

Applied Mathematical Sciences

Wolfgang Hackbusch

Iterative Solution of Large Sparse Systems of Equations

Second Edition



Springer

Applied Mathematical Sciences

Volume 95

Editors

S.S. Antman, Institute for Physical Science and Technology, University of Maryland, College Park, MD, USA

Leslie Greengard, Courant Institute of Mathematical Sciences, New York University, New York, NY, USA

P.J. Holmes, Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ, USA

Advisors

J. Bell, Lawrence Berkeley National Lab, Center for Computational Sciences and Engineering, Berkeley, CA, USA

P. Constantin, Department of Mathematics, Princeton University, Princeton, NJ, USA

R. Durrett, Department of Mathematics, Duke University, Durham North, NC, USA

J. Keller, Department of Mathematics, Stanford University, Stanford, CA, USA

R. Kohn, Courant Institute of Mathematical Sciences, New York University, New York, USA

R. Pego, Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA, USA

L. Ryzhik, Department of Mathematics, Stanford University, Stanford, CA, USA

A. Singer, Department of Mathematics, Princeton University, Princeton, NJ, USA

A. Stevens, Department of Applied Mathematics, University of Münster, Münster, Germany

A. Stuart, Mathematics Institute, University of Warwick, Coventry, UK

S. Wright, Computer Sciences Department, University of Wisconsin, Madison, WI, USA

Founding Editors

Fritz John, Joseph P. LaSalle and Lawrence Sirovich

More information about this series at <http://www.springer.com/series/34>

Wolfgang Hackbusch

Iterative Solution of Large Sparse Systems of Equations

Second Edition



Springer

Wolfgang Hackbusch
Max Planck Institute for Mathematics
in the Sciences
Leipzig
Germany

ISSN 0066-5452
Applied Mathematical Sciences
ISBN 978-3-319-28481-1
DOI 10.1007/978-3-319-28483-5

ISSN 2196-968X (electronic)
ISBN 978-3-319-28483-5 (eBook)

Library of Congress Control Number: 2016940360

Mathematics Subject Classification (2010): 65F10, 65N22, 65N55

© Springer International Publishing Switzerland 1994, 2016

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Preface

The numerical treatment of partial differential equations splits into two different parts. The first part are the discretisation methods and their analysis. This led to the author's monograph *Theory and Numerical Treatment of Elliptic Differential Equations* also published by Springer. The second part is the treatment of the equations obtained by the discretisation process. The arising system of linear (or even nonlinear) equations is of large size, only bounded by the available storage of the computers. Nowadays, systems of several millions of equations and variables must be solved. Another characteristic of the arising systems is the sparsity of the system matrix; i.e., only $\mathcal{O}(n)$ entries of the $n \times n$ matrix are different from zero. The classical Gauss elimination requires up to $\mathcal{O}(n^3)$ operations. Because of the large size of n , algorithms of this complexity are hopeless. Even methods requiring a cost of $\mathcal{O}(n^2)$ take a too long run time. Instead, one needs solution algorithms of complexity $\mathcal{O}(n)$ or $\mathcal{O}(n \log^* n)$.

This book grew out of a series of lectures given by the author at the Christian Albrecht University of Kiel to students of mathematics. The first German edition was published in 1991 by Teubner, Stuttgart. The second German edition in 1993 mainly corresponds to the first English edition at Springer in 1994. Since that time new methods have developed. Therefore the present second edition differs significantly from the first one.

Although special attention is devoted to the modern effective algorithms (multigrid iterations, domain decomposition methods, and the hierarchical LU iteration), the theory of classical iterative methods should not be neglected. One reason is that these iterations indirectly re-appear in modern methods.

This volume requires basic mathematical knowledge in analysis and linear algebra. The necessary facts from linear algebra and matrix theory are summarised in the Appendices A–C of this book in order to provide as complete a presentation as possible and present a formulation and notation needed here. Similarly, the basics of finite element discretisation are summarised in Appendix E.

Part I covers the introduction and the classical linear iterations. Part II describes the semi-iterative methods including the popular conjugate gradient method. The subjects of these two parts should be understood as two orthogonal methods: a linear

iteration is accelerated by a semi-iterative approach. Part III contains more recent linear iterations.

The new Chapter 5 in Part I is devoted to the algebra in the set of linear iterations. These operations are important for the generation of new iterations. Part III contains two new chapters. Chapter 13 describes the \mathcal{H} -LU iteration which is based on the technique of hierarchical matrices introduced in Appendix D. In many cases, this iteration is a very efficient and robust method of black-box type. Finally, in Chapter 14, tensor-based iterative methods are briefly mentioned.

The discussion of the various methods is illustrated by many numerical examples, mostly for the Poisson model problem. Since these calculations are taken from the first edition, the problem sizes are small compared with modern computers. However, these sizes are completely sufficient to demonstrate the asymptotic behaviour.

The author also wishes to express his gratitude to the publisher Springer for their friendly cooperation.

Leipzig and Molfsee, October 2015

Wolfgang Hackbusch

Contents

Part I Linear Iterations

1	Introduction	3
1.1	Historical Remarks Concerning Iterative Methods	3
1.2	Model Problem: Poisson Equation	4
1.3	Notation	7
1.3.1	Index Sets, Vectors, and Matrices	7
1.3.2	Star Notation	9
1.4	A Single System Versus a Family of Systems	10
1.5	Amount of Work for the Direct Solution of a Linear System	10
1.6	Examples of Iterative Methods	12
1.7	Sparse Matrices Versus Fully Populated Matrices	15
2	Iterative Methods	17
2.1	Consistency and Convergence	17
2.1.1	Notation	17
2.1.2	Fixed Points	18
2.1.3	Consistency	19
2.1.4	Convergence	19
2.1.5	Convergence and Consistency	20
2.1.6	Defect Correction as an Example of an Inconsistent Iteration	20
2.2	Linear Iterative Methods	21
2.2.1	Notation, First Normal Form	21
2.2.2	Consistency and Second Normal Form	22
2.2.3	Third Normal Form	23
2.2.4	Representation of the Iterates x^m	23
2.2.5	Convergence	24
2.2.6	Convergence Speed	26
2.2.7	Remarks Concerning the Matrices M , N , and W	28
2.2.8	Three-Term Recursions, Two- and Multi-Step Iterations	29
2.3	Efficacy of Iterative Methods	30

2.3.1	Amount of Computational Work	30
2.3.2	Efficacy	31
2.3.3	Order of Linear Convergence	32
2.4	Test of Iterative Methods	32
2.4.1	Consistency Test	32
2.4.2	Convergence Test	33
2.4.3	Test by the Model Problem	34
2.4.4	Stopping Criterion	34
3	Classical Linear Iterations in the Positive Definite Case	35
3.1	Eigenvalue Analysis of the Model Problem	35
3.2	Traditional Linear Iterations	37
3.2.1	Richardson Iteration	37
3.2.2	Jacobi Iteration	38
3.2.3	Gauss–Seidel Iteration	39
3.2.4	SOR Iteration	41
3.3	Block Versions	42
3.3.1	Block Structure	42
3.3.2	Block-Jacobi Iteration	43
3.3.3	Block-Gauss–Seidel Iteration	44
3.3.4	Block-SOR Iteration	45
3.4	Computational Work of the Iterations	45
3.4.1	Case of General Sparse Matrices	45
3.4.2	Amount of Work in the Model Case	46
3.5	Convergence Analysis	47
3.5.1	Richardson Iteration	47
3.5.2	Convergence Criterion for Positive Definite Iterations	54
3.5.3	Jacobi Iteration	55
3.5.4	Gauss–Seidel and SOR Iterations	56
3.5.5	Convergence of the Block Variants	62
3.6	Convergence Rates in the Case of the Model Problem	62
3.6.1	Richardson and Jacobi Iteration	62
3.6.2	Block-Jacobi Iteration	63
3.6.3	Numerical Examples for the Jacobi Variants	65
3.6.4	SOR and Block-SOR Iteration with Numerical Examples	66
4	Analysis of Classical Iterations Under Special Structural Conditions	69
4.1	2-Cyclic Matrices	69
4.2	Preparatory Lemmata	72
4.3	Analysis of the Richardson Iteration	74
4.4	Analysis of the Jacobi Iteration	76
4.5	Analysis of the Gauss–Seidel Iteration	77
4.6	Analysis of the SOR Iteration	78
4.6.1	Consistently Ordered Matrices	78
4.6.2	Theorem of Young	81

4.6.3	Order Improvement by SOR	84
4.6.4	Practical Handling of the SOR Method	85
4.6.5	p -Cyclic Matrices	85
4.7	Application to the Model Problem	86
4.7.1	Analysis in the Model Case	86
4.7.2	Gauss–Seidel Iteration: Numerical Examples	87
4.7.3	SOR Iteration: Numerical Examples	88
5	Algebra of Linear Iterations	89
5.1	Adjoint, Symmetric, and Positive Definite Iterations	90
5.1.1	Adjoint Iteration	90
5.1.2	Symmetric Iterations	92
5.1.3	Positive Definite Iterations	93
5.1.4	Positive Spectrum of NA	95
5.2	Damping of Linear Iterations	95
5.2.1	Definition	95
5.2.2	Damped Jacobi Iteration	96
5.2.3	Accelerated SOR	97
5.3	Addition of Linear Iterations	97
5.4	Product Iterations	99
5.4.1	Definition and Properties	99
5.4.2	Constructing Symmetric Iterations	101
5.4.3	Symmetric Gauss–Seidel and SSOR	103
5.5	Combination with Secondary Iterations	103
5.5.1	First Example for Secondary Iterations	104
5.5.2	Second Example for Secondary Iterations	105
5.5.3	Convergence Analysis in the General Case	106
5.5.4	Analysis in the Positive Definite Case	108
5.5.5	Estimate of the Amount of Work	110
5.5.6	Numerical Examples	111
5.6	Transformations	112
5.6.1	Left Transformation	112
5.6.2	Right Transformation	115
5.6.3	Kaczmarz Iteration	116
5.6.4	Cimmoni Iteration	118
5.6.5	Two-Sided Transformation	119
5.6.6	Similarity Transformation	122
6	Analysis of Positive Definite Iterations	123
6.1	Different Cases of Positivity	123
6.2	Convergence Analysis	125
6.2.1	Case 1: Positive Spectrum	125
6.2.2	Case 2: Positive Definite NA	126
6.2.3	Case 3: Positive Definite Iteration	127
6.2.4	Case 4: Positive Definite $W+W^H$ or $N+N^H$	128

6.2.5	Case 5: Symmetrised Iteration \varPhi^{sym}	129
6.2.6	Case 6: Perturbed Positive Definite Case	131
6.3	Symmetric Gauss–Seidel Iteration and SSOR	132
6.3.1	The Case $A > 0$	132
6.3.2	SSOR in the 2-Cyclic Case	134
6.3.3	Modified SOR	135
6.3.4	Unsymmetric SOR Method	136
6.3.5	Numerical Results for the SSOR Iteration	136
7	Generation of Iterations	137
7.1	Product Iterations	137
7.2	Additive Splitting Technique	139
7.2.1	Definition and Examples	139
7.2.2	Regular Splittings	141
7.2.3	Applications	144
7.2.4	P-Regular Splitting	147
7.3	Incomplete Triangular Decompositions	148
7.3.1	Introduction and ILU Iteration	148
7.3.2	Incomplete Decomposition with Respect to a Star Pattern	151
7.3.3	Application to General Five-Point Formulae	152
7.3.4	Modified ILU Decompositions	154
7.3.5	Existence and Stability of the ILU Decomposition	154
7.3.6	Properties of the ILU Decomposition	159
7.3.7	ILU Decompositions Corresponding to Other Patterns	161
7.3.8	Approximative ILU Decompositions	162
7.3.9	Blockwise ILU Decomposition	163
7.3.10	Numerical Examples	163
7.3.11	Remarks	164
7.4	Preconditioning	165
7.4.1	Idea of Preconditioning	165
7.4.2	Examples	166
7.4.3	Preconditioning in the Wider Sense	167
7.4.4	Rules for Condition Numbers and Spectral Equivalence	167
7.4.5	Equivalent Bilinear Forms	170
7.5	Time-Stepping Methods	171
7.6	Nested Iteration	172

Part II Semi-Iterations and Krylov Methods

8	Semi-Iterative Methods	175
8.1	First Formulation	175
8.1.1	Notation	175
8.1.2	Consistency and Asymptotic Convergence Rate	176
8.1.3	Error Representation	177
8.1.4	Krylov Space	179

8.2	Second Formulation of a Semi-Iterative Method	181
8.2.1	General Representation	181
8.2.2	Three-Term Recursion	183
8.3	Optimal Polynomials	184
8.3.1	Minimisation Problem	184
8.3.2	Discussion of the Second Minimisation Problem	185
8.3.3	Chebyshev Polynomials	187
8.3.4	Chebyshev Method (Solution of the Third Minimisation Problem)	187
8.3.5	Order Improvement by the Chebyshev Method	192
8.3.6	Optimisation Over Other Sets	193
8.3.7	Cyclic Iteration	194
8.3.8	Two- and Multi-Step Iterations	195
8.3.9	Amount of Work of the Semi-Iterative Method	195
8.4	Application to Iterations Discussed Above	196
8.4.1	Preliminaries	196
8.4.2	Semi-Iterative Richardson Method	197
8.4.3	Semi-Iterative Jacobi and Block-Jacobi Method	198
8.4.4	Semi-Iterative SSOR and Block-SSOR Iteration	198
8.5	Method of Alternating Directions (ADI)	201
8.5.1	Application to the Model Problem	201
8.5.2	General Representation	203
8.5.3	ADI in the Commutative Case	205
8.5.4	ADI Method and Semi-Iterative Methods	208
8.5.5	Amount of Work and Numerical Examples	209
9	Gradient Method	211
9.1	Reformulation as Minimisation Problem	211
9.1.1	Minimisation Problem	211
9.1.2	Search Directions	212
9.1.3	Other Quadratic Functionals	213
9.1.4	Complex Case	214
9.2	Gradient Method	215
9.2.1	Construction	215
9.2.2	Properties of the Gradient Method	216
9.2.3	Numerical Examples	218
9.2.4	Gradient Method Based on Other Basic Iterations	219
9.2.5	Numerical Examples	223
9.3	Method of the Conjugate Directions	224
9.3.1	Optimality with Respect to a Direction	224
9.3.2	Conjugate Directions	225
9.4	Minimal Residual Iteration	228

10 Conjugate Gradient Methods and Generalisations	229
10.1 Preparatory Considerations	229
10.1.1 Characterisation by Orthogonality	229
10.1.2 Solvability	231
10.1.3 Galerkin and Petrov–Galerkin Methods	231
10.1.4 Minimisation	232
10.1.5 Error Statements	232
10.2 Conjugate Gradient Method	234
10.2.1 First Formulation	234
10.2.2 CG Method (Applied to Richardson’s Iteration)	237
10.2.3 Convergence Analysis	238
10.2.4 CG Method Applied to Positive Definite Iterations	241
10.2.5 Numerical Examples	244
10.2.6 Amount of Work of the CG Method	245
10.2.7 Suitability for Secondary Iterations	246
10.2.8 Three-Term Recursion for p^m	247
10.3 Method of Conjugate Residuals (CR)	250
10.3.1 Algorithm	250
10.3.2 Application to Hermitian Matrices	251
10.3.3 Stabilised Method of Conjugate Residuals	252
10.3.4 Convergence Results for Indefinite Matrices	253
10.3.5 Numerical Examples	255
10.4 Method of Orthogonal Directions	256
10.5 Solution of Nonsymmetric Systems	258
10.5.1 Generalised Minimal Residual Method (GMRES)	258
10.5.2 Full Orthogonalisation Method (FOM)	261
10.5.3 Biconjugate Gradient Method and Variants	262
10.5.4 Further Remarks	262

Part III Special Iterations

11 Multigrid Iterations	265
11.1 Introduction	266
11.1.1 Smoothing	266
11.1.2 Hierarchy of Systems of Equations	268
11.1.3 Prolongation	269
11.1.4 Restriction	271
11.1.5 Coarse-Grid Correction	272
11.2 Two-Grid Method	274
11.2.1 Algorithm	274
11.2.2 Modifications	274
11.2.3 Iteration Matrix	274
11.2.4 Numerical Examples	275
11.3 Analysis for a One-Dimensional Example	276
11.3.1 Fourier Analysis	276

11.3.2	Transformed Quantities	278
11.3.3	Convergence Results	279
11.4	Multigrid Iteration	281
11.4.1	Algorithm	281
11.4.2	Numerical Examples	282
11.4.3	Computational Work	284
11.4.4	Iteration Matrix	286
11.5	Nested Iteration	287
11.5.1	Discretisation Error and Relative Discretisation Error	287
11.5.2	Algorithm	288
11.5.3	Error Analysis	288
11.5.4	Application to Optimal Iterations	290
11.5.5	Amount of Computational Work	291
11.5.6	Numerical Examples	291
11.5.7	Comments	292
11.6	Convergence Analysis	293
11.6.1	Summary	293
11.6.2	Smoothing Property	293
11.6.3	Approximation Property	298
11.6.4	Convergence of the Two-Grid Iteration	301
11.6.5	Convergence of the Multigrid Iteration	301
11.6.6	Case of Weaker Regularity	303
11.7	Symmetric Multigrid Methods	304
11.7.1	Symmetric and Positive Definite Multigrid Algorithms	304
11.7.2	Two-Grid Convergence for $\nu_1 > 0, \nu_2 > 0$	306
11.7.3	Smoothing Property in the Symmetric Case	307
11.7.4	Strengthened Two-Grid Convergence Estimates	308
11.7.5	V-Cycle Convergence	310
11.7.6	Unsymmetric Multigrid Convergence for all $\nu > 0$	311
11.8	Combination of Multigrid Methods with Semi-Iterations	313
11.8.1	Semi-Iterative Smoothers	313
11.8.2	Damped Coarse-Grid Corrections	315
11.8.3	Multigrid as Basic Iteration of the CG Method	315
11.9	Further Comments	316
11.9.1	Multigrid Method of the Second Kind	316
11.9.2	Robust Methods	317
11.9.3	History of the Multigrid Method	317
11.9.4	Frequency Filtering Decompositions	318
11.9.5	Nonlinear Systems	320
12	Domain Decomposition and Subspace Methods	325
12.1	Introduction	325
12.2	Overlapping Subdomains	327
12.2.1	Introductory Example	327
12.2.2	Many Subdomains	329

12.3	Nonoverlapping Subdomains	329
12.3.1	Dirichlet–Neumann Method	329
12.3.2	Lagrange Multiplier Based Methods	330
12.4	Schur Complement Method	332
12.4.1	Nonoverlapping Domain Decomposition with Interior Boundary	332
12.4.2	Direct Solution	332
12.4.3	Preconditioners of the Schur Complement	334
12.4.4	Multigrid-like Domain Decomposition Methods	335
12.5	Subspace Iteration	336
12.5.1	General Construction	336
12.5.2	The Prolongations	337
12.5.3	Multiplicative and Additive Schwarz Iterations	338
12.5.4	Interpretation as Gauss–Seidel and Jacobi Iteration	339
12.5.5	Classical Schwarz Iteration	340
12.5.6	Approximate Solution of the Subproblems	340
12.5.7	Strengthened Estimate $A \leq \Gamma W$	342
12.6	Properties of the Additive Schwarz Iteration	344
12.6.1	Parallelism	344
12.6.2	Condition Estimates	344
12.6.3	Convergence Statements	347
12.7	Analysis of the Multiplicative Schwarz Iteration	349
12.7.1	Convergence Statements	349
12.7.2	Proofs of the Convergence Theorems	352
12.8	Examples	357
12.8.1	Schwarz Method With Proper Domain Decomposition	357
12.8.2	Additive Schwarz Iteration with Coarse-Grid Correction	358
12.8.3	Formulation in the Case of Galerkin Discretisation	358
12.9	Multigrid Iterations as Subspace Decomposition Method	359
12.9.1	Braess’ Analysis without Regularity	360
12.9.2	V-Cycle Interpreted as Multiplicative Schwarz Iteration	362
12.9.3	Proof of V-Cycle Convergence	364
12.9.4	Hierarchical Basis Method	366
12.9.5	Multilevel Schwarz Iteration	369
12.9.6	Further Approaches	369
13	\mathcal{H}-LU Iteration	371
13.1	Approximate LU Decomposition	371
13.1.1	Triangular Matrices	372
13.1.2	Solution of $LUX = b$	372
13.1.3	Matrix-Valued Solutions of $LX = Z$ and $XU = Z$	373
13.1.4	Generation of the LU Decomposition	375
13.1.5	Cost of the \mathcal{H} -LU Decomposition	376
13.2	\mathcal{H} -LU Decomposition for Sparse Matrices	376
13.2.1	Finite Element Matrices	376

13.2.2 Separability of the Matrix	377
13.2.3 Construction of the Cluster Tree	378
13.2.4 Application to Inversion	380
13.2.5 Admissibility Condition	381
13.2.6 LU Decomposition	381
13.3 UL Decomposition of the Inverse Matrix	381
13.4 \mathcal{H} -LU Iteration	382
13.4.1 General Construction	382
13.4.2 Algebraic LU Decomposition	384
13.5 Further Applications of Hierarchical Matrices	384
14 Tensor-based Methods	385
14.1 Tensors	385
14.1.1 Introductory Example: Lyapunov Equation	385
14.1.2 Nature of the Underlying Problems	386
14.1.3 Definition of Tensor Spaces	387
14.1.4 Case of Grid Functions	388
14.1.5 Kronecker Products of Matrices	389
14.1.6 Functions on Cartesian Products	389
14.2 Sparse Tensor Representation	390
14.2.1 r -Term Format (Canonical Format)	390
14.2.2 A Particular Example	391
14.2.3 Subspace Format (Tucker Format)	394
14.2.4 Hierarchical Tensor Format	395
14.3 Linear Systems	396
14.3.1 Poisson Model Problem	396
14.3.2 A Parametrised Problem	396
14.3.3 Solution of Linear Systems	398
14.3.4 CG-Type Methods	398
14.3.5 Multigrid Approach	398
14.3.6 Convergence	399
14.3.7 Parabolic Problems	399
14.4 Variational Approach	400
A Facts from Linear Algebra	401
A.1 Notation for Vectors and Matrices	401
A.2 Systems of Linear Equations	402
A.3 Eigenvalues and Eigenvectors	403
A.4 Block Vectors and Block Matrices	407
A.5 Orthogonality	409
A.5.1 Elementary Definitions	409
A.5.2 Orthogonal and Unitary Matrices	410
A.5.3 Sums of Subspaces and Orthogonal Complements	410
A.6 Normal Forms	411
A.6.1 Schur Normal Form	411

