,

2021, special Issue on Computational Advances in Short Term Power System Operations.

[85] A. W. A and L. T. Biegler, "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming," *Mathematical Programming*, vol. 106, no. 1, pp. 25–57, 2006.

[86] J. Carpentier, "Contribution to the economic dispatch problem," *Bulletin de la Societe Francoise des Electriciens*, vol. 3, no. 8, pp. 431–447, 1962, in French.

[87] H. Dommel and W. Tinney, "Optimal power flow solutions," *Power Apparatus and Systems, IEEE Transactions on*, vol. PAS-87, no. 10, pp. 1866–1876, Oct. 1968.

[88] J. A. Momoh, *Electric Power System Applications of Optimization*, ser. Power Engineering. Markel Dekker Inc.: New York, USA, 2001.

[89] M. Huneault and F. D. Galiana, "A survey of the optimal power flow literature," *IEEE Trans. on Power Systems*, vol. 6, no. 2, pp. 762–770, 1991.

[90] J. A. Momoh, M. E. El-Hawary, and R. Adapa, "A review of selected optimal power flow literature to 1993. Part I: Nonlinear and quadratic programming approaches," *IEEE Trans. on Power Systems*, vol. 14, no. 1, pp. 96–104, 1999.

[91] ——, "A review of selected optimal power flow literature to 1993. Part II: Newton, linear programming and interior point methods," *IEEE Trans. on Power Systems*, vol. 14, no. 1, pp. 105 – 111, 1999.

[92] K. S. Pandya and S. K. Joshi, "A survey of optimal power flow methods," *J. of Theoretical and Applied Information Technology*, vol. 4, no. 5, pp. 450–458, 2008.

[93] S. Frank, I. Steponavice, and S. Rebennack, "Optimal power flow: a bibliographic survey, I: formulations and deterministic methods," *Energy Systems*, vol. 3, pp. 221–258, September 2012.

[94] ——, "Optimal power flow: a bibliographic survey, II: nondeterministic and hybrid methods," *Energy Systems*, vol. 3, pp. 259–289, September 2013.

[95] M. B. Cain, R. P. O'Neill, and A. Castillo, "History of optimal power flow and formulations (OPF Paper 1)," US FERC, Tech. Rep., December 2012.

[96] R. P. O'Neill, A. Castillo, and M. B. Cain, "The IV formulation and linear approximations of the AC optimal power flow problem (OPF Paper 2)," US FERC, Tech. Rep., December 2012.

[97] ——, "The computational testing of AC optimal power flow using the current voltage formulations (OPF Paper 3)," US FERC, Tech. Rep., December 2012.

[98] A. Castillo and R. P. O'Neill, "Survey of approaches to solving the ACOPF (OPF Paper 4)," US FERC, Tech. Rep., March 2013.

[99] ——, "Computational performance of solution techniques applied to the ACOPF (OPF Paper 5)," US FERC, Tech. Rep., March 2013.

[100] S. H. Low, "Convex relaxation of optimal power flow, I: formulations and relaxations," *IEEE Trans. on Control of Network Systems*, vol. 1, no. 1, pp. 15–27, March 2014.

[101] D. K. Molzahn and I. A. Hiskens, "A survey of relaxations and approximations of the power flow equations," *Foundations and Trends in Electric Energy Systems*, vol. 4, no. 1–2, pp. 1–221, February 2019.

[102] D. Bienstock and A. Verma, "Strong NP-hardness of AC Power Flows Feasibility," *arXiv:1512.07315*, Dec. 2015.

[103] E. Dall'Anese, H. Zhu, and G. Giannakis, "Distributed optimal power flow for smart microgrids," *IEEE Trans. on Smart Grid*, vol. 4, no. 3, pp. 1464–1475, Sep. 2013.

[104] L. Gan and S. H. Low, "Convex relaxations and linear approximation for optimal power flow in multiphase radial networks," in *Proc. of the 18th Power Systems Computation Conference (PSCC)*, Wroclaw, Poland, August 2014.

[105] C. Zhao, E. Dall'Anese, and S. H. Low, "Convex relaxation of OPF in multiphase radial networks with delta connections," in *Proceedings of the 10th Bulk Power Systems Dynamics and Control Symposium*, 2017.

[106] R. Baldick, *Applied Optimization: Formulation and Algorithms for Engineering Systems*. Cambridge University Press, 2009.

[107] J. A. Taylor, *Convex optimization of power systems*. Cambridge University Press, 2015.

[108] R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas, "MATPOWER's extensible optimal power flow architecture," in *Proc. IEEE PES General Meeting*, 2009, pp. 1–7.