A.6.2	Jordan Normal Form	412
A.6.3	Diagonalisability	414
A.6.4	Singular Value Decomposition	416
B	Facts from Normed Spaces	417
B.1	Norms	417
B.1.1	Vector Norms	417
B.1.2	Equivalence of All Norms	418
B.1.3	Corresponding Matrix Norms	419
B.1.4	Condition and Spectral Condition Number	421
B.2	Hilbert Norm	422
B.2.1	Elementary Properties	422
B.2.2	Spectral Norm	422
B.3	Correlation Between Norms and Spectral Radius	424
B.3.1	Spectral Norm and Spectral Radius	424
B.3.2	Matrix Norms Approximating the Spectral Radius	425
B.3.3	Geometrical Sum of Matrices	426
B.3.4	Numerical Radius of a Matrix	427
C	Facts from Matrix Theory	431
C.1	Positive Definite Matrices	431
C.1.1	Definition and Notation	431
C.1.2	Rules and Criteria for Positive Definite Matrices	432
C.1.3	Remarks Concerning Positive Definite Matrices	433
C.2	Graph of a Matrix and Irreducible Matrices	435
C.3	Positive Matrices	438
C.3.1	Definition and Notation	438
C.3.2	Perron–Frobenius Theory of Positive Matrices	440
C.3.3	Diagonal Dominance	443
C.4	M-Matrices	445
C.4.1	Definition	445
C.4.2	M-Matrices and the Jacobi Iteration	446
C.4.3	M-Matrices and Diagonal Dominance	447
C.4.4	Further Criteria	449
C.5	H-Matrices	452
C.6	Schur Complement	452
D	Hierarchical Matrices	453
D.1	Introduction	453
D.1.1	Fully Populated Matrices	453
D.1.2	Rank- r Matrices	455
D.1.3	Model Format	456
D.2	Construction	459
D.2.1	Cluster Trees	459
D.2.2	Block Cluster Tree	462

D.2.3	Partition	462
D.2.4	Admissible Blocks	463
D.2.5	Use of Bounding Boxes for X_τ	464
D.2.6	Set of Hierarchical Matrices	465
D.2.7	\mathcal{H}^2 -Matrices	465
D.2.8	Storage	465
D.2.9	Accuracy	467
D.3	Matrix Operations	469
D.3.1	Matrix-Vector Multiplication	469
D.3.2	Truncations	470
D.3.3	Addition	470
D.3.4	Agglomeration	471
D.3.5	Matrix-Matrix Multiplication	471
D.3.6	Inversion and LU Decomposition	472
E	Galerkin Discretisation of Elliptic PDEs	473
E.1	Variational Formulation of Boundary Value Problems	473
E.2	Galerkin Discretisation	475
E.3	Subdomain Problems and Finite Element Matrix	477
E.4	Relations Between the Continuous and Discrete Problems	478
E.5	Error Estimates	480
E.6	Relations Between Two Discrete Problems	482
References	483
Index	501

List of Symbols and Abbreviations

Symbols

$\mathbf{1}$	vector $(1, 1, \dots, 1)^T$
A^T, A^H	transposed and Hermitian transposed matrix; cf. §A.1
A^{-T}, A^{-H}	inverse of A^T, A^H ; cf. §A.1
W^\perp	orthogonal complement of a W ; cf. §A.5
$U \oplus V$	direct sum of subspaces; cf. §A.5.3
$M _b$	restriction of the matrix to the block b ; cf. Notation D.6
$M ^b$	extension of the matrix to the block b ; cf. §D.3.4
Δ	Laplace operator; cf. (1.1a)
$\langle \cdot, \cdot \rangle$	(Euclidean) scalar product; cf. (1.1a–c)
$\langle \cdot, \cdot \rangle_A$	energy scalar product; cf. (C.5b)
$\ \cdot\ , \ \cdot\ $	norm (of vectors or matrices)
$\ \cdot\ _A$	energy norm; cf. (C.5a)
$\ \cdot\ _2$	Euclidean norm, cf. (B.2); spectral norm, cf. (B.21a)
$\ \cdot\ _\infty$	maximum norm, cf. (B.2); row sum norm, cf. (B.8)
$\ \cdot\ _{Y \leftarrow X}$	norm of a mapping (matrix) from X into Y ; cf. (B.11)
$\ \cdot\ _T$	transformed vector or matrix norm; cf. (B.10a,b)
$ \cdot $	absolute value, in §C.3 applied to matrices and vectors; cf. page 438
$<, \leq, >, \geq$	in connection with matrices, the order relation from §C.1.2; only in §§C.3–C.4 (and §7.3.5) it denotes the order relation of (C.9a,b)
$\dot{\cup}$	disjoint union
\subset	$A \subset B$: A is a subset of B , not necessarily a proper subset
\subsetneq	$A \subsetneq B$: A is a proper subset of B
\otimes	tensor product; cf. §14.1
\odot	Hadamard product; cf. Lemma 5.60c
\oplus_r	addition of hierarchical matrices with truncation to rank r ; cf. p. 457
\odot_r	multiplication of \mathcal{H} -matrices with truncation to rank r ; cf. §D.3.4
\circ	product $\Phi \circ \Psi$ of iterations or transformation of iterations; cf. §5
$\#S$	cardinality of a set S

Greek Letters

α, β, γ	indices of the index set; cf. §1.3
γ	in §11: number of secondary multi-grid steps for the coarse-grid equation; cf. (11.33d ₂)
γ, Γ	lower and upper eigenvalue bounds of $W^{-1}A$; cf. (9.18a)
δ_{ij}	Kronecker symbol: $\delta_{ij} = 1$ for $i = j$, $\delta_{ij} = 0$ otherwise
Δ	Laplace operator; cf. (1.1a)
ζ	often contraction number; cf. §2.2.6, (11.30b), (11.48)
η	characteristic factor involved in the admissibility condition (D.10)
$\eta(\nu)$	zero sequence for smoothing property; cf. (11.58b)
$\eta_0(\nu)$	special function, defined in Lemma 11.23
ϑ, Θ	damping factor; cf. §3.2.1 and §5.2
$\kappa(A)$	spectral condition number (B.13)
λ, Λ	eigenvalue bounds of A ; cf. Theorem 9.10, Theorem 3.30
$\lambda_{\max}(A)$	maximal eigenvalue of a matrix A if $\sigma(A) \subset \mathbb{R}$
$\lambda_{\min}(A)$	minimal eigenvalue of a matrix A if $\sigma(A) \subset \mathbb{R}$
ν, ν_1, ν_2	in §11: number of the smoothing steps; cf. (11.21) and (11.22a)
$\rho(A)$	spectral radius of a matrix A ; cf. Definition A.17
$\rho_{m+k,m}$	convergence factors; cf. (2.23a,b)
$\sigma(A)$	spectrum of the matrix A ; cf. §A.3
τ, σ	symbols representing clusters; cf. §D.2.1
$\Phi(x, b, A)$	function describing an iteration; cf. (2.3)
$\Upsilon[\Phi]$	semi-iterative method with Φ as basic iteration
$\Upsilon_{a,b}^{\text{Cheb}}$	Chebyshev method; cf. Notation 8.29
$\Upsilon_{\text{CG}}, \Upsilon_{\text{CR}}, \Upsilon_{\text{OD}}, \Upsilon_{\text{GMRES}}$	conjugate gradient methods and variants; cf. §10.2
Υ_{grad}	gradient method; cf. §9.2.1 and §9.2.4
ω	relaxation parameter; cf. (1.22) and §3.2.4
Ω	underlying domain of a boundary value problem; cf. (1.1a), §E.1
Ω_h	grid; cf. (1.3)

Latin Letters

$a(\cdot, \cdot)$	bilinear or sesquilinear form; cf. Definition (E.1)
a, b	bounds for $\sigma(M)$; cf. (8.26a)
A, A_ℓ	matrix of the linear system; cf. (1.5), (11.6a)
$A^{\kappa\lambda}, A^{ij}$	block of A ; cf. (A.8b,c)
$A_{\alpha\beta}, a_{\alpha\beta}, A_{ij}, a_{ij}$	entries of the matrix A
b, b_ℓ	right-hand side of the linear system; cf. (1.5), (11.6a)
$\text{blockdiag}\{\dots\}$	block-diagonal matrix; cf. page 408
$\text{blockdiag}_B\{A\}$	block-diagonal part of A with respect to the block structure B ; cf. (4.2')
$\text{blocktridiag}\{\dots\}$	block-tridiagonal matrix; cf. (A.9)

\mathbb{C}	complex numbers
cond, cond ₂	condition of a matrix; cf. (B.12)
d^m	defect $Ax^m - b$; cf. (2.17)
D, D', \dots	(block-) diagonal matrix
$\mathfrak{D}(\Phi)$	domain of the iteration Φ ; cf. Definition 2.2a
$\deg_X(v)$	degree of a vector v ; cf. Definition 8.10
degree(\cdot)	degree of a polynomial
depth(T)	depth of the tree T ; cf. (D.7)
det	determinant
diag{...}	diagonal matrix or diagonal part; cf. (A.1)
diam(τ)	diameter of a cluster; cf. (D.9a)
dist(τ, σ)	distance between clusters; cf. (D.9b)
e^m	error $x^m - x$ of the m -th iterate; cf. (2.15)
\mathbf{e}_α	unit vector
E	strictly lower triangular matrix; cf. (1.16)
Eff(Φ)	effective amount of work; cf. (2.31a)
F	strictly upper triangular matrix; cf. (1.16)
\mathcal{F}	matrices in full format; cf. Definition D.2a
$G(A)$	graph of a matrix A ; cf. Definition C.12
h, h_ℓ	grid size; cf. (1.2)
\mathcal{H}^2	see §D.2.7
\mathcal{H}_p	model format; cf. §D.1.3
i, j, k	indices of the ordered index set $I = \{1, \dots, n\}$
I	identity matrix
I	index set (not necessarily ordered)
I_κ	subset of block indices; cf. (A.7)
$Init(\Phi, A)$	cost for initialising the iteration Φ applied to the system $Ax = b$
$It(\Phi)$	cf. (2.30a)
\mathbb{K}	the field \mathbb{R} or \mathbb{C}
\mathbb{K}^I	space of the vectors corresponding to the index set I
$\mathbb{K}^{I \times I}$	space of the matrices corresponding to the index set I
\mathcal{K}	integral operator; cf. §D.2.9
$\mathcal{K}_m(X, v)$	Krylov space; cf. Definition 8.7
ker	kernel of a mapping or matrix
ℓ	level number in the discretisation hierarchy; cf. (3.15a)
L, L', \hat{L}	lower (block-)triangular matrix
\mathcal{L}	set of consistent linear iterations; cf. (2.11)
\mathcal{L}_{pos}	set of positive definite iterations; cf. Definition 5.8
$\mathcal{L}_{\text{semi}}$	set of positive semidefinite iterations; cf. Definition 5.11
\mathcal{L}_{sym}	set of symmetric iterations; cf. Definition 5.3
$\mathcal{L}_{>}$	set of directly positive definite iterations; cf. Definition 5.14
$\mathcal{L}(T)$	set of leaves of the tree T ; cf. §D.2.1
level(τ)	level-number of a cluster; cf. §D.2.1
log	natural logarithm
log ₂	dual logarithm, logarithm with respect to the basis 2

$\log^*(\cdot)$	some power of $\log(\cdot)$; cf. Footnote 9 on page 15
m	iteration number; cf. e^m, x^m
M, M^{xyz}	iteration matrix (of the iteration ‘xyz’); cf. §2.2.1
$M[A]$	iteration matrix for the system $Ax = b$; cf. Definition 2.9
n, n_ℓ	dimension of the linear system; cf. §2.3, (11.6b)
n_{\min}	minimal size of clusters; cf. §D.2.1
N	number of the grid points per row or column; cf. (1.2)
N, N^{xyz}	matrix of the 2nd normal form (of the iteration ‘xyz’); cf. (2.10)
$N[A]$	matrix N for the system $Ax = b$; cf. Definition 2.9
\mathbb{N}	natural numbers $\{1, 2, 3, \dots\}$
\mathbb{N}_0	$\mathbb{N} \cup \{0\} = \{0, 1, 2, \dots\}$
N_{xyz}	number of arithmetic operations required for ‘xyz’; cf. pages 455ff
$\mathcal{O}(\cdot)$	Landau symbol: $f(\alpha) = \mathcal{O}(g(\alpha))$ if $ f(\alpha) \leq C g(\alpha) $ for the underlying limit process $\alpha \rightarrow 0$ or $\alpha \rightarrow \infty$. The notation $f(\eta) = 1 - \mathcal{O}(\eta^\tau)$ is more special and means that $f(\eta) \leq 1 - C\eta^\tau$ with fixed $C > 0$ for $\eta \rightarrow 0$.
p	prolongation; cf. §11.1.3, (12.7)
P	partition of a hierarchical matrix; cf. §D.2.3
P^+, P^-	subsets of the partition P ; cf. §D.2.6
\mathcal{P}_m	space of polynomials; cf. Definition 8.2
Q	often unitary matrix
$Q_{\min}(\cdot)$	bounding box; cf. §D.2.1.2
r	restriction; cf. §11.1.4, (12.14a)
r	representation rank of matrices in \mathcal{R}_r ; cf. (D.2)
$r(A)$	numerical radius of the matrix A ; cf. §B.3.4
\mathbb{R}	real numbers
$\text{range}(\cdot)$	range (image space) of a mapping
\mathcal{R}_r	rank- r matrices or tensors; cf. Definition D.2b and page 390
$\text{root}(T)$	root of the tree T ; cf. §D.2.1
S_ℓ	iteration matrix of the smoother S_ℓ ; cf. Lemma 11.11
\mathcal{S}_ℓ	smoothing iteration; cf. §11.1.1 and §11.2.1
$\text{span}\{\dots\}$	linear space spanned by $\{\dots\}$
$\text{supp}(\cdot)$	support of a function; Footnote 6 on page 463
T_ℓ, T_r	left- and right-sided transformation; cf. (5.32), (5.39)
$T(I)$	cluster tree corresponding to the index set I ; cf. §D.2.1
$T^{(\ell)}(I)$	subset of $T(I)$; cf. (D.7)
$T(I \times J)$	block cluster tree corresponding to the index set I ; cf. §D.2.1
$\mathcal{T}_r, \mathcal{T}_r^{\mathcal{R}}, \mathcal{T}_r^{\mathcal{H}}, \mathcal{T}_{r;\text{pairw}}^{\mathcal{R}}$	truncation operator; cf. §D.3.2
$\text{tridiag}\{\dots\}$	tridiagonal matrix; cf. (A.2)
u_{ij}	components of the grid function u ; cf. (1.6b)
U, U', \hat{U}	upper (block-)triangular matrix
W, W_Φ	matrix of the third normal form (of the iteration Φ); cf. (2.12)
$W[A]$	matrix W for the system $Ax = b$; cf. Definition 2.9
$\text{Work}(\Phi, A)$	amount of work of the iteration Φ applied to $Ax = b$; cf. (2.29)
x	vector; often solution of the equation $Ax = b$

x^*	solution of the equation $Ax = b$ if the symbol x is used as a variable
x_ℓ, x_ℓ^*	vectors x, x^* at the level ℓ ; cf. (11.6a)
x^0	starting value of the iteration
x^m	m -th iterate
x^α, x^i	block of x corresponding to the index α or i ; cf. (A.8a)
x_α, x_i	components of a vector x
x, y	spatial variables $(x, y) \in \Omega$; cf. (1.1a)
X_τ	support of the cluster τ ; cf. §D.2.1.2 and (D.8)
\mathbb{Z}	set of integers

Abbreviations and Algorithms

ALS	alternating least squares method cf. §14.4
AMG	algebraic multigrid method
AMLI	algebraic multilevel iteration; cf. page 335
AOR	accelerated overrelaxation
ART	algebraic reconstruction technique
BCG, BiCG	biconjugate gradient method; cf. §10.5.3
BEM	boundary element method
Bi-CGSTAB	biconjugate gradient stabilised method; cf. §10.5.3
BPX	additive multigrid iteration; cf. §12.9.6
CG	method of conjugate gradients; cf. §10.2
CGS	conjugate gradient squared method; cf. §10.5.3
CR	method of conjugate residuals; cf. §10.3
DDM	domain decomposition method
FEM	finite element method
FFT	fast Fourier transform
FOM	full orthogonalisation method; cf. §10.5.2
GMRES	generalised minimal residual method; cf. §10.5.1
\mathcal{H} -matrix	hierarchical matrix
\mathcal{H} -LU	hierarchical LU decomposition
HOSVD	higher order singular value decomposition; cf. §14.2.3
MAOR	modified accelerated overrelaxation
MINRES	minimal residual method
MSOR	modified successive overrelaxation
OD	method of orthogonal directions; cf. §10.4
ORTHODIR, ORTHOMIN, ORTHORES	cf. §10.5.4
SAOR	symmetric accelerated overrelaxation
SIRT	simultaneous iterative reconstruction technique; cf. page 94
SOR	successive overrelaxation; cf. §3.2.4
SVD	singular value decomposition; cf. §A.6.4
SYMMLQ	symmetric LQ method; cf. page 257
USSOR	unsymmetric successive overrelaxation

Part I

Linear Iterations

The core of iterative methods for linear systems are *linear iterations*. Different from direct methods, an infinite sequence of iterates is produced. Since, in practice, only a finite number of iteration steps is performed, the unavoidable iteration error depends crucially on the speed of convergence.

Chapter 1 starts with historical remarks. It introduces the Poisson model problem which will be a test example for the iterative methods described later on. Vector and matrix notations are provided in §1.3. In the case of discretisations of partial differential equations, it is important to consider the family of systems obtained for different discretisation parameters (§1.4). A crucial question is whether the convergence speed deteriorates with increasing matrix size. To get a first idea of an iterative method, the Gauss–Seidel and SOR methods are presented in §1.6 with numerical results for the model problem. These examples involve sparse matrices (§1.7). Except for Chapter D, we shall always assume that the underlying matrices are sparse. This assumption ensures that the cost of one iteration step is proportional to the matrix size; however, sparsity is not needed for convergence analysis.

Chapter 2 introduces general iterative methods. The concepts of consistency and convergence are described in §2.1. The class of linear iterations is specified in §2.2. For its description three normal forms are introduced. A first important result is the convergence theorem in §2.2.5. The quality of a linear iteration depends on both cost and convergence speed. The resulting efficacy is discussed in §2.3. Section 2.4 demonstrates how to test iterative methods numerically.

The convergence of a linear iteration depends on the properties of the underlying matrix. Chapter 3 investigates classical iterations (Richardson, Jacobi, Gauss–Seidel, SOR) applied to positive definite matrices. The corresponding analysis of general linear iterations is presented in Chapter 6.

Chapter 4 considers classical iterations assuming other structural matrix properties. In particular, §4.6 contains Young’s theorem on SOR for consistently ordered matrices. It describes the improvement of the convergence order in explicit form.

The set of linear iterations forms an algebra containing various operations as described in Chapter 5. Section 5.1 introduces the definition of an adjoint iteration. This enables the construction of symmetric or even positive definite iterations. Damping of linear iterations is discussed in §5.2. Addition of linear iterations is the subject of §5.3, while the product of linear iterations is investigated in §5.4. Another combination of iterative methods is the secondary iteration (§5.5). The left, right, or two-sided transformations are studied in §5.6. Kaczmarz’ iteration (§5.6.3) and Cimmoni’s iteration (§5.6.4) can be obtained by suitable transformations.

Chapter 6 collects the convergence results for positive definite iterations (including possible perturbations of the positive definite matrix). In particular, the symmetric Gauss–Seidel method and symmetric SOR are studied (§6.3).

Chapter 7 is concerned with the generation of linear iterations. A classical technique is additive splitting of the underlying matrix (§7.2). Incomplete LU decomposition (ILU, §7.3) is another possibility to generate an iteration by matrix data only. Preconditioning in §7.4 is a particular case of a transformation aiming at improving the convergence.

Modern linear iterations will be treated in Part III.

Chapter 1

Introduction

I recommend this [iterative] method to you for imitation. You will hardly ever again eliminate directly, at least not when you have more than 2 unknowns. The indirect procedure can be done while half asleep, or while thinking about other things.

(C.F. Gauss in a letter to Gerling [148], Dec. 1823).

Abstract After some historical comments in Section 1.1, we introduce a model problem (Section 1.2) serving as a first test example of the various iterative methods. Deliberately, a simply structured problem is chosen since this allows us to determine all required quantities explicitly. The role of the ordering of the unknowns is explained. Often no ordering is needed. Section 1.3 introduces notation for vectors and matrices. Furthermore, the description of difference schemes by stencils is explained. Besides the behaviour of an iterative method for a single system, its behaviour with respect to a whole family of systems is often more interesting (Section 1.4). In Section 1.5, the cost of the direct solution by the Gauss elimination is determined. This cost can be compared with the cost of the iterative methods introduced later. In Section 1.6, the Gauss–Seidel and SOR iteration are presented as first examples of linear iterations. Finally, in Section 1.7, sparsity of the underlying matrix discussed.

1.1 Historical Remarks Concerning Iterative Methods

Iterative methods are almost 200 years old. The first iterative method for systems of linear equations is due to Carl Friedrich Gauß (simplified spelling: Gauss). His method of least squares led him to a system of equations that was too large for the use of direct Gauss elimination.¹ Today the iterative method described in Gauss [147]² would be called the blockwise Gauss–Seidel method. The value that Gauss attributed to his iterative method can be seen in the excerpt from his letter [148]³ at the top of the page.

¹ The Gauss elimination is known since ancient times; the Chinese text *Jiu Zhang Suanshu: Nine Chapters on the Mathematical Art* is written about 200 BC.

² A translation of the neo-Latin title is ‘Supplement to the theory on the combination of observations subject to minimal errors’.

³ See also the English translation by Forsythe [137].

Carl Gustav Jacobi [227]⁴ described a very similar method in 1845. In 1874 Phillip Ludwig Seidel, a student of Jacobi, wrote about ‘a method, to solve the equations arising from the least squares method as well as general linear equations by successive approximation’ [337].

Since the time that electronic computers became available for solving systems of equations, the number of equations has increased by many orders of magnitude and the methods mentioned above have proved to be too slow. After more than 100 years of stagnation in this field, Southwell [346, 347, 348, 349] experimented with variants of the Gauss–Seidel method⁵ and, in 1950, David M. Young, Jr. [411] succeeded in a breakthrough. His modification of the Gauss–Seidel method leads to an important acceleration of the convergence. This so-called SOR iteration will be described in §1.6 as an example of an iterative method. Since then, numerous other methods have been developed. The modern ones will be described in Part III.

Concerning a historical view to the development of iterative techniques, we recommend, e.g., the articles by Stiefel [353] (1952), Forsythe [138] (1953), Axelsson [10, §7.1] (1976), and Young [413] (1989).

1.2 Model Problem: Poisson Equation

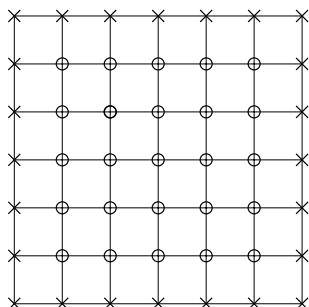


Fig. 1.1 Grid Ω_h with inner grid points (o) and boundary points (x).

During the time of Gauss, Jacobi and Seidel, the equations of the least squares method have led to a larger number of equations (e.g., obtained from geodesic measurements). Today, in particular the discretisations of partial differential equations give rise to systems of a large number of equations.

Since the discretisation error is smaller the larger the dimension of the system is, one is interested in systems of millions of unknowns⁶. In the following we shall often refer to a model problem representing the simplest nontrivial example of a boundary value problem. It is the Poisson equation with Dirichlet boundary values:

$$-\Delta u(x, y) = f(x, y) \quad \text{for } (x, y) \in \Omega, \quad (1.1a)$$

$$u(x, y) = \varphi(x, y) \quad \text{on } \Gamma = \partial\Omega. \quad (1.1b)$$

⁴ The English translation of the title is ‘On a new solution method of linear equations arising from the least squares method’.

⁵ Southwell used the term *relaxation*, since he considered the system of equations as a mechanical arrangement. The solution of the system characterises the equilibrium. Otherwise, forces act between nodes. One partial step of the Gauss–Seidel method leads to the local equilibrium in one node, i.e., this node is *relaxed*.

⁶ The concrete size is time dependent, since it increases with the available computer capacity.

Here $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ abbreviates the two-dimensional⁷ Laplace operator. As the underlying domain Ω , we choose the unit square

$$\Omega = (0, 1) \times (0, 1). \quad (1.1c)$$

In (1.1a,b), the *source term* f and the *boundary values* φ are given, while the function u is unknown.

To discretise the differential equation (1.1a–c), the domain Ω is covered with a grid of step size h (cf. Figure 1.1). Each grid point (x, y) has the representation $x = ih$, $y = jh$ ($0 < i, j < N$), where

$$h = 1/N. \quad (1.2)$$

More precisely, the grid is the set of *inner* grid points:

$$\Omega_h := \{(x, y) = (ih, jh) : 1 \leq i, j \leq N - 1\}. \quad (1.3)$$

We abbreviate the desired values $u(x, y) = u(ih, jh)$ with u_{ij} . An approximation of the differential equation (1.1a) is given by the *five-point formula*

$$h^{-2} [4u_{ij} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1}] = f_{ij} \quad (1.4a)$$

with $f_{ij} := f(ih, jh)$ for $1 \leq i, j \leq N - 1$. The left-hand side in (1.4a) coincides with $-\Delta u(ih, jh)$ up to a consistency error $\mathcal{O}(h^2)$ when a sufficiently smooth solution u of (1.1a,b) is inserted (cf. Hackbusch [193, §4.5]). For grid values on the boundary, i.e., for $i = 0, i = N, j = 0$, or $j = N$, the values u_{ij} are known from the boundary data (1.1b):

$$u_{ij} := \varphi(ih, jh) \quad \text{for } i = 0, i = N, j = 0, \text{ or } j = N. \quad (1.4b)$$

The number of the unknowns u_{ij} is $n := (N - 1)^2$ and corresponds of the number of the inner grid points. In order to form the system of equations, we have to eliminate the boundary values (1.4b), which possibly may appear in (1.4a). For instance, if $N \geq 3$, the equation corresponding to the index $(i, j) = (1, 1)$ reads as

$$h^{-2}[4u_{11} - u_{12} - u_{21}] = g_{11} \quad \text{with} \quad g_{11} := f(h, h) + h^{-2}[\varphi(0, h) + \varphi(h, 0)].$$

To write the equations in the common matrix formulation

$$Ax = b \quad (1.5)$$

with an $n \times n$ matrix A and n -dimensional vectors x and b with $n = (N - 1)^2$, one is forced to represent the doubly indexed unknowns u_{ij} by a singly indexed vector x . This implies that the (inner) grid points must be enumerated in some way.

⁷ The one-dimensional Laplace equation $-u'' = f$ leads to a too simple system which is not suited as a test example. The two-dimensional problem has already the typical properties. The three-dimensional counterpart would not be better.

	21	22	23	24	25	
	16	17	18	19	20	
	11	12	13	14	15	
	6	7	8	9	10	
	1	2	3	4	5	

	11	24	12	25	13	
	21	9	22	10	23	
	6	19	7	20	8	
	16	4	17	5	18	
	1	14	2	15	3	

Fig. 1.2 Left: lexicographical ordering of the grid points. Right: chequer-board ordering.

Figure 1.2 (left) shows the lexicographical ordering. The exact definition of the matrix A and of the right-hand side b can be seen from the following definition of the matrix A and of the vector b by the lexicographical ordering for the Poisson model problem with step size $h = 1/N$:

$A := 0;$ {all entries of A are initialised by zero} (1.6a) $k := 0;$ { $1 \leq k \leq n$ is the index with respect to the lexicographical ordering} for $j := 1$ to $N - 1$ do for $i := 1$ to $N - 1$ do begin $k := k + 1;$ $a_{kk} := 4 \cdot h^2;$ $b_k := f(ih, jh);$ if $i > 1$ then $a_{k-1,k} := -h^2$ else $b_k := b_k + h^2 \cdot \varphi(0, jh);$ if $i < N - 1$ then $a_{k+1,k} := -h^2$ else $b_k := b_k + h^2 \cdot \varphi(1, jh);$ if $j > 1$ then $a_{k,k-(N-1)} := -h^2$ else $b_k := b_k + h^2 \cdot \varphi(ih, 0);$ if $j < N - 1$ then $a_{k,k+(N-1)} := -h^2$ else $b_k := b_k + h^2 \cdot \varphi(ih, 1)$ end;

Vice versa, the solution x of $Ax = b$ has to be interpreted as

$$\begin{aligned} x_k &= u_{ij} = u(ih, jh) \text{ for} \\ k &= i + (j - 1)(N - 1) \end{aligned} \quad (1 \leq i, j \leq N - 1). \quad (1.6b)$$

When x is interpreted as a grid function, we use the notation u_{ij} or $u(x, y)$ with $x = ih$, $y = jh$.

Remark 1.1. The reformulation of the two-dimensionally ordered unknowns into a one-dimensionally ordered vector is rather unnatural. The reason should not be sought in the two-dimensional nature of the problem, but rather in the questionable idea of enumerating the vector components by indices 1 to n . We shall see that the matrix A will never be required in the full presentation (1.6a).

If, nevertheless, one wants to represent the matrix A as $(a_{ij})_{1 \leq i,j \leq N}$, A should be written as a block matrix. The vector x decomposes naturally into $N - 1$ blocks

$$x^j := \begin{bmatrix} x_{k+1} \\ \vdots \\ x_{k+N-1} \end{bmatrix} = \begin{bmatrix} u_{1,j} \\ \vdots \\ u_{N-1,j} \end{bmatrix} \quad \begin{array}{l} \text{with } k := (j-1)(N-1) \\ \text{for } j = 1, \dots, N-1, \end{array} \quad (1.7)$$

corresponding to the j -th row in the grid Ω_h . Accordingly, A takes the form of a block-tridiagonal matrix built from $(N-1) \times (N-1)$ blocks T , which again are tridiagonal $(N-1) \times (N-1)$ matrices:

$$A = h^{-2} \begin{bmatrix} T & -I & & & \\ -I & T & -I & & \\ & \ddots & \ddots & \ddots & \\ & & -I & T & -I \\ & & & -I & T \end{bmatrix}, \quad T = \begin{bmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 4 \end{bmatrix}. \quad (1.8)$$

I is the $(N-1) \times (N-1)$ identity matrix. Unmarked matrix entries or blocks are zeros or zero blocks, respectively. The representation (1.8) proves the next remark.

Remark 1.2. For the lexicographical ordering of the unknowns, the matrix A has a block-tridiagonal structure.

The lexicographical ordering is by no means the only ordering one can think of. Another frequently used approach is the *chequer-board ordering* (cf. Fig. 1.2, right).

In that case, the components u_{ij} with an even sum $i + j$ ('black squares') are enumerated first and thereafter those with an odd sum $i + j$ ('red squares') are numbered lexicographically. In the course of the next chapters further orderings will be mentioned. A broad collection of orderings of practical interest is given by Duff–Meurant [117].

Exercise 1.3. In the case of the chequer-board ordering, A decomposes into two blocks corresponding to the 'red' and 'black' indices. Prove that A has the block structure (1.9) with a rectangular submatrix B and identity matrices I_r , I_b whose block sizes are given by the numbers of the red and black grid points:

$$A = \begin{bmatrix} D_r & B \\ B^T & D_b \end{bmatrix}, \quad D_r = 4h^{-2}I_r, \quad D_b = 4h^{-2}I_b. \quad (1.9)$$

1.3 Notation

1.3.1 Index Sets, Vectors, and Matrices

According to Remark 1.1, the indices of the vectors are considered as unordered (unless we refer explicitly to a particular ordering). The (always finite) index set is denoted by I . The elements of I are often denoted by Greek letters, e.g., $\alpha \in I$.

In the case of the model problem, the indices $\alpha \in I$ are either the pairs $\alpha = (i, j)$ of the integers $1 \leq i, j \leq N - 1$ or the grid points $(x, y) = (ih, jh)$. We denote the *cardinality* of I , i.e., the number elements of I , by $\#I$.

In general, we use the field \mathbb{C} of complex numbers. This includes the standard case of the real field \mathbb{R} . For real matrices, the Hermitian transposed matrix A^H may be replaced by A^T . The neutral notation \mathbb{K} stands for \mathbb{R} or \mathbb{C} :

$$\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}. \quad (1.10)$$

A vector $b \in \mathbb{K}^I$ is a mapping $b : I \rightarrow \mathbb{K}$ into the field \mathbb{K} . The value of b at $\alpha \in I$ is denoted as vector component b_α . In programs, the notation $b[\alpha]$ is often used. If the index is a pair, e.g., $\alpha = (i, j)$, we write $b_{i,j} = b[i, j]$. A vector, composed of its components b_α , is written in the form

$$b = (b_\alpha)_{\alpha \in I}.$$

If the index set is ordered, we identify the indices with $1, 2, \dots, n := \#I$. While the indices $\alpha, \beta, \gamma, \dots$ are used for nonordered indices, we use Latin letters i, j, k, \dots in the ordered case.

In general, subscripts indicate the components of a vector. Sometimes, a subscript enumerates vectors; e.g., the first column vector of a matrix A may be written as a_1 . In order to avoid confusion with vector components, indexed vectors will be written in boldface as in the previous example. If not defined differently, e_α abbreviates the α -unit vector with the components $(e_\alpha)_\beta = \delta_{\alpha\beta}$. Here,

$$\delta_{\alpha\beta} = \begin{cases} 1 & \text{for } \alpha = \beta \\ 0 & \text{for } \alpha \neq \beta \end{cases} \quad (\alpha, \beta \in I) \quad (1.11)$$

is the Kronecker symbol.

Square matrices are mappings of the set $I \times I$ of index pairs into \mathbb{K} . The set of these matrices is denoted by $\mathbb{K}^{I \times I}$. Matrices are always symbolised by upper-case letters. The matrix entry of A corresponding to the index pair $(\alpha, \beta) \in I \times I$ is written as $a_{\alpha\beta}$ or $a_{\alpha,\beta}$, and occasionally as $A_{\alpha\beta}$. Alternatively, the notation $A[\alpha, \beta] = a[\alpha, \beta]$ is used. In particular, $(A + B)_{\alpha\beta}$, $(A^{-1})_{\alpha\beta}$, etc. is written for the components of matrix expressions. The matrix composed of the entries $a_{\alpha\beta}$ is denoted by

$$A = (a_{\alpha\beta})_{\alpha, \beta \in I}.$$

The symbol

$$I = (\delta_{\alpha\beta})_{\alpha, \beta \in I}$$

abbreviates the *identity matrix*, since it cannot be confused with the index set I .

In the case of rectangular (sub)matrices, the indices α and β belong to different sets I and J : $A = (a_{\alpha\beta})_{\alpha \in I, \beta \in J}$ is an $I \times J$ matrix. The set of these matrices is denoted by $\mathbb{K}^{I \times J}$.

For an ordered index set I , e.g., $I = \{1, \dots, n\}$, we use the standard index notation a_{ij} or A_{ij} .

1.3.2 Star Notation

In §1.2 the index set $I = \Omega_h$ is used. In the following, Ω_h can be more general than in (1.3). It may be an arbitrary subset of the two-dimensional infinite grid $\{(x, y) = (ih, jh) : i, j \in \mathbb{Z}\}$. The vector $x \in \mathbb{K}^I$ can be interpreted as a *grid function*, i.e., of a mapping defined at the grid points. Since the letter x represents the vector as well as the first component in the point $(x, y) \in \Omega_h$, we write u instead of $x \in \mathbb{K}^I$ in accordance with the equations (1.1a,b):

$$x_\alpha = u(x, y) \quad \text{for } \alpha = (x, y) \in I = \Omega_h. \quad (1.12)$$

If it seems to be more favourable, the argument $(x, y) = (ih, jh)$ is replaced with the indices ‘ ij ’:

$$u(ih, jh) = u_{ij} \quad \text{for } (ih, jh) \in \Omega_h.$$

The first index component x or i corresponds to the grid row (oriented from left to right), the second component y or j to the grid column (from the bottom to the top).

Mappings (matrices) defined in \mathbb{K}^I with $I = \Omega_h$ can conveniently be described by using the *star* or *stencil notation*. The *nine-point formula*

$$\begin{bmatrix} a_{-1,1} & a_{0,1} & a_{1,1} \\ a_{-1,0} & a_{0,0} & a_{1,0} \\ a_{-1,-1} & a_{0,-1} & a_{1,-1} \end{bmatrix} \quad (1.13a)$$

represents a matrix containing the nine coefficients a_{pq} ($-1 \leq p, q \leq 1$) of (1.13a) in each row. The component of Ax associated with the index $(ih, jh) \in \Omega_h$ is

$$\sum_{p,q=-1}^1 a_{pq} u_{i+p,j+q} \quad \text{or} \quad \sum_{p,q=-1}^1 a_{pq}^{ij} u_{i+p,j+q}, \quad (1.13b)$$

where $u = x$ according to (1.12). In the left part of (1.13b), the matrix entries are independent of the grid point (as, e.g., for the Poisson model problem), whereas, in the right part, they depend on $(ih, jh) \in \Omega_h$. $a_{00} = a_{00}^{ij}$ is the diagonal element $A_{(ij),(ij)}$ corresponding to the index ‘ ij ’. For example, the element $a_{1,0}$ in (1.13a) at the right position in the middle row is the matrix entry $A_{(i,j),(i+1,j)}$ by which the right neighbour $u_{i+1,j}$ —corresponding to the grid point $((i+1)h, jh) \in \Omega_h$ —has to be multiplied in (1.13b).

Although $(ih, jh) \in \Omega_h$, the index $(i + p, j + q)$ appearing in (1.13b)—more precisely the grid point $((i + p)h, (j + q)h)$ —may not belong to Ω_h . In this case, the term $a_{pq}^{ij} u_{i+p,j+q}$ in (1.13b) has to be ignored. The same effect is obtained by the formal definition $u_{i+p,j+q} := 0$.

The *five-point formula* of the Poisson model problem is

$$h^{-2} \begin{bmatrix} -1 \\ -1 & 4 & -1 \\ -1 \end{bmatrix}. \quad (1.14)$$

Unmarked entries a_{pq} (as at the positions $p, q = \pm 1$ in (1.14)) are defined by zero.

1.4 A Single System Versus a Family of Systems

Usually, a discretisation matrix A is embedded into a family $\{A_h\}_{h \in H}$. Here H is an infinite set with accumulation point $0 \in \overline{H}$. For instance, the Poisson model problem is defined for all $N \in \mathbb{N} \setminus \{1\}$ and the corresponding step sizes $h := \frac{1}{N} \rightarrow 0$. Statements about the convergence speed may be of the form $1 - \mathcal{O}(h^\kappa)$ (i.e., $\leq 1 - Ch^\kappa$ for some fixed C). Such expressions only make sense if there is a limit process $h \rightarrow 0$.

Given a family $\{A_\eta\}_{\eta \in F}$ of matrices, one is interested in the behaviour of the convergence rates (or of the computational cost for obtaining a certain accuracy) with respect to a limit process $\eta \rightarrow 0$ (or $\eta \rightarrow \infty$). If the iteration method leads to convergence estimates which are uniform with respect to η , we say that the iteration method is *robust* with respect to $\eta \in F$. A standard parameter is the discretisation size $h \rightarrow 0$, but it is not the only one. For instance, increasing anisotropy can be described by $A_\eta := B + \eta C$ for $\eta \rightarrow 0$, where B and C are discretisations of $\partial^2/\partial x^2$ and $\partial^2/\partial y^2$, respectively. Increasing convection is modelled by $A_\eta := \eta B + C$ ($\eta \rightarrow 0$), where C is a discretisations of a differential operator of first order.

1.5 Amount of Work for the Direct Solution of a Linear System

Methods are called *direct* if they terminate after finitely many operations with an exact solution (up to floating-point errors). The best known direct method is the Gauss elimination. In the case of the model problem in §1.2, one may perform this method without pivoting (cf. §C.4.4).

Concerning the valuation of the amount of computational work, we do not distinguish between additions, subtractions, multiplications, or divisions. Each is counted as one (arithmetic) operation. Traditionally, arithmetic operations for indices, data transfer, and similar activities are not counted (cf. Björck [48, §1.1.4]).

Remark 1.4. In the general case, the Gauss elimination solving a system $Ax = b$ of n equations requires $2n^3/3 + \mathcal{O}(n^2)$ operations. The storage amounts to $n^2 + n$.

Proof. During the i -th elimination step, the i -th row contains $n - i$ nonzero elements, whose multiples have to be subtracted from $n-i-1$ matrix rows. Summation of these $2(n-i)^2 + \mathcal{O}(n)$ operations over $1 \leq i \leq n$ yields the statement. \square

In the model case, $n = (N - 1)^2 = h^{-2} + \mathcal{O}(h^{-1})$ implies the following.

Conclusion 1.5. A naive application of the Gauss elimination to the model problem in §1.2 leads to $2N^6/3 + \mathcal{O}(N^5) = 2h^{-6}/3 + \mathcal{O}(h^{-5})$ operations and requires storage of $N^4 + \mathcal{O}(N^3) = h^{-4} + \mathcal{O}(h^{-3})$.

Halving the grid size h , yields the 64-fold computational work. Assuming one second for the solution of grid size h , the same computation for the quartered grid size $h/4$ consumes more than one hour!

However, the amount of work is less if the system matrix $A \in \mathbb{R}^{n \times n}$ is a *band matrix*. Here we assume the ordered index set $I = \{1, \dots, n\}$.

Definition 1.6. A is a band matrix of band width $w \in \mathbb{N}_0$ if $a_{ij} = 0$ holds for all $|i - j| > w$.

A band matrix has at maximum $2w$ nonvanishing off-diagonals besides the main diagonal. Concerning the properties of band matrices, we refer to Berg [44].

Remark 1.7. The matrix A arising from the model problem with lexicographical ordering according to (1.7) is a band matrix of band width $w = N - 1$.

The major part of the amount of work given in Remark 1.4 consists of unnecessary multiplications and additions by zeros. During the i -th elimination step the i -th row contains $w + 1$ nonzero elements. It is sufficient to eliminate the next w rows. This leads to $2w^2$ operations. In total, one obtains the next result.

Remark 1.8. The amount of work for the Gauss elimination without pivoting for solving a system with an $n \times n$ matrix of band width w amounts to

$$2nw^2 + \mathcal{O}(nw + w^3).$$

The storage requirement reduces to $2n(w + 1)$ when only the $2w + 1$ diagonals of A and the right-hand side b are stored.

Conclusion 1.9. *In the case of the model problem in §1.2, w is equal to $N - 1$. Therefore, the banded Gauss elimination requires $2N^4 + \mathcal{O}(N^3) = h^{-4} + \mathcal{O}(h^{-3})$ operations and storage of $2N^3 + \mathcal{O}(N^2)$.*

In the latter version, $2w + 1$ diagonals of A are used, although the matrix A in (1.8) has only five diagonals: the main diagonal, two side-diagonals at distance 1, and two further ones at distance $N - 1$. Unfortunately, one cannot exploit this property for the Gauss elimination.

Remark 1.10. The zeros in the second to $(N - 2)$ -th side-diagonals of the matrix A in (1.8) are completely filled during the elimination process by nonzeros (with the exception of the first block).

This occurrence is called *fill-in* and indicates a principal disadvantage of Gauss elimination when applied to sparse matrices. Here, we call an $n \times n$ matrix *sparse*, if the number of nonzero entries is by far smaller than n^2 . Otherwise, the matrix is called a *fully populated* or *dense* matrix. Because of the equivalence of Gauss elimination to the triangular or LU decomposition (cf. Quarteroni–Sacco–Saleri [314, §3], Björck [48, §1.2]), the same difficulties holds for the LU decomposition.

Conclusion 1.11. The decomposition $A = LU$ into a lower triangular matrix L and an upper triangular matrix U for the sparse matrix A in (1.8) yields factors L and U , which are full band matrices of width $w = N - 1$. The same holds for Cholesky decomposition.