[109] R. Y. Zhang and J. Lavaei, "Sparse semidefinite programs with guaranteed near-linear time complexity via dualized clique tree conversion," *Mathematical programming*, vol. 188, no. 1, pp. 351–393, 2021.

[110] K. G. Murty and S. N. Kabadi, "Some NP-complete problems in quadratic and nonlinear programming," *Mathematical Programming*, vol. 39, pp. 117–129, 1987.

[111] H.-D. Chiang and C.-Y. Jiang, "Feasible region of optimal power flow: characterization and applications," *IEEE Trans. Power Systems*, vol. 33, no. 1, January 2018.

[112] R. Grone, C. R. Johnson, E. M. Sá, and H. Wolkowicz, "Positive definite completions of partial Hermitian matrices," *Linear Algebra and its Applications*, vol. 58, pp. 109–124, 1984.

[113] S. Bose, S. H. Low, and M. Chandy, "Equivalence of branch flow and bus injection models," in *50th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, October 2012.

[114] M. Fukuda, M. Kojima, K. Murota, and K. Nakata, "Exploiting sparsity in semidefinite programming via matrix completion I: General framework," *SIAM Journal on Optimization*, vol. 11, pp. 647–674, 1999.

[115] K. Nakata, K. Fujisawa, M. Fukuda, M. Kojima, and K. Murota, "Exploiting sparsity in semidefinite programming via matrix completion II: Implementation and numerical results," *Mathematical Programming*, vol. 95, no. 2, pp. 303–327, 2003.

[116] D. R. Fulkerson and O. A. Gross, "Incidence matrices and interval graphs," *Pacific Journal of Mathematics*, vol. 15, no. 3, pp. 835–855, 1965.

[117] D. J. Rose, R. E. Tarjan, and G. S. Lueker, "Algorithmic aspects of vertex elimination on graphs," *SIAM Journal on Computing*, vol. 5, no. 2, pp. 266–283, 1976.

[118] B. Kocuk, S. Dey, and X. A. Sun, "Strong SOCP relaxations of the optimal power flow problem," *Operations Research*, vol. 64, no. 6, pp. 1177–1196, 2016.

[119] D. Bertsimas, E. Litvinov, X. A. Sun, J. Zhao, and T. Zheng, "Adaptive robust optimization for the security constrained unit commitment problem," *IEEE Transactions on Power Systems*, vol. 28, no. 1, pp. 52–63, 2013.

[120] B. Zeng and L. Zhao, "Solving two-stage robust optimization problems using a column-and-constraint generation method," *Operations Research Letters*, vol. 41, pp. 457–461, 2013.

[121] S. Kim and M. Kojima, "Exact solutions of some nonconvex quadratic optimization problems via SDP and SOCP relaxations," *Computational Optimization and Applications*, vol. 26, no. 2, pp. 143–154, 2003.

[122] S. Bose, D. Gayme, K. M. Chandy, and S. H. Low, "Quadratically constrained quadratic programs on acyclic graphs with application to power flow," March 2012, arXiv:1203.5599v1.

[123] S. Bose, D. Gayme, S. H. Low, and K. M. Chandy, "Optimal power flow over tree networks," in *Proc. Allerton Conf. on Comm., Ctrl. and Computing*, Monticello, IL, October 2011.

[124] S. Sojoudi and J. Lavaei, "Physics of power networks makes hard optimization problems easy to solve," in *IEEE Power & Energy Society (PES) General Meeting*, San Diego, CA, July 2012.

[125] B. Zhang and D. Tse, "Geometry of the injection region of power networks," *IEEE Trans. Power Systems*, vol. 28, no. 2, pp. 788–797, 2013.

[126] S. Bose, D. Gayme, K. M. Chandy, and S. H. Low, "Quadratically constrained quadratic programs on acyclic graphs with application to power flow," *IEEE Trans. Control of Network Systems*, vol. 2, no. 3, pp. 278–287, 2015.

[127] S. Sojoudi and J. Lavaei, "Semidefinite relaxation for nonlinear optimization over graphs with application to power systems," 2013, preprint.

[128] J. Lavaei, D. Tse, and B. Zhang, "Geometry of Power Flows and Optimization in Distribution Networks," *IEEE Trans. Power Systems*, vol. 29, no. 2, pp. 572–583, March 2014.

[129] B. Zhang, A. Y. Lam, A. Domínguez-García, and D. Tse, "An Optimal and Distributed Method for Voltage Regulation in Power Distribution Systems," *IEEE Trans. Power Syst.*, vol. 30, no. 4, pp. 1714–1726, July 2015.