There are special direct methods solving the system described in §1.2 with an amount of work between $\mathcal{O}(n) = \mathcal{O}(N^2)$ and $\mathcal{O}(n \log n) = \mathcal{O}(N^2 \log N)$. Examples are the *Buneman algorithm* and the *method of total reduction*, both described in Meis–Marcowitz [281, 282] (see also Bank [25], Bjørstad [49], Buneman [87], Buzbee et al. [90], Duff–Erisman–Reid [116], Golub [154], Hockney [223], and Schröder–Trottenberg [333]).

1.6 Examples of Iterative Methods

For the iterative solution of a system, one starts with an arbitrary starting vector x^0 and computes a sequence of iterates x^m for $m = 1, 2, \dots$:

$$x^0 \mapsto x^1 \mapsto x^2 \mapsto \dots \mapsto x^m \mapsto x^{m+1} \mapsto \dots$$

In the following, x^{m+1} is only dependent on x^m , so that the mapping $x^m \mapsto x^{m+1}$ determines the iteration method. The choice of the starting value x^0 is not part of the iteration method.

The already mentioned Gauss–Seidel iteration for solving the system $Ax = b$ reads as follows:

$$\text{for } i := 1 \text{ to } n \text{ do } x_i^{m+1} := \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{m+1} - \sum_{j=i+1}^n a_{ij} x_j^m \right) / a_{ii}. \quad (1.15)$$

Remark 1.12. (a) The Gauss–Seidel iteration (1.15) can be performed whenever all diagonal entries satisfy $a_{ii} \neq 0$.

(b) During the execution of the iteration, the variable x_i^m may be overwritten by the new value x_i^{m+1} .

(c) Different orderings (e.g., lexicographical or chequer-board ordering) yield different results.

Each matrix A can uniquely be decomposed into the sum

$$A = D - E - F, \quad \left\{ \begin{array}{l} D \text{ diagonal matrix,} \\ E \text{ strictly lower triangular matrix,} \\ F \text{ strictly upper triangular matrix.} \end{array} \right\} \quad (1.16)$$

Here, E is called a *lower triangular matrix* if $E_{ij} = 0$ for $j > i$, and a *strictly lower triangular matrix*, if $E_{ij} = 0$ for $j \geq i$. The (strictly) upper triangular matrix is defined analogously. The system of equations $Ax = b$ is equivalent to

$$(D - E)x = b + Fx. \quad (1.17)$$

Replacing x by x^m on the right-hand side and by x^{m+1} on the left-hand side, we obtain the iterative description (1.18a) or (1.18b):

$$(D - E)x^{m+1} = b + Fx^m, \quad \text{i.e.,} \quad (1.18a)$$

$$x^{m+1} = (D - E)^{-1}(b + Fx^m). \quad (1.18b)$$

Exercise 1.13. Prove that (1.18a,b) and (1.15) are equivalent, i.e., (1.18a) and (1.18b) are the vector representations of the Gauss–Seidel iteration, while (1.15) is the componentwise representation.

For a sparse matrix, one has to avoid defining A, b by (1.6a) and applying (1.15). For the model problem (1.4a,b), we should use the original data $f_{ij} = f[i, j]$ in (1.4a) and the boundary data $\varphi(ih, jh) = u_{ij} = u[i, j]$, which are stored at the boundary points of the array u . According to (1.6b), we use the variables u and f instead of x and b . The lexicographical Gauss–Seidel method for the model problem then takes the following form:

```

procedure GaussSeidel( $u, f$ );
begin for  $j := 1$  to  $N - 1$  do for  $i := 1$  to  $N - 1$  do
     $u[i, j] := (h^2 \cdot f[i, j] + u[i - 1, j] + u[i + 1, j] + u[i, j - 1] + u[i, j + 1]) / 4$ 
end; {lexicographical ordering}

```

In the double loop of (1.19), the matrix A is explicitly represented by its nonzero entries. The indexing is based on the ‘natural’ double indices. The lexicographical ordering of the grid points is a consequence of the arrangement of the loops. For the chequer-board ordering, the loop in (1.19) can be changed as follows:

```

 $w := 2$ ; {red squares} for  $j := 1$  to  $N - 1$  do (1.20)
begin  $w := 3 - w$ ; for  $i := w$  step 2 to  $N - 1$  do
     $u[i, j] := (h^2 \cdot f[i, j] + u[i - 1, j] + u[i + 1, j] + u[i, j - 1] + u[i, j + 1]) / 4$ 
end;
 $w := 1$ ; {black squares} for  $j := 1$  to  $N - 1$  do ... {same loop as in lines 2–4}

```

Since one may immediately store $h^2 \cdot f[i, j]$ instead of $f[i, j]$, the next remark follows.

Remark 1.14. In the case of the model problem, the Gauss–Seidel method (independently of the ordering) requires $5n$ operations ($4n$ additions and n divisions) per iteration.

Remark 1.15. In the case of $f_{ij} = -4$ and $\varphi(x, y) = x^2 + y^2$, the corresponding solution is $u_h(x, y) = x^2 + y^2$, i.e., $u_{ij} = (i^2 + j^2)h^2$. The simplest starting values are $u_{ij}^0 = 0$. In the following, the system (1.18a) with these data will be called the *Poisson model problem*. In the course of the next chapters, various iterative methods will be tested using this example.

Table 1.1 shows the error

$$\varepsilon_m := \max \{ |u_{ij}^m - (i^2 + j^2)h^2| : 1 \leq i, j \leq N-1 \}$$

of the m -th iterate for $h = \frac{1}{32}$ and the value $u_{16,16}^m$ at the midpoint $(16h, 16h) = (\frac{1}{2}, \frac{1}{2})$ of Ω . The values $u_{16,16}^m$ should converge to

$$u\left(\frac{1}{2}, \frac{1}{2}\right) = 0.5.$$

The listed values indicate the convergence of the Gauss–Seidel method, but its slowness is disappointing. After 100 iterations the first decimal of $u_{16,16}^m$ is still completely wrong! The third column contains the so-called *reduction factor*: the ratio $\varepsilon_{m-1}/\varepsilon_m$ of the successive errors. This factor indicates how fast the error decreases per iteration. A comparison of the data in Table 1.1 demonstrates that the ordering influences the results, but not the convergence speed.

The Gauss–Seidel iteration (1.15) is equivalent to the representation

$$\textbf{for } i := 1 \textbf{ to } n \textbf{ do } x_i^{m+1} := x_i^m - \left[\sum_{j=1}^{i-1} a_{ij}x_j^{m+1} + \sum_{j=i}^n a_{ij}x_j^m - b_i \right] / a_{ii}, \quad (1.21)$$

showing that the new iterate x_i^{m+1} is obtained from x_i^m by subtracting a correction. In contrast to (1.15), the second sum in (1.21) starts at $j = i$. A seemingly insignificant modification is the multiplication of this

correction by a factor ω . The obtained method is called the *successive over-relaxation method* and abbreviated with *SOR*. In the general case, it takes the form

m	lexicographical ordering			chequer-board ordering		
	$u_{16,16}^m$	ε_m	$\varepsilon_{m-1}/\varepsilon_m$	$u_{16,16}^m$	ε_m	$\varepsilon_{m-1}/\varepsilon_m$
0	0.0	1.877	-	0.0	1.877	-
1	-0.002	1.760	0.93756	-0.001	1.759	0.93704
2	-0.004	1.646	0.93563	-0.003	1.589	0.90323
9	-0.018	1.276	-	-0.017	1.202	-
10	-0.019	1.246	0.97637	-0.019	1.165	0.96903
99	+0.1102	0.404	-	+0.1353	0.380	-
100	+0.1135	0.400	0.98989	+0.1385	0.376	0.98994
199	+0.3479	0.152	-	+0.3585	0.142	-
200	+0.3494	0.151	0.99041	+0.3598	0.140	0.99041
299	+0.4421	0.058	-	+0.4461	0.054	-
300	+0.4426	0.057	0.99039	+0.4466	0.053	0.99039

Table 1.1 Results of the Gauss–Seidel iteration for $N = 32$.

m	$u_{16,16}^m$	ε_m	$\frac{\varepsilon_{m-1}}{\varepsilon_m}$	m	$u_{16,16}^m$	ε_m	$\frac{\varepsilon_{m-1}}{\varepsilon_m}$
0	0.0	1.877	-	39	0.4805	0.050	-
1	-0.016	1.777	0.9468	40	0.4838	0.043	0.8566
2	-0.027	1.680	0.9451	49	0.4964	0.0055	-
9	-0.065	1.046	-	50	0.4970	0.0049	0.8830
10	-0.068	0.962	0.9197	99	0.4999996	$9.05_{10}-7$	-
19	0.1111	0.399	-	100	0.4999997	$7.23_{10}-7$	0.7977
20	0.1486	0.365	0.9155	129	$0.5-1.5_{10}-9$	$3.57_{10}-9$	-
29	0.4198	0.166	-	130	$0.5-1.2_{10}-9$	$2.81_{10}-9$	0.7881

Table 1.2 SOR (lexicographical ordering, $N = 32$, $\omega = 1.821465$).

$$\text{for } i := 1 \text{ to } n \text{ do } x_i^{m+1} := x_i^m - \frac{\omega}{a_{ii}} \left[\sum_{j=1}^{i-1} a_{ij} x_j^{m+1} + \sum_{j=i}^n a_{ij} x_j^m - b_i \right]. \quad (1.22)$$

In the model case, the only change in (1.19) or (1.20) is the replacement of the assignment $u[i, j] := (\dots)/4$ by

$$u[i, j] := u[i, j] - \frac{\omega}{4} [4u[i, j] - u[i-1, j] - u[i+1, j] - u[i, j-1] - u[i, j+1] - h^2 \cdot f[i, j]].$$

In §4.6 we shall prove that $\omega = 2/(1 + \sin(\pi h))$ (i.e., $\omega = 1.821\dots$ for $N = 32$) is a suitable value. Table 1.2 shows the errors ε_m of the first 150 iterations for the same example as above. Convergence is evidently much faster than for the Gauss–Seidel method. Analysis of the above mentioned methods and constructing even faster iterations are the foci of the next chapters.

1.7 Sparse Matrices Versus Fully Populated Matrices

The matrix of the Poisson model problem is an example of a sparse matrix. Discretisation by finite differences and finite elements (cf. §E.2) generates sparse matrices with the property that the number of nonzero entries per row is bounded,⁸ i.e., the number

$$s(I) := \max_{\alpha \in I} \#\{\beta \in I : a_{\alpha\beta} \neq 0\} \quad (1.23)$$

is bounded independently of the matrix size $n = \#I$. This property is important for the storage cost which is $\mathcal{O}(n)$. Operations as matrix–vector multiplication and most of the operations involved by one step of an iteration method also have a computational cost of $\mathcal{O}(n)$.

Formally, the iterative schemes also work for fully populated matrices. In this case, the storage cost is n^2 and basic operations as matrix–vector multiplication cost $\mathcal{O}(n^2)$ arithmetic operations. For large-scale matrices, the quadratic order $\mathcal{O}(n^2)$ is too large. Fortunately, there are other techniques which allow reducing $\mathcal{O}(n^2)$ to⁹ $\mathcal{O}(n \log^* n)$ or even $\mathcal{O}(n)$. Appendix D will describe such a method.

There is a further reason why in the following we focus to sparse matrices. Fully populated matrices typically arise from discretising nonlocal operators (integral operators), e.g., by the boundary element method (cf. Sauter–Schwab [331]). The corresponding linear systems have other properties than, e.g., the Poisson model problem, and require other types of iterative methods (see the Picard iteration and the multigrid iteration of the second kind in §11.9.1).

Assume that A is a sparse matrix satisfying (1.23). In the regular case of a difference scheme, the data of A are organised by a nine-point star (1.13a) or a five-point formula (1.14). If the coefficients are constant as in the left-hand side of (1.13a) or

⁸ For the finite element method, this is a consequence of the shape regularity of the triangulation.

⁹ The asterix in $\log^* n$ indicates an unspecified power.

in (1.14), the data size is negligible. Procedure (1.19) shows how easily these data can be used.

However, the standard case are more general sparse matrices arising from finite element discretisations. In this case, the sparse matrix format is used for organising the matrix data. For each $i \in I$, there is a subset

$$I_\alpha := \{\beta \in I : a_{\alpha\beta} \neq 0\}$$

whose size is bounded by $s(I)$ (cf. (1.23)). The entries $a_{\alpha\beta} \neq 0$ (α fixed) form the vector $\mathbf{a}_\alpha \in \mathbb{K}^{I_\alpha}$. Then the matrix data are described by the set

$$\{(I_\alpha, \mathbf{a}_\alpha) : \alpha \in I\},$$

which can be implemented by a list. For instance, the SOR iteration (1.22) reads as

$$\text{for } \alpha \in I \text{ do } u[\alpha] := u[\alpha] - \frac{\omega}{\mathbf{a}_\alpha[\alpha]} \left[\sum_{\beta \in I_\alpha} \mathbf{a}_\alpha[\beta] \cdot u[\beta] - h^2 \cdot f[\alpha] \right],$$

where the loop follows the ordering of I .

Chapter 2

Iterative Methods

Abstract In this chapter we consider general properties of iterative methods. Such properties are *consistency*, ensuring the connection between the iterative method and the given system of equations, as well as *convergence*, guaranteeing the success of the iteration. The most important result of this chapter is the characterisation of the convergence of linear iterations by the spectral radius of the iteration matrix (cf. §2.1.4). Since we only consider iterative methods for systems with regular matrices, iterative methods for singular systems or those with rectangular matrices will not be studied.¹ The quality of a linear iteration depends on both the cost and the convergence speed. The resulting efficacy is discussed in Section 2.3. Finally, Section 2.4 explains how to test iterative methods numerically.

2.1 Consistency and Convergence

2.1.1 Notation

We want to solve the *system of linear equations*

$$Ax = b \quad (A \in \mathbb{K}^{I \times I} \text{ and } b \in \mathbb{K}^I \text{ given}) \quad (2.1)$$

(cf. (1.10)). To guarantee solvability for all $b \in \mathbb{K}^I$, we generally assume:

$$A \text{ is regular.} \quad (2.2)$$

An iterative method producing iterates x^1, x^2, \dots from the starting value x^0 can be characterised by a prescription $x^{m+1} := \Phi(x^m)$. Φ depends on the data A and b in (2.1). These parameters are explicitly expressed by the notation

¹ Concerning this topic, we refer, e.g., to Björck [47], Marek [275], Kosmol–Zhou [241], Berman–Plemmons [46], and Remark 5.17.

$$x^{m+1} := \Phi(x^m, b, A) \quad (m \geq 0, b \text{ in (2.1)}). \quad (2.3)$$

Since in most of the cases the matrix A is fixed, we usually write

$$x^{m+1} := \Phi(x^m, b)$$

instead of $\Phi(x^m, b, A)$. By $\Phi(\cdot, \cdot, A)$ we express the fact that we consider the iteration (2.3) exclusively for the matrix A .

Definition 2.1. An *iterative method* is a (in general nonlinear) mapping

$$\Phi : \mathbb{K}^I \times \mathbb{K}^I \times \mathbb{K}^{I \times I} \rightarrow \mathbb{K}^I.$$

By $x^m = x^m(x^0, b, A)$ we denote the *iterates* of the sequence generated by the prescription (2.3) with a starting value $x^0 = y \in \mathbb{K}^I$:

$$\begin{aligned} x^0(y, b, A) &:= y, \\ x^{m+1}(y, b, A) &:= \Phi(x^m(y, b, A), b, A) \quad \text{for } m \geq 0. \end{aligned} \quad (2.4)$$

If A is fixed, we write $x^m(y, b)$ instead of $x^m(y, b, A)$. If all parameters y, b, A are fixed, we write x^m .

If Φ is called an *iteration method*, we expect that the method is applicable to a whole class of matrices A . Here ‘applicable’ means that Φ is well defined (including the case that the sequence x^m diverges).

Definition 2.2. (a) $\mathfrak{D}(\Phi) := \{A : \Phi(\cdot, \cdot, A) \text{ well defined}\}$ is the *domain of Φ* .
(b) An iteration is called *algebraic* if the definition of $\Phi(\cdot, \cdot, A)$ can be based exclusively on the data of $A \in \mathfrak{D}(\Phi)$.

In the case of the Gauss–Seidel iteration Φ^{GS} in (1.15), the domain is defined by $\mathfrak{D}(\Phi^{\text{GS}}) = \{A \in \mathbb{K}^{I \times I} : a_{ii} \neq 0 \text{ for all } i \in I, I \text{ finite}\}$. Another extreme case is $\mathfrak{D}(\Phi) = \{A\}$, i.e., the iteration can only be applied to one particular matrix A .

2.1.2 Fixed Points

Definition 2.3. $x^* = x^*(b, A)$ is called a *fixed point* of the iteration Φ corresponding to $b \in \mathbb{K}^I$ and $A \in \mathfrak{D}(\Phi)$ (or shortly: a fixed point of $\Phi(\cdot, b, A)$) if

$$x^* = \Phi(x^*, b, A).$$

If the sequence $\{x^m\}$ of the iterates generated by (2.3) converges, we may form the limit in (2.3) and obtain the next lemma.

Lemma 2.4. Let the iteration Φ be continuous with respect to the first argument. If

$$x^* := \lim_{m \rightarrow \infty} x^m(y, b, A) \quad (\text{cf. (2.4)})$$

exists, x^* is a fixed point of $\Phi(\cdot, b, A)$.

2.1.3 Consistency

Lemma 2.4 states that possible results of the iteration method have to be sought in the set of fixed points. Therefore, a minimum condition is that the solution of system (2.1) with the right-hand side $b \in \mathbb{K}^I$ be a fixed point with respect to b . This property is the subject of the following definition.

Definition 2.5 (consistency). The iterative method Φ is called *consistent* to the system (2.1) with $A \in \mathfrak{D}(\Phi)$ if, for all right-hand sides $b \in \mathbb{K}^I$, any solution of $Ax = b$ is a fixed point of $\Phi(\cdot, b, A)$.

According to Definition 2.5, consistency means: For all $b, x \in \mathbb{K}^I$ and all matrices $A \in \mathfrak{D}(\Phi)$, the implication $Ax = b \Rightarrow x = \Phi(x, b, A)$ holds. The reverse implication would yield an *alternative* (nonequivalent) form of consistency:

$$Ax = b \quad \text{for all fixed points } x \text{ of } \Phi(\cdot, b, A) \text{ and for all } b \in \mathbb{K}^I, A \in \mathfrak{D}(\Phi). \quad (2.5)$$

Note that both variants of consistency do not require the regularity assumption (2.2). Even without (2.2), there may be a solution of $Ax = b$ for certain b . Then Definition 2.5 implies the existence of a fixed point of $\Phi(\cdot, b)$. Vice versa, (2.5) states the existence of a solution of $Ax = b$ as soon as $\Phi(\cdot, b, A)$ has a fixed point. The regularity of A will be discussed in Theorem 2.8.

2.1.4 Convergence

A natural definition of the convergence of an iterative method Φ seems to be

$$\lim_{m \rightarrow \infty} x^m(y, b, A) \quad \text{exists for all } y, b \in \mathbb{K}^I, \quad (2.6)$$

where $x^m(y, b, A)$ are the iterates defined in (2.4) corresponding to the starting value $x^0 := y$, while $A \in \mathfrak{D}(\Phi)$ is a fixed matrix. Since the starting value may be chosen arbitrarily, it may happen that an iteration satisfying (2.6) converges, but to different limits *depending* on the starting value. Therefore, the independence of the limit has to be incorporated into the definition of convergence. This yields the following definition, which is stronger than (2.6).

Definition 2.6. Fix $A \in \mathfrak{D}(\Phi)$. An iterative method $\Phi(\cdot, \cdot, A)$ is called *convergent* if for all $b \in \mathbb{K}^I$, there is a limit $x^*(b, A)$ of the iterates (2.4) independent of the starting value $x^0 = y \in \mathbb{K}^I$.

Note that consistency is a property of Φ for *all* $A \in \mathfrak{D}(\Phi)$, whereas convergence is required for a particular $A \in \mathfrak{D}(\Phi)$. Therefore $\Phi(\cdot, \cdot, A)$ may be convergent for some A , while $\Phi(\cdot, \cdot, A')$ diverges for another A' .

2.1.5 Convergence and Consistency

Remark 2.7. In the following, we shall often assume that the iterative method Φ is *convergent and consistent*. The term ‘convergent and consistent’ refers to a matrix $A \in \mathfrak{D}(\Phi)$ and means precisely: Φ is consistent and, for $A \in \mathfrak{D}(\Phi)$, the particular iteration $\Phi(\cdot, \cdot, A)$ is convergent.

It will turn out that the chosen definitions of the terms ‘convergence’ and ‘consistency’ of Φ are almost equivalent to the combination of the alternative definitions in (2.5) and (2.6).

Theorem 2.8. *Let Φ be continuous in the first argument. Then Φ is consistent and convergent if and only if A is regular and Φ fulfils the conditions (2.5) and (2.6).*

Proof. (i) Assume Φ to be consistent and convergent. (2.6) follows from Definition 2.6. If A is singular, the equation $Ax = 0$ would have a nontrivial solution $x^{**} \neq 0$ besides $x^* = 0$. By consistency, both are fixed points of Φ with respect to $b = 0$. Therefore, choosing the starting values $x^0 = x^*$ and $x^0 = x^{**}$, we obtain the constant sequences $x^m(x^*, 0) = x^*$ and $x^m(x^{**}, 0) = x^{**}$. The convergence definition states that the limits x^* and x^{**} coincide contrary to the assumption. Hence, A is regular. It remains to prove (2.5). The preceding argument shows that a convergent iterative method can have only *one* fixed point with respect to b . Because of the regularity of A , there is a solution of $Ax = b$ that, thanks to consistency, is the unique fixed point of Φ with b . Hence, (2.5) is proved.

(ii) Assume $\Phi(x, b)$ to be continuous in x and that (2.5) and (2.6) are fulfilled. Furthermore, let A be regular. Due to Lemma 2.4, $x^* := \lim x^m(y, b)$ is a fixed point of Φ with respect to b and therefore, by (2.5), a solution of $Ax = b$. Because of the regularity of A , the solution of the system is unique and hence also the limit of $x^m(y, b)$, which thereby cannot depend on y . Hence, Φ is convergent in the sense of Definition 2.6. Convergence leads to the uniqueness of the fixed point with respect to b (cf. part (i)). Since, by (2.5), this fixed point is the uniquely determined solution of $Ax = b$, Φ is consistent. \square

2.1.6 Defect Correction as an Example of an Inconsistent Iteration

In this monograph, all iterations will be assumed to be consistent. Usually, inconsistent iterations are an involuntary consequence of a bug in the implementation. However, there are examples where inconsistent iterations are of practical relevance. Assume that both $Ax = b$ and $By = c$ are discretisations of the same partial differential equation. Assume further that $Ax = b$ is simpler to solve than $By = c$, but the error of the discretisation by B is smaller than the discretisation error of A . Then there are combinations of both discretisations so that the overall treatment is as simple as for A but yielding the accuracy of B .

The standard defect correction $x^{m+1} = x^m - A^{-1}(Bx^m - c)$ can be stopped after a few iteration steps since the desired discretisation accuracy is reached (cf. [194, §14.2.2], [197, §7.5.9.2]). This is even true if the matrix B is singular or almost singular (this is the case of an unstable but consistent² discretisation). An extreme case of solving a problem with an unstable discretisation of high consistency order is demonstrated in [178].

Another mixing of both discretisation is described in [194, §14.3.3], where parts of the multigrid iteration for $Ax = b$ use B in the smoothing step. The limit x^* of the iterates solves neither $Ax^* = b$ nor $Bx^* = c$.

2.2 Linear Iterative Methods

One would expect iterative methods to be linear in x, b , since they solve linear equations. In fact, most of the methods described in this book are linear, but there are also important nonlinear iterations as, e.g., discussed in Part II.

2.2.1 Notation, First Normal Form

Definition 2.9 (linear iteration, iteration matrix). An iterative method Φ is called *linear* if $\Phi(x, b)$ is linear in (x, b) , i.e., if there are matrices M and N such that

$$\Phi(x, b, A) = M[A]x + N[A]b.$$

In most of the cases, A is fixed and we use the shorter form

$$\Phi(x, b) = Mx + Nb. \quad (2.7)$$

Here, the matrix $M = M[A]$ is called the *iteration matrix* of the iteration Φ .

Iteration (2.3) takes the form (2.8), which represents the *first normal form* of the iteration Φ : $x^{m+1} := Mx^m + Nb \quad (m \geq 0, b \text{ in (2.1)}).$ (2.8)

Whenever possible, we shall denote the iteration matrix of a specific iteration method ‘xyz’ by M^{xyz} ; e.g., M^{GS} belongs to the Gauss-Seidel method. Similarly for N^{xyz} . When we refer to the mapping Φ , we write M_Φ, N_Φ , etc.

Remark 2.10. Assume (2.2). If $N = N[A]$ is singular, there is some $x^* \neq 0$ with $Nx^* = 0$ and $b := Ax^* \neq 0$. Starting iteration (2.8) with $x^0 = 0$ yields $x^m = 0$ and hence $\lim x^m = 0$. In Corollary 2.17b we shall state that, in this case, the iteration is not convergent.

The iteration $\Phi(\cdot, \cdot, A)$ is algebraic in the sense of Definition 2.2b if and only if the matrices M and N are explicit functions of A .

² Concerning the terms ‘consistent’ and ‘consistency order’, we refer to Hackbusch [197, §§6,7].

2.2.2 Consistency and Second Normal Form

For a linear and consistent iteration Φ , each solution of $Ax = b$ must be a fixed point with respect to b : $x = Mx + Nb$. Each $x \in \mathbb{K}^I$ can be the solution of $Ax = b$ (namely, for $b := Ax$). Hence,

$$x = Mx + Nb = Mx + NAx$$

holds for all x and leads to the matrix equation

$$M[A] + N[A]A = I, \quad (2.9)$$

or in short,

$$M + NA = I,$$

establishing a relation between M and N in (2.8). This proves the next theorem.

Theorem 2.11 (consistency). *A linear iteration Φ is consistent if and only if the iteration matrix M can be determined from N by*

$$M[A] = I - N[A]A \quad \text{for all } A \in \mathfrak{D}(\Phi). \quad (2.9')$$

If, in addition, A is regular, N can be represented as a function of M :

$$N[A] = (I - M[A])A^{-1}. \quad (2.9'')$$

Combining formulae (2.8) and (2.9'), we can represent linear and consistent iterations in their *second normal form*:

$$x^{m+1} := x^m - N[A](Ax^m - b) \quad (m > 0, A, b \text{ in (2.1)}). \quad (2.10)$$

In the sequel, the matrix

$$N = N[A] = N_\Phi = N_\Phi[A]$$

will be called the ‘matrix of the second normal form of Φ ’. Equation (2.10) shows that x^{m+1} is obtained from x^m by a correction which is the *defect* $Ax^m - b$ of x^m multiplied by N . The fact that the defect of x^m vanishes if and only if it is a solution of $Ax = b$, proves the next remark.

Remark 2.12. The second normal form (2.10) with arbitrary $N \in \mathbb{K}^{I \times I}$ represents all linear and consistent iterations.

Since consistent linear iterations are the standard case, we introduce the following notation for the set of these iterations:

$$\mathcal{L} := \{\Phi : \mathbb{K}^I \times \mathbb{K}^I \times \mathbb{K}^{I \times I} \rightarrow \mathbb{K}^I \text{ consistent linear iteration, } \#I < \infty\}. \quad (2.11)$$

2.2.3 Third Normal Form

The *third normal form* of a linear iteration reads as follows:

$$W[A](x^m - x^{m+1}) = Ax^m - b \quad (m > 0, A, b \text{ in (2.1)}). \quad (2.12)$$

$W = W[A] = W_\Phi = W_\Phi[A]$ is called the ‘matrix of the third normal form of Φ ’. Equation (2.12) can be understood in the following algorithmic form:

$$\text{solve } W\delta = Ax^m - b \quad \text{and define} \quad x^{m+1} := x^m - \delta. \quad (2.12')$$

This represents a definition of x^{m+1} as long as W is regular. Under this assumption, one can solve for x^{m+1} . A comparison with (2.10) proves the following.

Remark 2.13. If W in (2.12) is regular, iteration (2.12) coincides with the second normal form (2.10), where N is defined by

$$N = W^{-1}. \quad (2.13)$$

Vice versa, the representation (2.10) with regular N can be rewritten as (2.12) with $W = N^{-1}$.

We shall see that for the interesting cases, N must be regular (cf. Remark 2.18). Combining (2.9') and (2.13) yields

$$M[A] = I - W[A]^{-1}A. \quad (2.13')$$

2.2.4 Representation of the Iterates x^m

By the notation $x^m(x^0, b, A)$ in (2.4) we express the dependency on the starting value x^0 and on the data b, A of the system (2.1). The explicit representation of x^m in terms of x^0 and b is given in (2.14).

Theorem 2.14. *The linear iteration (2.7) produces the iterates*

$$x^m(x^0, b, A) = M[A]^m x^0 + \sum_{k=0}^{m-1} M[A]^k N[A] b \quad (2.14)$$

for $m \geq 0$ and $A \in \mathfrak{D}(\Phi)$.

Proof. For the induction start at $m = 0$, Eq. (2.14) takes the form $x^0(x^0, b) = x^0$ in accordance with (2.4). Assuming (2.14) for $m - 1$, we obtain from (2.7) that

$$\begin{aligned} x^m(x^0, b) &= Mx^{m-1} + Nb = M \left(M^{m-1}x^0 + \sum_{k=0}^{m-2} M^k Nb \right) + Nb \\ &= M^m x^0 + \sum_{k=1}^{m-1} M^k Nb + Nb = M^m x^0 + \sum_{k=0}^{m-1} M^k Nb. \end{aligned} \quad \square$$

In the following, e^m denotes the (iteration) error of x^m :

$$e^m := x^m - x, \quad \text{where } x \text{ solves } Ax = b. \quad (2.15)$$

Assuming consistency, we have $x = Mx + Nb$ for the solution x in (2.15). Forming the difference with (2.8): $x^{m+1} = Mx^m + Nb$, we attain the simple relation

$$e^{m+1} = Me^m \quad (m \geq 0), \quad e^0 = x^0 - x, \quad (2.16a)$$

between two successive errors. Therefore the iteration matrix is the amplification matrix of the error. A trivial conclusion is

$$e^m = M^m e^0 \quad (m \geq 0). \quad (2.16b)$$

The expression $Ax - b$ is called the *defect* of a vector x . In particular,

$$d^m := Ax^m - b \quad (2.17)$$

denotes the defect of the m -th iterate x^m .

Exercise 2.15. Prove: (a) The defect $\bar{d} = A\bar{x} - b$ and the error $\bar{e} = \bar{x} - x$ fulfil the equation $A\bar{e} = \bar{d}$.

(b) Let $\Phi \in \mathcal{L}$ (cf. (2.11)) and assume that A is regular. Then the defects satisfy

$$d^{m+1} = AMA^{-1}d^m, \quad d^0 := Ax^0 - b, \quad d^m = (AMA^{-1})^m d^0.$$

2.2.5 Convergence

A necessary and sufficient convergence criterion can be formulated by the spectral radius $\rho(M)$ of the iteration matrix (cf. Definition A.17).

Theorem 2.16 (convergence theorem, convergence rate). A linear iteration (2.7) with the iteration matrix $M = M[A]$ is convergent if and only if

$$\rho(M) < 1. \quad (2.18)$$

$\rho(M)$ is called the convergence rate of the iteration $\Phi(\cdot, \cdot, A)$.

In the sequel, the terms *convergence rate*, *convergence speed*, and *iteration speed* are used synonymously for $\rho(M)$. Some authors define the convergence rate as the negative logarithm $-\log(\rho(M))$ (cf. (2.30a) and Varga [375], Young [412]).

Proof. (i) Let iteration (2.7) be convergent. In Definition 2.6 we may choose $b := 0$ and exploit the representation (2.14): $x^m = M^m x^0$. The starting value $x^0 := 0$ yields the limit $x^* = 0$, which by the convergence definition must hold for any starting value. If $\rho(M) \geq 1$, one could choose $x^0 \neq 0$ as the eigenvector corresponding to an eigenvalue λ with $|\lambda| = \rho(M) \geq 1$. The resulting sequence $x^m = \lambda^m x^0$ cannot converge to $x^* = 0$. Hence, inequality (2.18) is necessary for convergence.

(ii) Now let (2.18) be valid: $\rho(M) < 1$. By Lemma B.28, $M^m x^0$ converges to zero, while Theorem B.29 proves $\sum_{k=0}^{m-1} M^k \rightarrow (I - M)^{-1}$. Thanks to the representation (2.14), x^m tends to

$$x^* := (I - M)^{-1}Nb. \quad (2.19)$$

Since this limit does not depend on the starting value, the iteration is convergent. \square

The proof already contains the first statement of the following corollary.

Corollary 2.17. (a) If the iterative method (2.7) is convergent, the iterates converge to $(I - M)^{-1}Nb$.

(b) If the iteration is convergent, then A and $N = N[A]$ are regular.

(c) If, in addition, the iteration is consistent, the iterates x^m converge to the unique solution $x = A^{-1}b$.

Proof. (b) If either A or N are singular, the product AN is singular and $ANx = 0$ holds for some $x \neq 0$. As $M = I - NA$, x is an eigenvector of M with the eigenvalue 1. Hence $\rho(M) \geq 1$ proves the divergence of the iteration. This proves part (b).

(c) By consistency and part (b), there is a representation (2.10) with regular N and A , so that $(I - M)^{-1}N = A^{-1}$ follows from (2.9). (2.19) proves part (c). \square

Remark 2.18. Since only convergent and consistent iterations are of interest and since in this case, by Corollary 2.17b, A and N are regular, the representation (2.9'') of N and the third normal form (2.4) hold with the matrix $W = N^{-1}$.

The convergence $x^m \rightarrow x$ is an asymptotic statement for $m \rightarrow \infty$ that allows no conclusion concerning the error $e^m = x^m - x$ for some fixed m . The values of $u_{16,16}^m$ given in Tables 1.1–1.2 even deteriorate during the first steps before they converge monotonically to the limit $\frac{1}{2}$. Often, one would like to have a statement for a *fixed* iteration number m . In this case, the convergence criterion (2.18) has to be replaced with a norm estimate.

Theorem 2.19. Let $\|\cdot\|$ be a corresponding matrix norm. A sufficient condition for convergence of an iteration is the estimate

$$\|M\| < 1 \quad (2.20)$$

of the iteration matrix M . If the iteration is consistent, the error estimates (2.21) hold:

$$\|e^{m+1}\| \leq \|M\| \|e^m\|, \quad \|e^m\| \leq \|M\|^m \|e^0\|. \quad (2.21)$$

Proof. (2.20) implies (2.18) (cf. (B.20b)). (2.21) is a consequence of (2.16a,b). \square

$\|M\|$ is called the *contraction number* of the iteration (with respect to the norm $\|\cdot\|$). In the case of (2.20), the iteration is called *monotonically convergent* with respect to the norm $\|\cdot\|$, since $\|e^{m+1}\| < \|e^m\|$. If the norm $\|\cdot\|$ fulfills the equality $\rho(M) = \|M\|$, the terms ‘convergence’ and ‘monotone convergence’ coincide.

2.2.6 Convergence Speed

Inequality (2.21), i.e., $\|e^{m+1}\| \leq \zeta \|e^m\|$ with $\zeta := \|M\| < 1$, describes linear convergence. Faster convergence than linear convergence is only attainable by non-linear methods (cf. §10.2.3). The contraction number ζ depends on the choice of the norm. According to (B.20b), the contraction number ζ is always larger or equal to the convergence rate $\rho(M)$. On the other hand, Lemma B.26 ensures that for a suitable choice of the norm, the contraction number ζ approximates the convergence rate $\rho(M)$ arbitrarily well.

The contraction number as well as the convergence rate determine the quality of an iterative method. Both quantities can be determined from the errors e^m as follows.

Remark 2.20. The contraction number is the maximum of the ratios $\|e^1\|/\|e^0\|$ taken over all starting values x .

Proof. Use (2.16b) for $m = 1$ and Exercise B.10d. \square

Exercise 2.21. Prove: (a) In general, Remark 2.20 becomes wrong if $\|e^1\|/\|e^0\|$ is replaced with $\|e^{m+1}\|/\|e^m\|$ for some $m > 0$.

(b) The latter quotient takes the maximum

$$\zeta_{m+1} := \begin{cases} \max\{\|Mx\| / \|x\| : x \in \text{range}(M^m) \setminus \{0\}\} & \text{if } M^m \neq 0, \\ 0 & \text{otherwise,} \end{cases}$$

which can be interpreted as the matrix norm of the mapping $x \mapsto Mx$ restricted to the subspace $V_m := \text{range}(M^m) := \{M^m x : x \in \mathbb{K}^I\}$.

- (c) The inclusion $V_{m+1} \subset V_m$ holds with an equality sign at least for $m \geq \#I$.
- (d) $\rho(M) \leq \zeta_{m+1} \leq \zeta_m \leq \zeta_0 = \zeta := \|M\|$ holds for $m \geq 0$.
- (e) For regular M , one has $\zeta_m = \zeta$ for all m .

Exercise 2.21 demonstrates that the contraction number is a somewhat too coarse term: It may happen that the contraction number gives a too pessimistic prediction of the convergence speed. A more favourable estimate can be obtained by the numerical radius $r(\cdot)$ of the matrix M^m (cf. §B.3.4). The inequalities

$$\|M^m\|_2 \leq 2r(M^m) \quad (\text{cf. (B.28d)}) \tag{2.22a}$$

and (2.16b) yield the error estimate

$$\|e^m\|_2 \leq 2r(M^m) \|e^0\|_2 \quad (m \geq 0) \tag{2.22b}$$

with respect to the Euclidean norm. If $\|\cdot\|_C$ is the norm defined by (C.5a) with a positive definite matrix C , one analogously proves the inequality

$$\|e^m\|_C \leq 2r(C^{1/2}M^m C^{-1/2}) \|e^0\|_C \quad (m \geq 0). \tag{2.22c}$$

For the practical judgment of the convergence speed from ‘experimental data’, i.e., from a sequence of errors e^m belonging to a special starting value x^0 , one may use the *reduction factors*

$$\rho_{m+1,m} := \|e^{m+1}\|/\|e^m\|. \quad (2.23a)$$

These numbers can, e.g., be found in the last column of Tables 1.1–1.2. More interesting than a single value $\rho_{m+1,m}$ is the geometric mean

$$\rho_{m+k,m} := [\rho_{m+k,m+k-1} \cdot \rho_{m+k-1,m+k-2} \cdots \rho_{m+1,m}]^{1/k},$$

which due to definition (2.23a) can more easily be represented by

$$\rho_{m+k,m} := [\|e^{m+k}\|/\|e^m\|]^{1/k}. \quad (2.23b)$$

The properties of $\rho_{m+k,m}$ are summarised below.

Remark 2.22. (a) Denote the dependence of the magnitude $\rho_{m+k,m}$ on the starting value x^0 by $\rho_{m+k,m}(x^0)$. Then

$$\lim_{k \rightarrow \infty} \max\{\rho_{m+k,m}(x^0) : x^0 \in \mathbb{K}^I\} = \rho(M) \quad \text{for all } m.$$

(b) Even without maximisation over all $x^0 \in \mathbb{K}^I$,

$$\lim_{k \rightarrow \infty} \rho_{m+k,m}(x^0) = \rho(M) \quad \text{for all } m \quad (2.23c)$$

holds, provided that x^0 does not lie in the subspace $U \subset \mathbb{K}^I$ of dimension $< \#I$ spanned by all eigenvectors and possibly existing principal vectors of the matrix M corresponding to eigenvalues λ with $|\lambda| < \rho(M)$. (2.23c) holds almost always because a stochastically chosen starting value x^0 lying in a fixed lower dimensional subspace has probability zero.

(c) The reduction factors $\rho_{m+1,m}(x^0)$ tend to the spectral radius of M :

$$\lim_{m \rightarrow \infty} \rho_{m+1,m}(x^0) = \rho(M) \quad (2.23d)$$

for all $x^0 \notin U$ with U in part (b) if and only if there is exactly one eigenvalue $\lambda \in \sigma(M)$ with $|\lambda| = \rho(M)$, and if, for this eigenvalue, the geometric and algebraic multiplicities coincide. Sufficient conditions are: (i) $\lambda \in \sigma(M)$ with $|\lambda| = \rho(M)$ is a single eigenvalue, or (ii) M is a positive matrix (cf. (C.11a)).

(d) Choose a norm $\|\cdot\| = \|\cdot\|_C$ with $C > 0$ (cf. (2.22c)) in (2.23a). If $C^{\frac{1}{2}}MC^{-\frac{1}{2}}$ is Hermitian, $\rho_{m+1,m}(x^0)$ ($x^0 \notin U$) converges monotonically increasing to $\rho(M)$.

Proof. (i) Use

$$\rho(M) \leq \max_{x^0 \in \mathbb{K}^I} \rho_{m+k,m}(x^0) \leq \max_{x^0 \in \mathbb{K}^I} \rho_{k,0}(x^0) \leq \|M^k\|^{1/k}$$

and $\|M^k\|^{1/k} \rightarrow \rho(M)$ according to Theorem B.27. This proves part (a).

(ii) Let $I_0 \subset I$ be the nonempty index subset $I_0 := \{i \in I : |J_{ii}| = \rho(M)\}$, where J_{ii} are the diagonal elements of the Jordan normal form $M = T J T^{-1}$ (cf. (A.15a,b)). The subspace $U := \{x : (T^{-1}x)_i = 0 \text{ for all } i \in I_0\}$ is the maximal subspace with the property $\lim_{m \rightarrow \infty} [\|M^m x\| / \|x\|]^{1/m} < \rho(M)$. Its dimension is $\dim(U) = \#I - \#I_0 < \#I$.

(iii) Define $\hat{M} = C^{1/2}MC^{-1/2}$ and $\hat{e}^m := C^{1/2}e^m$. Since the norms are related by $\|e^m\|_C = \|\hat{e}^m\|_2$, we obtain for $m \geq 1$ that

$$\begin{aligned}\|\hat{e}^m\|_2^2 &= \|\hat{M}^m \hat{e}^0\|_2^2 = \langle \hat{M}^m \hat{e}^0, \hat{M}^m \hat{e}^0 \rangle = \langle \hat{M}^{m+1} \hat{e}^0, \hat{M}^{m-1} \hat{e}^0 \rangle \\ &= \langle \hat{e}^{m+1}, \hat{e}^{m-1} \rangle \leq \|\hat{e}^{m+1}\|_2 \|\hat{e}^{m-1}\|_2.\end{aligned}$$

Hence it follows that $\rho_{m+1,m} = \frac{\|e^{m+1}\|}{\|e^m\|} = \frac{\|\hat{e}^{m+1}\|_2}{\|\hat{e}^m\|_2} \geq \frac{\|\hat{e}^m\|_2}{\|\hat{e}^{m-1}\|_2} = \rho_{m,m-1}$. \square

Remark 2.22 allows us to view the value $\rho_{m+k,m}$ and possibly also $\rho_{m+1,m}$ for sufficiently large m as a good approximation of the spectral radius. This viewpoint can be reversed.

Remark 2.23. The convergence rate $\rho(M)$ is a suitable measure for judging (asymptotically) the convergence speed. This holds even if convergence is required with respect to a specific norm.

Proof. By Theorem B.27, for each $\varepsilon > 0$ there is some m_0 such that $m \geq m_0$ implies that $\rho(M) \leq \|M^m\|^{1/m} \leq \rho(M) + \varepsilon$ and $\|e^m\| \leq (\rho(M) + \varepsilon)^m \|e^0\|$. \square

2.2.7 Remarks Concerning the Matrices M , N , and W

Considerations in §§2.2.5–2.2.6 show the close connection between the iteration matrix M and the convergence speed. M directly describes the *error reduction* or amplification (cf. (2.16a)). Roughly speaking, the convergence is better the smaller M is. $M = 0$ would be optimal. However, then Φ is a direct method, since x^1 is already the exact solution (its error is $e^1 = Me^0 = 0$).

The matrix N transforms the defect $Ax^m - b$ into the correction $x^m - x^{m+1}$. The optimal case³ $M = 0$ mentioned above corresponds to $N[A] = A^{-1}$. Therefore, one may regard $N[A]$ as an *approximate inverse* of A .

Concerning implementation, often the matrix W of the third normal form (2.12) is the important one. By the relation $W = N^{-1}$ (cf. (2.13)), $W = A$ would be optimal. However, then computing the correction $x^m - x^{m+1}$ is equivalent to the direct solution of the original equation. Therefore, one has to find approximations W of A , so that the solution of the system $W\delta = d$ is sufficiently easy.