[130] R. Jabr, "Radial Distribution Load Flow Using Conic Programming," *IEEE Trans. on Power Systems*, vol. 21, no. 3, pp. 1458–1459, Aug 2006.

[131] X. Bai, H. Wei, K. Fujisawa, and Y. Wang, "Semidefinite programming for optimal power flow problems," *Int'l J. of Electrical Power & Energy Systems*, vol. 30, no. 6-7, pp. 383–392, 2008.

[132] X. Bai and H. Wei, "A semidefinite programming method with graph partitioning technique for optimal power flow problems," *Int'l J. of Electrical Power & Energy Systems*, vol. 33, no. 7, pp. 1309–1314, 2011.

[133] R. A. Jabr, "Exploiting sparsity in SDP relaxations of the OPF problem," *Power Systems, IEEE Transactions on*, vol. 27, no. 2, pp. 1138–1139, 2012.

[134] A. Lam, B. Zhang, and D. N. Tse, "Distributed algorithms for optimal power flow problem," in *IEEE CDC*, 2012, pp. 430–437.

[135] D. Molzahn, J. Holzer, B. Lesieutre, and C. DeMarco, "Implementation of a large-scale optimal power flow solver based on semidefinite programming," *IEEE Transactions on Power Systems*, vol. 28, no. 4, pp. 3987–3998, November 2013.

[136] F. Zhou, A. S. Zamzam, S. H. Low, and N. D. Sidiropoulos, "Exactness of OPF relaxation on three-phase radial networks with Delta connections," *IEEE Transactions on Smart Grid*, vol. 12, no. 4, pp. 3232–3241, July 2021.

[137] N. Li, L. Chen, and S. Low, "Exact convex relaxation of opf for radial networks using branch flow models," in *IEEE International Conference on Smart Grid Communications*, Tainan City, Taiwan, November 2012.

[138] L. Gan, N. Li, U. Topcu, and S. H. Low, "On the exactness of convex relaxation for optimal power flow in tree networks," in *Prof. 51st IEEE Conference on Decision and Control*, Maui, HI, December 2012.

[139] ——, "Optimal power flow in distribution networks," in *Proc. 52nd IEEE Conference on Decision and Control*, December 2013, in arXiv:12084076.

[140] M. Farivar, C. R. Clarke, S. H. Low, and K. M. Chandy, "Inverter var control for distribution systems with renewables," in *Proceedings of IEEE SmartGridComm Conference*, Brussels, Belgium, October 2011.

[141] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex analysis and minimization algorithms, I: fundamentals*. Springer Verlag, 1993.

[142] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on stochastic programming: modeling and theory*, 2nd ed. SIAM, 2014.

[143] J. R. Birge and F. Louveaux, *Introduction to stochastic programming*, 2nd ed. Springer, 2011, springer Series in Operations Research and Financial Engineering.

[144] G. C. Calafiore and M. C. Campi, "Uncertain convex programs: Randomized solutions and confidence levels," *Math. Program.*, vol. 102, no. 1, pp. 25–46, 2005.

[145] G. C. Calafiore, "Random convex programs," *SIAM J. Optim.*, vol. 20, no. 6, pp. 3427–3464, 2010.

[146] M. C. Campi and S. Garatti, "The exact feasibility of randomized solutions of uncertain convex programs," *SIAM J. Optim.*, vol. 19, no. 3, pp. 1211–1230, 2008.

[147] G. C. Calafiore, "A note on the expected probability of constraint violation in sampled convex programs," in *2009 IEEE Control Applications, (CCA) & Intelligent Control, (ISIC)*, 2009, pp. 1788–1791.

[148] P. M. Esfahani, T. Sutter, and J. Lygeros, "Performance bounds for the scenario approach and an extension to a class of non-convex programs," *IEEE Transactions on Automatic Control*, vol. 60, no. 1, pp. 46–58, January 2015.

[149] D. W. Walkup and R. J.-B. Wets, "Stochastic programs with recourse," *SIAM J. Appl. Math.*, vol. 15, no. 5, pp. 1299–1314, September 1967.

[150] R. J.-B. Wets, "Stochastic programming," in *Optimization (Handbooks in Operations Research andManagement Science)*, G. Nemhauser, A. R. Kan, and M. Todd, Eds. North-Holland, Amsterdam, Netherlands, 1990, vol. 1.

[151] A. Ben-Tal, L. E. Ghaoui, and A. Nemirovski, *Robust optimization*. Princeton University Press, 2009.