In the case of some of the classical iterations discussed in §3, we have explicit expressions for N or W and may use these matrices for the computation. On the other hand, there will be iterative methods, for which the algorithm is implemented differently without reference to the matrices M , N , W (see, e.g., Propositions 3.13 or 5.25).

³ Consistent linear iterations with $M = 0$ can be called direct solvers. Vice versa, any direct solver defines a linear iteration with $M = 0$.

2.2.8 Three-Term Recursions, Two- and Multi-Step Iterations

So far we considered *one-step iterations*, i.e., x^{m+1} is computed in one step from x^m . Sometimes linear iterations occur, in which computing x^{m+1} involves x^m and x^{m-1} :

$$x^{m+1} = M_0 x^m + M_1 x^{m-1} + N_0 b \quad (m \geq 1). \quad (2.24)$$

For the starting procedure, one needs two initial values x^0 and x^1 . Such *two-step iterations* are also called *three-term recursions* since they involved the three terms x^{m+1} , x^m , x^{m-1} . Formally, a three-term recursion can be reduced to a standard one-step iteration acting in the space $\mathbb{K}^I \times \mathbb{K}^I$:

$$\begin{bmatrix} x^{m+1} \\ x^m \end{bmatrix} = \mathbf{M} \begin{bmatrix} x^m \\ x^{m-1} \end{bmatrix} + \begin{bmatrix} N_0 b \\ 0 \end{bmatrix} \quad \text{with } \mathbf{M} := \begin{bmatrix} M_0 & M_1 \\ I & 0 \end{bmatrix}. \quad (2.25)$$

Now the convergence condition

$$\rho(\mathbf{M}) < 1 \quad (2.26a)$$

ensures that recursion (2.25) has a limit that is also the fixed point. The consistency condition takes the form

$$I - M_0 - M_1 = N_0 A. \quad (2.26b)$$

Exercise 2.24. The limit of the iteration (2.25) has the general form $\begin{bmatrix} \xi \\ \eta \end{bmatrix} \in \mathbb{K}^I \times \mathbb{K}^I$. Show that the conditions (2.26a,b) imply $\xi = \eta = A^{-1}b$.

Exercise 2.25. Given an iteration $x^{m+1} = Mx^m + Nb$, define the matrices M_0 , M_1 , N_0 in (2.24) by

$$\begin{aligned} M_0 &:= \Theta M + \vartheta I, \\ M_1 &:= (1 - \Theta - \vartheta) I, \\ N_0 &:= \Theta N \end{aligned}$$

with $\Theta, \vartheta \in \mathbb{R}$. The three-term recursion (2.24) takes the form

$$x^{m+1} = \Theta [(Mx^m + Nb) - x^{m-1}] + \vartheta(x^m - x^{m-1}) + x^{m-1}. \quad (2.27)$$

Prove that (a) \mathbf{M} has the spectrum

$$\sigma(\mathbf{M}) = \left\{ \frac{1}{2} (\Theta \lambda + \vartheta) \pm \sqrt{1 - \Theta - \vartheta + \frac{1}{4} (\Theta \lambda + \vartheta)^2} : \lambda \in \sigma(M) \right\}.$$

(b) Conclude from $\rho(M) < 1$ and $\Theta > 0$, $\vartheta \geq 0$, $\Theta + \vartheta \leq 1$ that $\rho(\mathbf{M}) < 1$.

2.3 Efficacy of Iterative Methods

The convergence rate cannot be the only criterion for the quality of an iterative method because one has also to take into account the amount of computational work of Φ .

2.3.1 Amount of Computational Work

The representation (2.12') suggests that any iteration requires at least computing the defect $Ax^m - b$. For a general $n \times n$ matrix $A \in \mathbb{K}^{I \times I}$ ($n = \#I$), multiplying Ax^m would require $2n^2$ operations. However, as discussed in §1.7, it is more realistic to assume that A is sparse; i.e., the number $s(n)$ of the nonzero elements of A is distinctly smaller than n^2 . For matrices arising from discretisations of partial differential equations, one has

$$s(n) \leq C_A n, \quad (2.28)$$

where C_A is a constant with respect to n , but depends on the matrix A . For the five-point formula (1.4a) of the model problem, inequality (2.28) holds with $C_A = 5$. Under assumption (2.28), one can perform matrix-vector multiplication in $2C_A n$ operations.

After evaluating $d := Ax^m - b$, one has still to solve the system $W\delta = d$ in (2.12'). For any practical iterative method, we should require that this part consumes only $\mathcal{O}(n)$ operations, so that the total amount of work is also of the order $\mathcal{O}(n)$. We relate the constant in $\mathcal{O}(n)$ to C_A in (2.28) and obtain the following formulation:

$$\begin{aligned} &\text{The number of arithmetic operations per iteration} \\ &\text{step of the method } \Phi \text{ is } \text{Work}(\Phi, A) \leq C_\Phi C_A n. \end{aligned} \quad (2.29)$$

Here, $\text{Work}(\Phi, A)$ is the amount of work of the Φ iteration applied to $Ax = b$. Note that C_Φ depends on the iteration Φ but not on A , whereas $C_A n$ indicates the degree of sparsity of A . Therefore, the constant C_Φ may be called the *cost factor* of the iteration Φ .

So far we only discussed the cost arising by performing one iteration step of Φ . Depending on the method, some *initialisation* may be necessary for precomputing some quantities required by Φ . Let $\text{Init}(\Phi, A)$ be the corresponding cost.

Remark 2.26. If m iteration steps are performed, the effective cost per iteration is

$$\text{Work}(\Phi, A) + \text{Init}(\Phi, A)/m.$$

In the standard case, the initialisation uses only the data of A . Therefore it pays if many systems $Ax^i = b^i$ are solved with different right-hand sides b^i but the same matrix A .

2.3.2 Efficacy

An iteration Φ can be called ‘more effective’ than Ψ if for the same amount of work Φ is faster, or if Φ has the same convergence rate, but consumes less work than Ψ . To obtain a common measure, we ask for the amount of work that is necessary to reduce the error by a fixed factor. This factor is chosen as $1/e$, since the natural logarithm is involved. According to Remark 2.23, we use the convergence rate $\rho(M)$ for the (asymptotic) description of the error reduction per iteration step. After m iteration steps, the asymptotic error reduction is $\rho(M)^m$. In order to ensure $\rho(M)^m \leq 1/e$, we have to choose $m \geq -1/\log(\rho(M))$, provided that convergence holds: $\rho(M) < 1 \Leftrightarrow \log(\rho(M)) < 0$. Therefore, we define

$$\text{It}(\Phi) := -1/\log(\rho(M)). \quad (2.30a)$$

$\text{It}(\Phi)$ represents the (asymptotic) number of the iteration steps for an error reduction by the factor of $1/e$. Note that, in general, $\text{It}(\Phi)$ is not an integer.

- Remark 2.27.** (a) Convergence of Φ is equivalent to $0 \leq \text{It}(\Phi) < \infty$. The value $\text{It}(\Phi) = 0$ corresponds to $\rho(M) = 0$, i.e., to a direct method.
 (b) Let $\Phi \in \mathcal{L}$. To reduce the iteration error (asymptotically) by a factor of $\varepsilon < 1$, we need the following number of iteration steps:

$$\text{It}(\Phi, \varepsilon) := -\text{It}(\Phi) \log(\varepsilon) \quad (2.30b)$$

- (c) If $\rho(M) = \|M\|$ or $\rho(M)$ in (2.30a) is replaced with $\|M\| < 1$, one can guarantee (not only asymptotically) that

$$\|e^{m+k}\| \leq \varepsilon \|e^m\| \quad \text{for } k \geq \text{It}(\Phi, \varepsilon). \quad (2.30c)$$

- (d) If $r(M) < 1$ holds for the numerical radius of M introduced in §B.3.4, definition (2.30b) can be replaced with $\text{It}(\Phi, \varepsilon) := \log(\varepsilon/2)/\log(r(M))$. Then, inequality (2.30c) holds with respect to the Euclidean norm.

The amount of work corresponding to the error reduction by $1/e$ is the product $\text{It}(\Phi) \text{Work}(\Phi, A) \leq \text{It}(\Phi) C_\Phi C_{An}$ (cf. (2.29)). As a characteristic quantity we choose the *effective amount of work*

$$\text{Eff}(\Phi) := \text{It}(\Phi) C_\Phi = -C_\Phi / \log(\rho(M)). \quad (2.31a)$$

$\text{Eff}(\Phi)$ measures the amount of work for an error reduction by $1/e$ in the unit ‘ C_{An} arithmetic operations’. Correspondingly, the effective amount of work for the error reduction by the factor of $1/e$ is given by

$$\text{Eff}(\Phi, \varepsilon) := -\text{It}(\Phi) C_\Phi \log(\varepsilon) = C_\Phi \log(\varepsilon) / \log(\rho(M)). \quad (2.31b)$$

Example 2.28. In the case of the model problem, the cost factor of the Gauss–Seidel iteration is $C_\Phi = 1$ (because of $C_A = 5$, cf. Remark 1.14). The numerical values in Table 1.1 suggest $\rho(M) = 0.99039$ for the grid size $h = 1/32$. Thus, the effective amount of work equals $\text{Eff}(\Phi) = 103.6$. Using $\rho(M) = 0.82$ for the SOR method and $C_\Phi = 7/5$, we deduce an effective amount of work of $\text{Eff}(\Phi) = 7.05$ for the SOR method with $h = 1/32$.

2.3.3 Order of Linear Convergence

The convergence rates $\rho(M)$ in Example 2.28 are typically close to one; i.e., the convergence is rather slow. Therefore, we may use the ansatz

$$\rho(M) = 1 - \eta \quad (\eta \text{ small}). \quad (2.32a)$$

The Taylor expansion yields $\log(1-\eta) = -\eta + \mathcal{O}(\eta^2)$ and $\frac{-1}{\log(1-\eta)} = \frac{1}{\eta(1+\mathcal{O}(\eta))} = 1/\eta + \mathcal{O}(1)$, since $1/(1-\zeta) = 1 + \zeta + \mathcal{O}(\zeta^2)$. Assuming (2.32a), we obtain the following effective amount of work:

$$\text{Eff}(\Phi) = C_\Phi/\eta + \mathcal{O}(1). \quad (2.32b)$$

For instance, the respective numbers in Example 2.28 yield $C_\Phi/\eta = 104$ for the Gauss–Seidel iteration and 7.8 for SOR.

For most of the methods we are going to discuss, assumption (2.32a) holds in the case of the model problem. More precisely, η is related to the grid size $h = 1/N = 1/(1 + \sqrt{n})$ by (2.32c) with some exponent $\tau > 0$ and a constant C_η :

$$\eta = C_\eta h^\tau + \mathcal{O}(h^{2\tau}), \quad \text{i.e., } \rho(M) = 1 - C_\eta h^\tau + \mathcal{O}(h^{2\tau}) \quad \text{with } \tau > 0 \quad (2.32c)$$

Inserting this relation into (2.32b), we obtain

$$\text{Eff}(\Phi) = C_{\text{eff}} h^{-\tau} + \mathcal{O}(1) \quad \text{with } C_{\text{eff}} := C_\Phi/C_\eta. \quad (2.32d)$$

Remark 2.29. (a) The exponent τ in (2.32c) is called the *order of convergence rate*. If an iteration Φ has a higher order than an iteration Ψ , Φ is more expensive than Ψ for sufficiently small step size h . The smaller the order, the better the method.

(b) If Φ_1 and Φ_2 have the same order but different constants $C_{\text{eff},1} < C_{\text{eff},2}$, then Φ_2 is more expensive by a factor of $C_{\text{eff},2}/C_{\text{eff},1}$.

2.4 Test of Iterative Methods

In later chapters numerous iterative methods will be defined. For the judgement and presentation of numerical results, one may ask how iterations should be tested.

2.4.1 Consistency Test

Because of a bug in the implementation, it may happen that an iterative method is nicely converging, but to a wrong solution. The reason is a violation of consistency. For that reason, one should choose some nontrivial vector $x \in \mathbb{K}^I$ (e.g., defined by random) and compute $b := Ax$. In that case, the solution x of $Ax = b$ is known and one can observe the errors $e^m = x^m - x$.

2.4.2 Convergence Test

The quality of an iteration is (at least asymptotically) determined by the effective amount of work $\text{Eff}(\Phi)$. The amount of computational work per iteration is obtained by counting the operations.⁴ It remains to determine the convergence speed experimentally. The following trivial remark emphasises the fact that one need not test the method with different right-hand sides b (and thereby with different solutions x).

Remark 2.30. A linear iteration applied to the two systems $Ax = b$ and $Ax' = b'$ results in the same errors $x^m - x$ and $x'^m - x'$ if the starting values x^0 and x'^0 are related by $x^0 - x = x'^0 - x'$.

Conclusion 2.31. *Without loss of generality, one may always choose $x = b = 0$, together with an arbitrary starting value $x^0 \neq 0$.*

According to Remark 2.30, the test of an iteration can be based on the errors $e^m = x^m - x$ and the ratio of their norms,

$$\rho_{m+1,m} := \|e^{m+1}\| / \|e^m\| \quad (\text{cf. (2.23a)}),$$

for one or more starting vectors e^0 .

Different starting values yield different errors. However, since the geometrical mean $\rho_{m+k,m} = (\|e^{m+k}\| / \|e^m\|)^{1/k}$ (cf. (2.23b)) converges to $\rho(M)$ for $k \rightarrow \infty$, the ratios can show remarkable deviations only during the first iteration steps. However, note the following remark.

Remark 2.32. In the exceptional case that the starting error $e^0 = x^0 - x$ lies in the subspace U defined in Remark 2.22b, the numbers $\rho_{m+k,m}$ approximate a value smaller than $\rho(M)$.

In practice, meeting this exceptional case is unlikely, in particular, when the solution x is unknown. Furthermore, the usual floating-point errors prevent the iterate x^m from staying in the described subspace.

Computing $\rho_{m+1,m} = \|e^{m+1}\| / \|e^m\|$ requires the knowledge of the exact solution. If we choose $b = 0$ and $x = 0$ according to Conclusion 2.31, $\rho_{m+1,m} = \|x^{m+1}\| / \|x^m\|$ holds. If one wishes to estimate the convergence rate during the iterative computation of an unknown solution x , one may use

$$\hat{\rho}_{m+1,m} = \|x^{m+1} - x^m\| / \|x^m - x^{m-1}\|$$

and $\hat{\rho}_{m+k,m} := (\hat{\rho}_{m+k,m+k-1} \cdot \dots \cdot \hat{\rho}_{m+1,m})^{1/k}$ instead of $\rho_{m+k,m}$.

Exercise 2.33. Prove: In spite of $1 \in \sigma(M)$, $\hat{\rho}_{m+k,m} \rightarrow \rho < 1 \leq \rho(M)$ may happen for $k \rightarrow \infty$. If $1 \notin \sigma(M)$, $\hat{\rho}_{m+k,m} \rightarrow \rho(M)$ is valid for all starting errors $e^0 \notin U$ with U defined in Remark 2.22b.

⁴ Alternatively, the number of iterations may be replaced with the CPU time.

2.4.3 Test by the Model Problem

Deviating from the proposal $x = b = 0$ but according to the choice in §1.6, we define the solution x of the Poisson model problem as the grid function with the components

$$u_{ij} = (ih)^2 + (jh)^2 \quad (1 \leq i, j \leq N - 1) \quad (2.33a)$$

corresponding to the right-hand side (2.33b) (cf. Remark 1.15):

$$b \text{ defined by (1.6a) with } f = -4. \quad (2.33b)$$

We recall that u and x are different representations of the same quantity (1.6b). The vector b coincides with f in grid points not neighboured to the boundary; otherwise boundary data are added in (1.6a).

2.4.4 Stopping Criterion

A comment has to be added concerning the desirable size of the (unavoidable) iteration error $\|e^m\|$. For an unlimited iterative process, the rounding errors prevent the iteration error from converging to zero. Instead, the error will oscillate around $\text{const} \cdot \|x\| \cdot \text{eps}$ (eps : relative machine precision). For testing an iteration, one may approach this lower limit; in practice, however, there is almost never a reason for such high accuracy.

Remark 2.34. The (exact) solution x of the Poisson model problem in §1.2 is only approximating the true solution of the boundary with a discretisation error, which in this case has the order $\mathcal{O}(h^2)$ (cf. Hackbusch [193, §4.5]). Therefore, an additional iteration error of the same order $\mathcal{O}(h^2)$ is acceptable.

The algorithm in §11.5 will automatically yield an approximation for which the discretisation and iteration errors are similar in size.

A more accurate approximation x^m is needed if, e.g., x^m is the starting point of an error estimation (cf. Verfürth [379]) or for the extrapolation to the limit $h \rightarrow 0$ ('Richardson extrapolation', cf. Richardson–Gaunt [325], [194, §14.1.1]).

Often, the stopping criterion is based on the defect $Ax^m - b$ (or the residual $b - Ax^m$). Here caution must be exercised: $\|b - Ax^m\|_2 \leq 10^{-16}$ might hold, in spite of $\|e^m\|_2 \approx 1$.

Remark 2.35. In general, the sizes of $\|b - Ax^m\|_2$ and $\|e^m\|_2$ are not comparable. Their ratio depends not only on the condition $\text{cond}_2(A)$ (cf. §5.6.5.2 and Proposition B.14) but also on the scaling of the vectors x and b .

Chapter 3

Classical Linear Iterations in the Positive Definite Case

Abstract The Jacobi and Gauss–Seidel iterations and the SOR method are closely connected, and therefore they will be analysed simultaneously. The analysis, however, is essentially different for the case of positive definite matrices A discussed below and other cases studied in Chapter 4. The introductory Section 3.1 underlines the fact that the positive definite case is of practical interest. The Poisson model matrix is an example of a positive definite matrix. Section 3.2 describes the iterations of Richardson, Jacobi, Gauss–Seidel, and the SOR iteration. Block versions of these iterations are discussed in Section 3.3. The required computational work is described in 3.4. Qualitative and quantitative convergence results are given in Section 3.5. The convergence analysis of the Richardson iteration in §3.5.1 leads to convergence criteria for general positive definite iterations (cf. §3.5.2). The Gauss–Seidel and SOR iteration is analysed in §3.5.4. In particular the improvement of the order of convergence by SOR is investigated. The convergence statements for the Poisson model case are illustrated in Section 3.6 by numerical examples.

3.1 Eigenvalue Analysis of the Model Problem

The eigenvalues of the matrix A in §1.2 (e.g., in the representation (1.8)) can be described explicitly.

Theorem 3.1. *The $n \times n$ matrix A of the Poisson model problem in §1.2 with $n = (N - 1)^2$ has the following n eigenvalues:*

$$\lambda_{\alpha\beta} = 4h^{-2} [\sin^2(\alpha\pi h/2) + \sin^2(\beta\pi h/2)] \quad (1 \leq \alpha, \beta \leq N - 1), \quad (3.1a)$$

not all being different. The multiplicity of $\lambda = \lambda_{\alpha\beta}$ is given by the number of pairs $(\alpha', \beta') \in [1, N - 1]^2$ with coinciding values $\lambda_{\alpha\beta} = \lambda_{\alpha'\beta'}$. The minimal eigenvalue is attained for $\alpha = \beta = 1$, the maximal for $\alpha = \beta = N - 1$:

$$\lambda_{\min} = \|A^{-1}\|_2^{-1} = 8h^{-2} \sin^2(\pi h/2), \quad (3.1b)$$

$$\lambda_{\max} = \|A\|_2 = 8h^{-2} \cos^2(\pi h/2). \quad (3.1c)$$

In particular, A is a positive definite matrix.

Proof. (1.8) shows $A = A^T = A^H$. The positive definiteness is a consequence of (3.1a) and Lemma C.3. Equations (3.1b,c) follows from (3.1a). $\lambda_{\alpha\beta}$ in (3.1a) depends monotonically on $\alpha, \beta \in \{1, \dots, N-1\}$; hence,

$$\lambda_{\max} = \lambda_{N-1, N-1} = \rho(A) = \|A\|_2, \quad \lambda_{\min} = \lambda_{1,1} = 1/\rho(A^{-1}) = 1/\|A^{-1}\|_2.$$

The eigenvalues (3.1a) follow from Lemma 3.2. \square

Lemma 3.2. *The n linearly independent and even orthonormal eigenvectors of the matrix A in §1.2 corresponding to the eigenvalues $\lambda_{\alpha\beta}$ in (3.1a) are the vectors $e^{\alpha\beta}$ with the components*

$$(e^{\alpha\beta})_{\nu\mu} = 2h \sin(\alpha h \nu \pi) \sin(\beta h \mu \pi) \quad (1 \leq \alpha, \beta \leq N-1). \quad (3.2)$$

Proof. $e^{\alpha\beta}$ can be viewed as the tensor product $e^\alpha \otimes e^\beta$ of the vectors

$$e^k \in \mathbb{R}^{N-1}, \quad (e^k)_\nu = \sqrt{2h} \sin(kh\nu\pi) \quad (1 \leq k, \nu \leq N-1),$$

since $(e^{\alpha\beta})_{\nu\mu} = (e^\alpha)_\nu (e^\beta)_\mu$. The scalar product in \mathbb{R}^n is

$$\begin{aligned} \langle e^{\alpha\beta}, e^{k\ell} \rangle &= \sum_{\nu, \mu} (e^{\alpha\beta})_{\nu\mu} (e^{k\ell})_{\nu\mu} = \sum_{\nu, \mu} e_\nu^\alpha e_\mu^\beta e_\nu^k e_\mu^\ell = \sum_\nu e_\nu^\alpha e_\nu^k \sum_\mu e_\mu^\beta e_\mu^\ell \\ &= \langle e^\alpha, e^k \rangle \langle e^\beta, e^\ell \rangle, \end{aligned}$$

where the last two scalar products are those in \mathbb{R}^{N-1} . This identity shows that $\{e^{\alpha\beta} : 1 \leq \alpha, \beta \leq N-1\}$ is an orthonormal basis, provided that $\{e^\alpha : 1 \leq \alpha \leq N-1\}$ is an orthonormal basis of \mathbb{R}^{N-1} . The latter statement is the subject of Exercise 3.3.

Exercise 3.3. Assume $hN = 1$ and $1 \leq k, \ell \leq N-1$. Prove that

$$\sum_{\nu=1}^{N-1} \sin(kh\nu\pi) \sin(\ell h\nu\pi) = \begin{cases} 1/(2h) & \text{for } k = \ell, \\ 0 & \text{otherwise.} \end{cases}$$

For the model matrix A , the values of the grid function $Ae^{\alpha\beta}$ at the inner grid points $(x, y) = (kh, \ell h) \in \Omega_h$ are

$$(Ae^{\alpha\beta})(x, y) = h^{-2} 2h [4 \sin(\alpha x \pi) \sin(\beta y \pi) \quad (3.3)$$

$$\begin{aligned} &- \sin(\alpha(x+h)\pi) \sin(\beta y \pi) - \sin(\alpha(x-h)\pi) \sin(\beta y \pi) \\ &- \sin(\alpha x \pi) \sin(\beta(y+h)\pi) - \sin(\alpha x \pi) \sin(\beta(y-h)\pi)]. \end{aligned}$$

The sine addition theorem yields

$$\begin{aligned}\sin(\alpha(x+h)\pi) + \sin(\alpha(x-h)\pi) &= 2\sin(\alpha x\pi)\cos(\alpha h\pi), \\ \sin(\beta(y+h)\pi) + \sin(\beta(y-h)\pi) &= 2\sin(\beta y\pi)\cos(\beta h\pi)\end{aligned}$$

and therefore $(Ae^{\alpha\beta})(x, y) = \frac{2}{h} \sin(\alpha x\pi) \sin(\beta y\pi) [4 - 2\cos(\alpha h\pi) - 2\cos(\beta h\pi)]$. From the identity

$$1 - \cos \xi = 2 \sin^2(\xi/2),$$

we conclude the assertion (3.1a): $Ae^{\alpha\beta} = \lambda_{\alpha\beta} e^{\alpha\beta}$. Equation (3.3) requires additional consideration. If all neighbours $(x-h, y)$ and $(x, y-h)$ of the grid point (x, y) belong again to Ω_h , Eq. (3.3) represents directly the component of the vector $Ae^{\alpha\beta}$ corresponding to the point (x, y) . If, however, one neighbour, say $Q = (x-h, y)$, is not an inner grid point because of $x = h$, the term $-h^{-2}e^{\alpha\beta}(Q)$ should not appear. In this case, $e^{\alpha\beta}(Q) = 0$ holds, and therefore, Eq. (3.3) is still valid. \square

The following exercise discusses two generalisations of the model problem. The grid in $\Omega = (0, L_x) \times (0, L_y)$ may use different step sizes h_x and h_y with respect to the x and y directions ('anisotropic discretisation'). The corresponding numbers of subintervals in each direction are N_x and N_y .

Exercise 3.4. Discretise the model problem in $\Omega = (0, L_x) \times (0, L_y)$ with the step sizes $h_x := L_x/N_x$ and $h_y := L_y/N_y$. Prove that the discretisation matrix A has the eigenvalues

$$\lambda_{\alpha\beta} = 4 \left(h_x^{-2} \sin^2 \frac{\alpha h_x \pi}{2} + h_y^{-2} \sin^2 \frac{\beta h_y \pi}{2} \right) \quad \text{for } \begin{cases} 1 \leq \alpha \leq N_x - 1, \\ 1 \leq \beta \leq N_y - 1. \end{cases}$$

The eigenvectors $e^{\alpha\beta}$ are the tensor products $e_x^\alpha \otimes e_y^\beta$ with

$$\begin{aligned}(e_x^\alpha)_\nu &= \sqrt{2h_x} \sin(\alpha h_x \nu \pi) \quad (1 \leq \alpha, \nu \leq N_x - 1) \quad \text{and} \\ (e_y^\beta)_\mu &= \sqrt{2h_y} \sin(\beta h_y \mu \pi) \quad (1 \leq \beta, \mu \leq N_y - 1).\end{aligned}$$

3.2 Traditional Linear Iterations

3.2.1 Richardson Iteration

The simplest choice of the matrix N of the second normal form (2.10) is the identity $N = I$ or a multiple of I . The resulting scheme reads as follows:

$$x^{m+1} = x^m - \Theta(Ax^m - b) \quad (\Theta \in \mathbb{C}). \quad (3.4)$$

This iteration method is called the *Richardson iteration*. It is denoted by \varPhi^{Rich} . In the original paper of Richardson [324, §3.2], the author describes a variant of (3.4) with varying constants which is called the semi-iterative or instationary Richardson iteration (see §8).

Proposition 3.5. (a) $\Phi_\Theta^{\text{Rich}} \in \mathcal{L}$ is algebraic (cf. Definition 2.2b), and the domain $\mathfrak{D}(\Phi_\Theta^{\text{Rich}})$ is the set of all matrices without any exception.
 (b) The iteration matrix of the Richardson iteration is

$$M_\Theta^{\text{Rich}} := I - \Theta A.$$

The matrices of the second and third normal forms are

$$N_\Theta^{\text{Rich}} := \Theta I, \quad W_\Theta^{\text{Rich}} := \frac{1}{\Theta} I.$$

(c) The Richardson method is independent of the ordering of indices. This fact is helpful for parallel implementations.

Although Richardson's iteration seems to be the simplest possible, it will turn out in Proposition 5.44 that the Richardson iteration with $\Theta = 1$ is the prototype of any linear iteration $\Phi \in \mathcal{L}$.

To apply iteration (3.4), we have to choose the parameter Θ . In Theorem 3.23 we shall discuss for which values we obtain convergence and what the best choice is. The following remarks holds for all methods involving one or more parameters.

Remark 3.6. Methods requiring the user to choose a suitable parameter may cause a practical problem. Even if it is known what the optimal parameter is, this value may depend on data (e.g., spectral data of the matrix) which are not known (or their computation is more expensive than the original problem). The difficulty is increased in the presence of two or more parameters to be tuned.

3.2.2 Jacobi Iteration

The iteration described by C. G. Jacobi in 1845 (also called ‘total-step process’, ‘Gesamtschrittverfahren’) results from (7.4) by the choice $W := D := \text{diag}\{A\}$:

$$x^{m+1} = \Phi^{\text{Jac}}(x^m, b) = x^m - D^{-1}(Ax^m - b). \quad (3.5)$$

The Jacobi iteration and the iterations discussed below are defined for matrices in

$$\mathfrak{D}(\Phi^{\text{Jac}}) = \mathfrak{D}(\Phi^{\text{GS}}) = \mathfrak{D}(\Phi_\omega^{\text{SOR}}) = \{A \in \mathbb{K}^{I \times I} : a_{\alpha\alpha} \neq 0 \text{ for } \alpha \in I\}. \quad (3.6)$$

Proposition 3.7. The matrices associated with $\Phi^{\text{Jac}} \in \mathcal{L}$ are

$$x^{m+1} = M^{\text{Jac}} x^m + N^{\text{Jac}} b \quad \text{with} \quad (3.7a)$$

$$M^{\text{Jac}} = D^{-1}(D - A) = I - D^{-1}A, \quad (3.7b)$$

$$N^{\text{Jac}} = D^{-1}, \quad W^{\text{Jac}} = D. \quad (3.7c)$$

The Jacobi iteration is algebraic and does not depend on the ordering of the indices.

In many cases, e.g., in the case of the Poisson model problem, the diagonal entries are constant, i.e., $D = c \cdot I$. Then the following remark applies.

Remark 3.8. If $D = cI$ is a multiple of the identity, the Jacobi iteration coincides with the Richardson iteration for $\Theta := 1/c$.

In the following algorithmic description of Φ^{Jac} , it is important that x^m and x^{m+1} (named below x and y) are stored separately.¹

```

function  $\Phi^{\text{Jac}}(x, b: \text{vector}): \text{vector};$  {Jacobi iteration}
  var  $y: \text{vector};$ 
  begin for all  $\alpha \in I$  do  $y[\alpha] := \left[ b[\alpha] - \sum_{\beta \in I \setminus \{\alpha\}} a[\alpha, \beta] x[\beta] \right] / a[\alpha, \alpha];$ 
     $\Phi^{\text{Jac}} := y$ 
  end;

```

(3.8)

Exercise 3.9. Let $\Delta \in \mathbb{K}^{I \times I}$ be any regular diagonal matrix. (a) Scaling the system $Ax = b$ by Δ yields $\Delta Ax = \Delta b$, i.e., $A'x = b'$ for $A' = \Delta A$ and $b' = \Delta b$. Prove that the Jacobi iteration applied to A' is identical with the Jacobi iteration applied to A . Hence, the Jacobi iteration is *invariant with respect to scaling*.

(b) A scaling by $A' = \Delta^{\frac{1}{2}}A\Delta^{\frac{1}{2}}$ preserves positive definiteness. The scaled system takes the form $A'x' = b'$ with $x' = \Delta^{-\frac{1}{2}}x$. Prove that the Jacobi iterations applied to $Ax = b$ and $A'x' = b'$ with starting values x^0 and $x'^0 = \Delta^{-\frac{1}{2}}x^0$ produce iterates related by $x'^m = \Delta^{-\frac{1}{2}}x^m$; i.e., they are identical up to the scaling.

3.2.3 Gauss–Seidel Iteration

The Gauss–Seidel iteration is mentioned by Gauss [147] in 1826. In the letter [148] (translated in [137]), Gauss uses an adaptive version of the Gauss–Seidel iteration together with a ‘symmetric trick’. The second name originates from the contribution of Seidel [337]. There are several other names for this method: ‘Liebmann method’ (cf. Liebmann [263]), the method of successive displacement (see ‘succesive Annäherung’ in the title of Seidel’s paper [337]), relaxation (see Footnote 5 on page 4), or ‘single-step process’ (Einzelschrittverfahren).

Let the index set I be ordered and identified with $\{1, \dots, n\}$. The algorithmic realisation looks very similar to the Jacobi iteration in (3.8) with the difference that the vector x is immediately overwritten by the new values:

```

function  $\Phi^{\text{GS}}(x, b: \text{vector}): \text{vector};$  {Gauss–Seidel iteration}
  begin for  $i := 1$  to  $n$  do  $x[i] := \left[ b[i] - \sum_{j \in I \setminus \{i\}} a[i, j] x[j] \right] / a[i, i];$ 
     $\Phi^{\text{GS}} := x$ 
  end;

```

(3.9)

¹ The function uses the representation of x^{m+1} by $D^{-1}[b - (A - D)x^m]$. The declaration ‘**var**’ denotes the declaration of the variable type. ‘**vector**’ indicates the type corresponding to \mathbb{K}^I .

Distinguishing more precisely the iterates x^m and x^{m+1} , we obtain

$$\textbf{for } i := 1 \text{ to } n \text{ do } x_i^{m+1} := \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{m+1} - \sum_{j=i+1}^n a_{ij} x_j^m \right) / a_{ii}; \quad (3.10)$$

This can be rewritten as

$$x_i^{m+1} = [b_i - (Ex^m)_i - (Fx^{m+1})_i] / a_{ii} = (D^{-1}[Ex^m - Fx^{m+1}])_i,$$

where D, E, F are the matrices obtained from the splitting (1.16) of A into

$$A = D - E - F \quad \text{with} \quad (3.11a)$$

$$D : \quad \text{diagonal matrix } (D := \text{diag}\{A\} \text{ is the diagonal part of } A), \quad (3.11b)$$

$$E : \quad \text{strictly lower triangular matrix}, \quad (3.11c)$$

$$F : \quad \text{strictly upper triangular matrix}. \quad (3.11d)$$

The previous componentwise equation proves the following representation (3.12).

Proposition 3.10. (a) $\Phi^{\text{GS}} \in \mathcal{L}$ is algebraic² with $\mathfrak{D}(\Phi^{\text{GS}})$ defined in (3.6) and the additional requirement that I is ordered.

(b) The matrices of the normal forms of the Gauss–Seidel iteration $\Phi^{\text{GS}}(x, b) = M^{\text{GS}}x + N^{\text{GS}}b$ are

$$M^{\text{GS}} = (D - E)^{-1}F, \quad N^{\text{GS}} = (D - E)^{-1}, \quad W^{\text{GS}} = D - E. \quad (3.12)$$

(c) Different orderings of I yield different iterations Φ^{GS} .

Because of part (c), a precise description of the Gauss–Seidel method must characterise the ordering if this is not already fixed by the problem. In the model case (1.4a), one has, e.g., the lexicographical Gauss–Seidel method $\Phi_{\text{lex}}^{\text{GS}}$ and the chequer-board Gauss–Seidel method $\Phi_{\text{cb}}^{\text{GS}}$, referring to the corresponding orderings of the grid points.

The componentwise representation (3.10) shows that the algorithm works sequentially. Although (3.9) seems to be simpler than (3.8), the parallel computation is hampered by the sequential loop (for the chequer-board variant of the model problem the parallel treatment of both colours is possible; cf. Niethammer [293]).

Remark 3.11. The chequer-board numbering can be generalised to the case of the nine-point formula defined in (1.13a) by introducing the *four-colour ordering*:

$$\begin{aligned} \Omega_{h1} &:= \{(x, y) = (ih, jh) \in \Omega_h : i, j \text{ even}\}, \\ \Omega_{h2} &:= \{(x, y) = (ih, jh) \in \Omega_h : i, j \text{ odd}\}, \\ \Omega_{h3} &:= \{(x, y) = (ih, jh) \in \Omega_h : i \text{ even}, j \text{ odd}\}, \\ \Omega_{h4} &:= \{(x, y) = (ih, jh) \in \Omega_h : i \text{ odd}, j \text{ even}\}. \end{aligned} \quad (3.13)$$

² Here we assume that the matrix data A are given in some ordering defining the ordering of I .

First, the points of Ω_{h1} are numbered, then those of Ω_{h2} , Ω_{h3} , and finally Ω_{h4} . Since $\Omega_{h1} \cup \Omega_{h2}$ represents the black squares and $\Omega_{h3} \cup \Omega_{h4}$ the red ones and since for the five-point formula the ordering inside of one colour is irrelevant, the four-colour numbering defined above coincides with the chequer-board ordering in the case of a five-point formula.

Exercise 3.12. Prove that the statements of Exercise 3.9 are also valid for the Gauss–Seidel iteration.

3.2.4 SOR Iteration

The SOR method (successive overrelaxation; cf. Young [411]) has already been defined in (1.22) by

$$\text{for } i := 1 \text{ to } n \text{ do } x_i^{m+1} := x_i^m - \omega \left[\sum_{j=1}^{i-1} a_{ij} x_j^{m+1} + \sum_{j=i}^n a_{ij} x_j^m - b_i \right] / a_{ii}. \quad (3.14)$$

The representation is very similar to (3.10). The essential difference is the introduced relaxation parameter ω . As in (3.9), the vector x can be overwritten by the newly computed values:

```

function  $\Phi_\omega^{\text{SOR}}(x, b);$ 
begin for  $i := 1$  to  $n$  do  $x[i] := x[i] - \frac{\omega}{a[i, i]} \left[ \sum_{j \in I} a[i, j] x[j] - b[i] \right];$ 
     $\Phi_\omega^{\text{SOR}} := x$ 
end;

```

Proposition 3.13. (a) For all ω , the iteration $\Phi_\omega^{\text{SOR}} \in \mathcal{L}$ is algebraic and the statement of Proposition 3.10a also applies to $\mathfrak{D}(\Phi_\omega^{\text{SOR}})$.

(b) Let $A = D - E - F$ be decomposed according to (3.11a–d). The matrices associated with the first and second normal forms of $\Phi_\omega^{\text{SOR}}(x, b)$ are

$$x^{m+1} = M_\omega^{\text{SOR}} x^m + N_\omega^{\text{SOR}} b, \quad (3.15a)$$

$$M_\omega^{\text{SOR}} = (I - \omega L)^{-1} \{(1 - \omega)I + \omega U\} = (D - \omega E)^{-1} \{(1 - \omega)D + \omega F\}, \quad (3.15b)$$

$$N_\omega^{\text{SOR}} = \omega(I - \omega L)^{-1} D^{-1} = \omega(D - \omega E)^{-1}, \quad \text{where} \quad (3.15c)$$

$$L := D^{-1}E, \quad U := D^{-1}F. \quad (3.15d)$$

The matrix of the third normal form is

$$W_\omega^{\text{SOR}} = \omega^{-1}(D - \omega E) = \omega^{-1}D - E. \quad (3.15e)$$

(c) For $\omega = 1$, SOR coincides with the Gauss–Seidel method: $\Phi_1^{\text{SOR}} = \Phi^{\text{GS}}$. For $0 < \omega < 1$, the precise name of Φ_ω^{SOR} is ‘underrelaxation method’, whereas the term ‘overrelaxation method’ suits for $\omega > 1$.

Proof. (i) Part (a) is easily seen from the representation (3.14).

(ii) Multiplication by $I - \omega L$ brings (3.15a) into the form

$$(I - \omega L)x^{m+1} = \{(1 - \omega)I + \omega U\}x^m + \omega D^{-1}b \quad \text{and} \quad (3.15f)$$

$$x^{m+1} = x^m - \omega [-Lx^{m+1} + (I - U)x^m - D^{-1}b].$$

We obtain the expression $[-Ex^{m+1} + (D - F)x^m - b]$ according to definition (3.15d) of L and U by moving D^{-1} in front of the bracket. The componentwise interpretation of this equation coincides with (3.14). \square

All comments about the Gauss–Seidel iteration in the lines after Proposition 3.10 also apply to the SOR method. In addition, Remark 3.6 applies to the relaxation parameter ω .

3.3 Block Versions

3.3.1 Block Structure

The Poisson model problem shows that the systems may have a natural block structure $\{I_i : i \in B\}$ (cf. §A.4). The representation

$$A = \begin{bmatrix} A^{11} & & & \\ & A^{22} & & \\ & & \ddots & \\ & & & A^{\beta\beta} \end{bmatrix}, \quad A^{ii} \in \mathbb{K}^{I_i \times I_i},$$

corresponds to the case of an ordered set $B = \{1, \dots, \beta\}$. The index subsets I_i need not be ordered. Vectors $x \in \mathbb{K}^I$ are similarly substructured by the vector blocks x^i ($i \in B$).

In the following, D does not denote the diagonal but the *block diagonal* of A (with respect to the block structure B):

$$D := \text{blockdiag}_B\{A\} := \text{blockdiag}\{A^{\kappa\kappa} : \kappa \in B\}. \quad (3.16)$$

Remark 3.14. The Poisson model problem in §1.2 offers different possibilities in defining the blocks.

(a) The columns of the grid ($x = ih$ constant) correspond to the blocks

$$u^i := (u_{i,1}, u_{i,2}, \dots, u_{i,N-1})^\top \quad (1 \leq i \leq N-1).$$

The block structure is described by the index subsets $I_i = \{(i, j) : 1 \leq j \leq N-1\}$ for all $i \in B := \{1, \dots, N-1\}$ (cf. §A.4).

(b) The rows of the grid points defined by $y = jh$ lead to the block structure $I_j = \{(i, j) : 1 \leq i \leq N-1\}$ for all $j \in B := \{1, \dots, N-1\}$ (cf. Example A.21).

In the case of the Poisson model problem, according to (1.8), the matrix blocks are

$$A^{ii} = D^{ii} = h^{-2} \begin{bmatrix} 4 & -1 & & \\ -1 & 4 & -1 & \\ & \ddots & \ddots & \ddots \\ & & -1 & 4 \end{bmatrix}, \quad \begin{aligned} A^{i,i\pm 1} &= -h^{-2}I, \\ A^{i,j} &= 0 \text{ otherwise.} \end{aligned} \quad (3.17)$$

Here we use the notation D^{ii} for the diagonal block.³ Concerning the diagonal blocks, we recall the result of Lemma C.4c.

Lemma 3.15. *If A is positive definite, all diagonal blocks A^{ii} are also positive definite.*

Independent of the choice (a) or (b) in Remark 3.14, B can be considered as a set with or without ordering. If the blocks should be ordered, there are again several possibilities.

Example 3.16. The elements of B above can be ordered as follows.

- (a) The *lexicographical ordering* is given by $1, 2, \dots, N-1$.
- (b) The *backward lexicographical ordering* is $N-1, N-2, \dots, 1$.
- (c) The *zebra ordering* $1, 3, 5, \dots, 2, 4, 6, \dots$ is an analogue to the chequer-board ordering.

3.3.2 Block-Jacobi Iteration

The block-Jacobi iteration $\Phi_{\text{block}}^{\text{Jac}}$ is the iteration (3.5) with the diagonal replaced with D in (3.16):

$$\text{for all } \kappa \in B \text{ do } y^\kappa := b^\kappa - (A^{\kappa\kappa})^{-1} \sum_{\lambda \in B \setminus \{\kappa\}} A^{\kappa\lambda} x^\lambda; \quad (3.18)$$

The block structures defined in Remark 3.14a,b yield the *column-block-Jacobi iteration* and the *row-block-Jacobi iteration*, respectively.

Remark 3.17. (a) Iteration $\Phi_{\text{block}}^{\text{Jac}} \in \mathcal{L}$ is algebraic, provided that the matrix data are organised blockwise. It is well-defined if and only if all diagonal blocks are regular: $\mathcal{D}(\Phi_{\text{block}}^{\text{Jac}}) = \{A \in \mathbb{K}^{I \times I} : A^{\kappa\kappa} \text{ regular for } \kappa \in B\}$. Here the statement of Lemma 3.15 is helpful.

- (b) The representations (3.7b,c) remain valid if D is defined by (3.16).

³ The term *diagonal block* means a block in diagonal position of a block-structured matrix.

(c) Each step of the loop in (3.18) requires solving a smaller system of the form $A^{\kappa\kappa}\delta^\kappa = c^\kappa$ ($\kappa \in B$). The exact solution of these systems should not be too expensive. For instance, $A^{\kappa\kappa}$ may be a tridiagonal or at least a band matrix with a small band width.

(d) The block-Jacobi iteration neither depends on the ordering of the blocks nor on the ordering of the indices inside of the blocks.

The requirement in Remark 3.17 is satisfied if $A^{\kappa\kappa}$ is tridiagonal (cf. (3.17)). Then one should determine the LU decomposition $A^{\kappa\kappa} = L^{\kappa\kappa} U^{\kappa\kappa}$. Solving $L U \delta = c$ costs $5\#\kappa$ arithmetic operations.

Exercise 3.18. Show that Exercise 3.9 remains valid when the regular diagonal matrix Δ is replaced by a regular block-diagonal matrix.