[152] C. L. Fortescue, "Method of symmetrical coordinates applied to the solution of polyphase networks," in *Proceedings to the 34th Annual Convention of the American Institute of Electrical Engineers*, Atlantic City, NJ, June 1918.

[153] N. D. Rao and H. N. R. Rao, "Study of symmetrical and related components through the theory of linear vector spaces," *Proc. IEE*, vol. 113, no. 6, June 1966.

[154] R. H. Park, "Two-reaction theory of synchronous machines: generalized method of analysis - part I," *Transactions of the American Institute of Electrical Engineers*, vol. 48, no. 3, pp. 716–727, July 1929.

[155] E. Clarke, *Circuit analysis of A-C power systems, Vol 1: Symmetrical and related components*. John Wiley & Sons, 1943.

[156] R. D. Zimmerman, "Comprehensive distribution power flow: modeling, formulation, solution algorithms and analysis," Ph.D. dissertation, Cornell University, 1995.

[157] J. R. Carson, "Wave propagation in overhead wires with ground return," *Bell System Technical Journal*, vol. 5, no. 4, pp. 539–554, 1926.

[158] T.-H. Chen, M.-S. Chen, K.-J. Hwang, P. Kotas, and E. A. Chebli, "Distribution system power flow analysis – a rigid approach," *EEE Transactions on Power Delivery*, vol. 6, no. 3, July 1991.

[159] R. C. Dugan, R. Gabrick, J. C. Wright, and K. W. Patten, "Validated techniques for modeling shell-form EHV transformers," *IEEE Trans. on Power Delivery*, vol. 4, no. 2, pp. 1070–1078, April 1989.

[160] M. Laughton, "Analysis of unbalanced polyphase networks by the method of phase co-ordinates. Part 1: System representation in phase frame of reference," *Proc. Inst. Electr. Eng.*, vol. 115, no. 8, 1968.

[161] M.-S. Chen and W. Dillon, "Power system modeling," *Proc. IEEE*, vol. 62, no. 7, pp. 901–915, July 1974.

[162] T.-H. Chen, M.-S. Chen, T. Inoue, P. Kotas, and E. A. Chebli, "Three-phase cogenerator and transformer models for distribution system analysis," *EEE Transactions on Power Delivery*, vol. 6, no. 4, pp. 1671–1681, October 1991.

[163] W. H. Kersting, W. H. Phillips, and W. Carr, "A new approach to modeling three-phase transformer connection," *IEEE Trans. on Industry Applications*, vol. 35, pp. 168–175, Jan/Feb 1999.

[164] S. H. Low, "Three-phase transformer modeling," in *Proc. 59th Allerton Conf. on Communication, Control and Computing (Allerton)*, Moticello, IL, September 2023.

[165] S. S. Moorthy and D. Hoadley, "A new phase-coordinate transformer model for Ybus analysis," *IEEE Trans. on Power Systems*, vol. 17, no. 4, pp. 951–956, November 2002.

[166] M. Coppo, F. Bignucolo, and R. Turri, "Generalised transformer modelling for power flow calculation in multi-phase unbalanced networks," *IET Generation, Transmission & Distribution*, vol. 11, no. 15, pp. 3843–3852, 2017.

[167] R. C. Dugan, "A perspective on transformer modeling for distribution system analysis," in *IEEE Power Energy Society General Meeting*, Toronto, Ont. Canada, 2003.

[168] S. Claeys, G. Deconinck, and F. Geth, "Decomposition of n-winding transformers for unbalanced optimal power flow," *IET Generation, Transmission & Distribution*, vol. 14, no. 24, pp. 5961–5969, 2020. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-gtd.2020.0776

[169] F. Dörfler and F. Bullo, "Kron reduction of graphs with applications to electrical networks," *IEEE Transactions on Circuits and Systems – I: Regular Papers*, vol. 60, no. 1, pp. 150–163, January 2013.

[170] A. H. El-Abiad and D. C. Tarsi, "Load flow study of untransposed EHV networks," in *In Proceedings of the IEEE Power Industry Computer Application (PICA) Conference*, Pittsburgh, PA, 1967, pp. 337–384.

[171] K.A.Birt, J. Graff, J. McDonald, and A. El-Abiad, "Three phase load flow program," *EEE Trans. on Power Apparatus and Systems*, vol. 95, no. 1, pp. 59–65, January/February 1976.

[172] J. Arrillaga and C. P. Arnold, "Fast-decoupled three phase load flow," *Proc. IEE*, vol. 125, no. 8, pp. 734–740, 1978.

[173] A. Bernstein, C. Wang, E. Dall'Anese, J. L. Boudec, and C. Zhao, "Load flow in multiphase distribution networks: Existence, uniqueness, non-singularity and linear models," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 5832–5843, 2018.

[174] F. Zhou, Y. Chen, and S. H. Low, "Sufficient conditions for exact semi-definite relaxation of optimal power flow in unbalanced multiphase radial networks," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, 2019, pp. 6227–6233.

[175] V. Kekatos, L. Zhang, G. B. Giannakis, and R. Baldick, "Voltage regulation algorithms for multiphase power distribution grids," *IEEE Transactions on Power Systems*, vol. 31, no. 5, pp. 3913–3923, 2016.

[176] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge University Press, 1985.

[177] R. E. Cline, "Representations for the generalized inverse of a partitioned matrix," *Journal of Society for Industrial and Applied Mathematics*, vol. 12, no. 3, pp. 588–600, September 1964.

[178] W. S. Hall and M. L. Newall, "The mean value theorem for vector-valued functions: a simple proof," *Mathematics Magazine*, vol. 52, pp. 157–158, 1979.

[179] N. Biggs, *Algebraic graph theory*. Cambridge University Press, 1993, cambridge Mathematical Library.

[180] F. H. Clarke, Y. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Nonsmooth analysis and control theory*. Springer, 1998.

[181] W. Schirotzek, *Nonsmooth anslysis*. Springer, 2007.

[182] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

Index

S-lemma, 750, 755

Azuma-Hoeffding inequality, 779

basis, 1149
  orthogonal, 1149
  orthonormal, 1149
  standard, 1149
  unit, 1149
  unit vector, 1149

Cauchy sequence, 1180
Cauchy-Schwarz inequality, 1181
chance constrained program, 759
chance constraint, 759
change of measure, 776, 777
Chebyshev's inequality, 769
complete normed linear space, 1180
complete recourse, 805
conjugate function, 681
constraint
  chance, 759
  probabilistic, 759
corrective action, 802

DC OPF, *see* economic dispatch
DC optimal power flow, *see* economic dispatch
DC power flow, 242
  shift factor, 837
direct sum, 1149
dual norm, 1182

economic dispatch, 484
  LMP, 484
  locational marginal price, 484

fixed recourse, 802
Frobenius inner product, 1187
Frobenius norm, 1186

Hölder's inequality, 1182
Hoeffding's lemma, 776, 777

induced norm, 1185
inequality
  Azuma-Hoeffding, 779
  Cauchy-Schwarz, 1181
  Chebyshev's, 769
  Hölder's, 1182
  Markov's, 768
inner product space, 1180

linear space, 1147
  complete normed linear space, 1180
  complex, 1148
  dimension, 1149
  inner product space, 1180
  isomorphic, 1149
  isomorphism, 1149
  normed linear space, 1178
  normed vector space, 1178
  real, 1148
  span, 1149
  subspace, 1148
linearly dependent, 1149
linearly independent, 1149
LMP, 484
locational marginal price, 484

Markov's inequality, 768
martingale, 779
matrix norm, 1185
moment-generating function, 769
  log moment-generating function, 769

norm, 1178
  L_1 norm, 1179
  L_2 norm, 1179
  L_∞ norm, 1179
  L_p norm, 1179
  l_1 norm, 1178
  l_2 norm, 1178
  l_∞ norm, 1178
  l_p norm, 1178
  dual, 1182
  Frobenius, 1186
  induced, 1185
  matrix, 1185
  operator, 1185
  spectral, 1186
  vector norm, 1178
normed linear space, 1178
normed vector space, 1178

operator norm, 1185
orthogonal basis, 1149
orthonormal basis, 1149

probabilistic constraint, 759
probability measure, 758
probability space, 758
  σ-algebra, 758
  absolutely continuous, 777
  distribution function, 758
  event, 758
  probability distribution function, 758
  probability measure, 758
  random variable, 758
  random vector, 758
  sample space, 758
program
  chance constrained, 759
  scenario, 782
  two-stage, 801

recourse action, 802
recourse function, 802
recourse matrix, 802
relative complete recourse, 805

safe approximation, 773
scenario program, 782
  violation probability, 783
shift factor, 837
spectral norm, 1186
stochastic linear program, 801
sub-Gaussian, 771

two-stage optimization, 801
  complete recourse, 805
  corrective action, 802
  fixed recourse, 802
  recourse action, 802
  recourse function, 802
  recourse matrix, 802
  relative complete recourse, 805
  second-stage expected value function, 802
  second-stage value function, 802
  stochastic linear program, 801

vector space, *see* linear space
violation probability, 783