3.3.3 Block-Gauss–Seidel Iteration

To obtain the block-Gauss–Seidel method, only the definitions (3.11b–d) have to be changed:

$$A = D - E - F, \quad (3.19a)$$

$$D : \text{block-diagonal matrix } \text{blockdiag}\{A\}, \quad (3.19b)$$

$$E : \text{strictly lower block-triangular matrix}, \quad (3.19c)$$

$$F : \text{strictly upper block-triangular matrix}. \quad (3.19d)$$

Using these matrices D, E, F in (3.12), we obtain the normal forms of the block-Gauss–Seidel method.

The loop in (3.9) becomes

$$\textbf{for } i := 1 \textbf{ to } n \textbf{ do } x^i := (D^{ii})^{-1} \left(x^i - \sum_{j \in B \setminus \{i\}} A^{i,j} x^j \right);$$

Remark 3.19. (a) The block-Gauss–Seidel method is consistent: $\Phi_{\text{block}}^{\text{GS}} \in \mathcal{L}$. It is well-defined under the same assumptions as for $\Phi_{\text{block}}^{\text{Jac}} \in \mathcal{L}$ (cf. Remark 3.17).

(b) The block-Gauss–Seidel method depends on the ordering of the blocks, but not on the ordering of the indices inside of the blocks.

To distinguish the standard Gauss–Seidel method from the block version, we shall use the term *pointwise Gauss–Seidel method* for the iteration in §3.2.3.

Since the block-Gauss–Seidel method depends on the ordering of the blocks, an iteration called, e.g., the *lexicographical row-block-Gauss–Seidel* combines the choice of the block structure in Remark 3.14b with the ordering of Example 3.16a.

3.3.4 Block-SOR Iteration

Define the matrices in $A = D - E - F$ by (3.19b–d) and determine L and U by (3.15d): $L = D^{-1}E$, $U = D^{-1}F$. Then the matrices in (3.15a–c) define the block-SOR method. The blockwise description reads as

$$\text{for } i := 1 \text{ to } \beta \text{ do } x^i := x^i + \omega (D^{ii})^{-1} \left(b^i - \sum_{j \in B} A^{ij} x^j \right),$$

where $B = \{1, \dots, \beta\}$ are the block indices.

In the case of the zebra block versions for the model problem (cf. Example 3.16c), parallel computations inside of the same ‘colour’ are possible (‘white’ for odd indices, ‘black’ for even indices).

Exercise 3.20. Prove that the statement of Exercise 3.18 can be generalised to the case of the block-SOR method.

3.4 Computational Work of the Iterations

3.4.1 Case of General Sparse Matrices

In the following, let $s(n) \leq C_A n$ be the number of the nonzero entries of A (cf. (2.28)). Since the diagonal entries vanish, the iteration matrix $M^{\text{Jac}} = D^{-1}(D - A)$ of the Jacobi iteration contains $s(n) - n \leq (C_A - 1)n$ nonzero elements. First, $\hat{b} := N^{\text{Jac}}b = D^{-1}b$ is computed and stored instead of b . The multiplication $M^{\text{Jac}}x$ requires $(C_A - 1)n$ multiplications and $(C_A - 2)n$ additions. Hence the work of $x^{m+1} = M^{\text{Jac}}x^m + \hat{b}$ amounts to

$$\text{Work}(\varPhi^{\text{Jac}}, A) \leq 2(C_A - 1)n. \quad (3.20a)$$

Since the Gauss–Seidel method differs from the Jacobi iteration only by the fact that, in part, components of x^{m+1} instead of x^m are used, we obtain the same amount of work:

$$\text{Work}(\varPhi^{\text{GS}}, A) \leq 2(C_A - 1)n. \quad (3.20b)$$

To save as many operations in the SOR method as possible, we move the term $a_{ij}x_j^m/a_{ii}$ for $j = i$ out of the bracket and obtain

$$x_i^{m+1} := (1 - \omega)x_i^m - \omega \left(\sum_{j=1}^{i-1} a_{ij}x_j^{m+1} + \sum_{j=i+1}^n a_{ij}x_j^m - b_i \right) / a_{ii},$$

where $\omega' := 1 - \omega$ is precomputed. Similarly, a_{ij}/a_{ii} and b_i/a_{ii} may be assumed to be available. This yields

$$\text{Work}(\varPhi^{\text{SOR}}, A) \leq 2(C_A + 1)n \quad \{ = 2C_A n, \text{ respectively} \}. \quad (3.20c)$$

The case $\{\dots\}$ in brackets refers to a further possibility. Beforehand, one may also multiply a_{ij}/a_{ii} and b_i/a_{ii} by ω . However note that, occasionally, ω can vary during the iteration (cf. §4.6.4).

For the Richardson method (3.4), one finds

$$\text{Work}(\Phi^{\text{Rich}}, A) \leq 2C_A n \quad \{ = (2C_A + 2)n, \text{ respectively}\}, \quad (3.20d)$$

if $I - \Theta A$ is available in this form. The value in brackets is valid if $x^{m+1} = x^m - \Theta(Ax^m - b)$ is evaluated via the defect $d := Ax^m - b$ and $x^{m+1} = x^m - \Theta d$.

The *cost factors* defined in §2.3.1 are

$$C_\Phi^{\text{Jac}} = C_\Phi^{\text{GS}} = 2 - 2/C_A, \quad C_\Phi^{\text{SOR}} = 2 + 1/C_A \quad \{ = 2, \text{ respectively}\}.$$

The amount of work of the block variants depends on the structure of the diagonal blocks. For further considerations, we assume

there are β blocks of size n/β ,

the amount of work for solving $A^{ii}u = z$ is $\leq C_\beta n/\beta$,

$A - D$ has $s_1(n) \leq C_{AD}n$ nonzero elements,

where $D = \text{blockdiag}\{A\}$. Then the operation count yields

$$\text{Work}(\Phi_{\text{block}}^{\text{Jac}}, A) = \text{Work}(\Phi_{\text{block}}^{\text{GS}}, A) \leq (C_B + 2C_{AD})n, \quad (3.20e)$$

$$\text{Work}(\Phi_{\text{block}}^{\text{SOR}}, A) \leq (C_B + 2C_{AD} + 3)n, \quad \{(C_B + 2C_{AD} + 2)n, \text{ resp.}\}, \quad (3.20f)$$

where the bracket refers to the case that the damping factor is already combined with the matrix entries, so that no multiplication by ω or Θ is necessary.

3.4.2 Amount of Work in the Model Case

For the Poisson model problem in §1.2, one needs less operations than described in (3.20a–f) with

$$C_A = 5, \quad C_B = 5, \quad C_{AD} = 2.$$

The reason is that multiplications by the coefficients -1 can be omitted. Computing $D^{-1}b$ beforehand corresponds to the replacement of f with $h^2 f$. Counting the operations, we obtain the following:

$$\begin{aligned} \text{Work}(\Phi^{\text{Jac}}, A) &= \text{Work}(\Phi^{\text{GS}}, A) \leq 5n, \\ \text{Work}(\Phi^{\text{SOR}}, A) &= \text{Work}(\Phi_\Theta^{\text{Rich}}, A) \leq 7n, \\ \text{Work}(\Phi_{\text{block}}^{\text{Jac}}, A) &= \text{Work}(\Phi_{\text{block}}^{\text{GS}}, A) \leq 7n, \\ \text{Work}(\Phi_{\text{block}}^{\text{SOR}}, A) &\leq 9n, \end{aligned} \quad (3.21)$$

where the number $9n$ of the last line holds in the case that instead of $h^2 A^{ii}$ the matrices $h^2 A^{ii}/\omega$ are decomposed into triangular LU factors.

In the model case, using the numbers in (3.21), the *cost factors* are equal to

$$C_{\bar{\phi}}^{\text{Jac}} = C_{\bar{\phi}}^{\text{GS}} = 1, \quad (3.22a)$$

$$C_{\bar{\phi}}^{\text{SOR}} = C_{\bar{\phi}}^{\text{Rich}} = C_{\bar{\phi}}^{\text{blockJac}} = C_{\bar{\phi}}^{\text{blockGS}} = 1.4, \quad (3.22b)$$

$$C_{\bar{\phi}}^{\text{blockSOR}} = 9/5 = 1.8. \quad (3.22c)$$

These numbers do not merit attention before we also know the respective convergence speeds. Then we are able to weigh whether, e.g., the block-Gauss–Seidel method is preferable to the pointwise Gauss–Seidel iteration in spite of its 1.4-fold amount of work.

3.5 Convergence Analysis

In the following, the convergence considerations are based on the assumption that A is positive definite or has weaker but related properties (positive definite Hermitian part or positive spectrum). In Chapter 4 and §7.2.2, other assumptions will be posed.

3.5.1 Richardson Iteration

The iteration matrix of the Richardson method is M_{Θ}^{Rich} :

$$\begin{aligned} x^{m+1} &= x^m - \Theta(Ax^m - b) \quad (\Theta \in \mathbb{C}), \\ M_{\Theta}^{\text{Rich}} &= I - \Theta A. \end{aligned}$$

We may express $M_{\Theta}^{\text{Rich}} = P(A)$ by the polynomial $P(\xi) = 1 - \Theta\xi$. If $\lambda_{\nu} \in \sigma(A)$ are the eigenvalues of A , $\mu_{\nu} = P(\lambda_{\nu}) = 1 - \Theta\lambda_{\nu}$ are those of M_{Θ}^{Rich} (cf. Lemma A.11a). Since the function $|1 - \Theta\xi|$ has no local maxima, we obtain the following statement.

Lemma 3.21. *Assume that A has only real eigenvalues. Let $\lambda_{\min} := \min\{\lambda \in \sigma(A)\}$ and λ_{\max} denote the extreme eigenvalues of A . Then the spectrum of M_{Θ}^{Rich} is real for any $\Theta \in \mathbb{R}$, i.e., $\sigma(M_{\Theta}^{\text{Rich}}) \subset \mathbb{R}$. The following characterisation also holds for complex Θ :*

$$\rho(M_{\Theta}^{\text{Rich}}) = \max \{ |1 - \Theta\lambda_{\min}|, |1 - \Theta\lambda_{\max}| \} \quad \text{for all } \Theta \in \mathbb{C}. \quad (3.23)$$

A first conclusion from (3.23) is that $\rho(M_{\Theta}^{\text{Rich}})$ improves when Θ is replaced with $\Re \Theta$. Therefore we restrict the following analysis to real values of Θ . Positive definite matrices have a positive spectrum (cf. Lemma C.3). Below we only need the latter property.

Theorem 3.22 (convergence of the Richardson iteration). Assume that A has only positive eigenvalues with $\lambda_{\max}(A)$ denoting the maximal eigenvalue. For real Θ , the Richardson method converges if and only if

$$0 < \Theta < 2/\lambda_{\max}(A). \quad (3.24)$$

The convergence rate is described by (3.23).

Proof. (i) For $0 < \Theta < 2/\lambda_{\max}$, we have $-1 < 1 - \Theta\lambda_{\max} \leq 1 - \Theta\lambda_{\min} < 1$. (3.23) yields $\rho(M_{\Theta}^{\text{Rich}}) < 1$, proving convergence.

(ii) Now assume convergence: $\rho(M_{\Theta}^{\text{Rich}}) < 1$. We conclude from

$$1 > \rho(M_{\Theta}^{\text{Rich}}) \geq_{(3.23)} |1 - \Theta\lambda_{\max}| \geq 1 - \Theta\lambda_{\max}$$

that $\Theta\lambda_{\max} > 0$ and therefore $\Theta > 0$ since λ_{\max} is positive. Similarly, the inequalities $-1 < -\rho(M_{\Theta}^{\text{Rich}}) \leq -|1 - \Theta\lambda_{\max}| \leq 1 - \Theta\lambda_{\max}$ show that $\Theta\lambda_{\max} < 2$ and prove that the second inequality in (3.24) is also necessary. \square

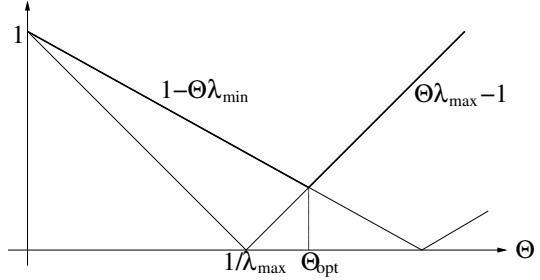


Fig. 3.1 Optimal Θ .

Equation (3.23) allows us to determine the factor Θ minimising the rate $\rho(M_{\Theta}^{\text{Rich}})$. The *optimal factor* Θ results as the intersection of the lines $y(\Theta) = \Theta\lambda_{\max} - 1$ and $y(\Theta) = 1 - \Theta\lambda_{\min}$ (see Fig. 3.1).

Theorem 3.23 (optimal Θ). Assume that A has only positive eigenvalues. λ_{\max} and λ_{\min} are the respective maximal and minimal eigenvalues of A . The optimal convergence rate of Richardson's method is attained for

$$\Theta_{\text{opt}} = \frac{2}{\lambda_{\max} + \lambda_{\min}}, \quad \text{so that } \rho(M_{\Theta_{\text{opt}}}^{\text{Rich}}) = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}. \quad (3.25)$$

Now we pose the stronger assumption that A is positive definite and prove norm estimates of $M_{\Theta_{\text{opt}}}^{\text{Rich}}$ instead of rates.

Corollary 3.24. Assume that the matrix A is positive definite and Θ is real. Then the Richardson method converges if and only if

$$0 < \Theta < 2/\|A\|_2. \quad (3.26a)$$

The convergence is monotone with respect to the Euclidean norm $\|\cdot\|_2$ and to the energy norm $\|\cdot\|_A$ defined in (C.5a,c) by $\|x\|_A = \|A^{1/2}x\|_2$. Furthermore, the convergence rate and the contraction numbers coincide:

$$\rho(M_{\Theta}^{\text{Rich}}) = \|M_{\Theta}^{\text{Rich}}\|_2 = \|M_{\Theta}^{\text{Rich}}\|_A. \quad (3.26b)$$

The optimal convergence rate (3.25) can be expressed as a function of the Euclidean condition and the condition number $\kappa(A) = \text{cond}_2(A) := \|A\|_2 \|A^{-1}\|_2$:

$$\rho(M_{\Theta_{\text{opt}}}^{\text{Rich}}) = \frac{\kappa(A) - 1}{\kappa(A) + 1} \quad \text{for } \Theta_{\text{opt}} = \frac{2\|A^{-1}\|_2}{\kappa(A) + 1}. \quad (3.26c)$$

Proof. (3.26a) follows from $\lambda_{\max} = \|A\|_2$. Using $\lambda_{\min} = 1/\|A^{-1}\|_2$ (cf. (B.15)) and $\kappa(A) = \lambda_{\max}/\lambda_{\min}$, we can transform (3.25) into (3.26c). As A is normal, $M := M_{\Theta}^{\text{Rich}}$ is also normal, proving $\rho(M) = \|M\|_2$ (cf. (B.21b)). The second equality in (3.26b) follows from $\|M\|_A = \|A^{\frac{1}{2}}MA^{-\frac{1}{2}}\|_2$ and the commutativity $A^{1/2}M = MA^{1/2}$ (cf. (C.5d) and Remark C.6b). \square

Assuming a real spectrum $\sigma(A)$, the signs of the eigenvalues must coincide. In the case of only negative eigenvalues, the previous statements stay valid if the sign of Θ is reversed.

Exercise 3.25. If A has at least one positive and one negative eigenvalue, the Richardson method diverges for any choice of $\Theta \in \mathbb{C}$.

Nevertheless, the assumption of positivity can be weakened.

Exercise 3.26. (a) Let the spectrum $\sigma(A)$ be contained in a closed circle around $\mu \in \mathbb{C} \setminus \{0\}$ with radius $r < |\mu|$. Prove that the choice $\Theta = 1/\mu$ leads to convergence of the Richardson method:

$$\rho(M_{\Theta}^{\text{Rich}}) \leq r/|\mu| < 1.$$

(b) Let γ be an arbitrary straight line in the complex plane passing through the origin $z = 0$. Then $\mathbb{C} \setminus \gamma$ consists of two half planes. If $\sigma(A)$ lies in one of the half planes, the Richardson iteration converges for a suitable Θ .

The next theorem offers a necessary and sufficient criterion concerning convergence in the general case; however, in practice it might be hard to apply.

Theorem 3.27. Let $\sigma(A)$ be the spectrum of a general matrix A . The convex hull of $\sigma(A)$ is defined by

$$\Sigma(A) := \left\{ \sum_{\lambda \in \sigma(A)} \alpha_{\lambda} \lambda : \alpha_{\lambda} \geq 0 \text{ with } \sum_{\lambda \in \sigma(A)} \alpha_{\lambda} = 1 \right\}.$$

Then the Richardson iteration converges for a suitable $\Theta \in \mathbb{C}$ if and only if $0 \notin \Sigma(A)$.

Proof. (i) For $\Theta = 0$, the Richardson iteration diverges since $\rho(M_0^{\text{Rich}}) = \rho(I) = 1$. In the following, we assume that $\Theta \neq 0$.

(ii) The spectrum of M_Θ^{Rich} is $\sigma(M_\Theta^{\text{Rich}}) = 1 - \Theta\sigma(A)$, where the right-hand side is the set $\{1 - \Theta\lambda : \lambda \in \sigma(A)\}$. One easily sees that its convex hull is $\Sigma(M_\Theta^{\text{Rich}}) = 1 - \Theta\Sigma(A)$. Obviously, $0 \notin \Sigma(A)$ is equivalent to $1 \notin \Sigma(M_\Theta^{\text{Rich}})$.

(iii) Assume $0 \in \Sigma(A)$. The conclusion $1 \in \Sigma(M_\Theta^{\text{Rich}})$ implies that there is some $\mu \in \sigma(M_\Theta^{\text{Rich}})$ with $\Re \mu \geq 1$. Therefore $\rho(M_\Theta^{\text{Rich}}) \geq |\mu| \geq 1$ proves that the Richardson iteration for an arbitrary Θ does not converge.

(iv) If $0 \notin \Sigma(A)$, the condition of Exercise 3.26b implies convergence for a suitable Θ . \square

In the non-Hermitian case, A can be split into a Hermitian and a skew-Hermitian part (cf. (B.29)):

$$A = A_0 + iA_1 \quad \text{with } A_0 := \frac{1}{2}(A + A^H), \quad A_1 := \frac{1}{2i}(A - A^H). \quad (3.27)$$

Together with suitable estimates of the Hermitian matrices A_0 and A_1 , convergence statements for the Richardson method can be formulated (cf. Theorems 3.28 and 3.30, and (3.33a,b), Samarskii–Nikolaev [330, §6.4]).

Theorem 3.28. *For A and A_0 in (3.27), constants $0 < \lambda \leq \Lambda$ are assumed to exist such that*

$$0 < \lambda I \leq A_0, \quad (3.28a)$$

$$A^H A \leq \Lambda A_0. \quad (3.28b)$$

Then, for Θ satisfying

$$0 < \Theta < \frac{2}{\Lambda}, \quad (3.28c)$$

Richardson's iteration converges monotonically with respect to the Euclidean norm:

$$\rho(M_\Theta^{\text{Rich}}) \leq \|M_\Theta^{\text{Rich}}\|_2 \leq \sqrt{1 - \Theta\lambda(2 - \Theta\Lambda)} < 1. \quad (3.29a)$$

The bound on the right-hand side is minimal for $\Theta' := 1/\Lambda$:

$$\rho(M_{\Theta'}^{\text{Rich}}) \leq \|M_{\Theta'}^{\text{Rich}}\|_2 \leq \sqrt{1 - \lambda/\Lambda}. \quad (3.29b)$$

Proof. (3.28a,b) leads to the estimate

$$\begin{aligned} (M_\Theta^{\text{Rich}})^H (M_\Theta^{\text{Rich}}) &= (I - \Theta A)^H (I - \Theta A) = I - \Theta(A + A^H) + \Theta^2 A^H A \\ &\stackrel{(3.28b)}{\leq} I - 2\Theta A_0 + \Theta^2 \Lambda A_0 = I - \Theta \underbrace{(2 - \Theta\Lambda)}_{>0} A_0 \\ &\stackrel{(3.28a)}{\leq} I - \Theta\lambda(2 - \Theta\Lambda)I, \end{aligned}$$

entailing $\|M_\Theta^{\text{Rich}}\|_2^2 = \|(M_\Theta^{\text{Rich}})^H (M_\Theta^{\text{Rich}})\|_2 \leq 1 - \Theta\lambda(2 - \Theta\Lambda)$ and therefore (3.29a) (cf. (C.3f)). The inequality $1 - \Theta\lambda(2 - \Theta\Lambda) < 1$ follows from (3.28c) and proves convergence. (3.29b) is easy to verify. \square

Condition (3.28b) can also be written as

$$\langle Ax, Ax \rangle \leq \langle A_0 x, x \rangle \quad \text{for all } x \in \mathbb{C}^I.$$

In the positive definite case, the inequalities (3.28a,b) are satisfied by $\lambda = \lambda_{\min}$, $\Lambda = \lambda_{\max}$. However, the optimal parameters Θ_{opt} and Θ' from Theorems 3.23 and 3.28 are different and lead to different bounds in (3.26c) and (3.29b).

Remark 3.29. Theorem 3.28 is a typical example where not the primary quantity $\rho(M_{\Theta}^{\text{Rich}})$ is optimised, but its bound (here, (3.29a)). Often the optimal parameter of an auxiliary problem is much easier to obtain than the true optimum. In any case, this technique also yields a bound for the optimal $\rho(M_{\Theta}^{\text{Rich}})$.

A stronger estimate than (3.29b) is possible if, in addition, an estimate of the skew-Hermitian part iA_1 can be used.

Theorem 3.30. Assume that for A_0, A_1 in (3.27) there are constants $0 < \lambda \leq \Lambda$ and $\tau \geq 0$ with

$$\lambda I \leq A_0 \leq \Lambda I, \tag{3.30a}$$

$$\|A_1\|_2 \leq \tau. \tag{3.30b}$$

Then the Richardson method converges for

$$0 < \Theta < \frac{2\lambda}{\lambda\Lambda + \tau^2} \tag{3.31a}$$

monotonically with respect to the Euclidean norm:

$$\|M_{\Theta}^{\text{Rich}}\|_2 \leq \frac{1}{2}\Theta(\Lambda - \lambda) + \sqrt{\left[1 - \frac{1}{2}\Theta(\Lambda + \lambda)\right]^2 + \Theta^2\tau^2} < 1. \tag{3.31b}$$

Optimising the upper bound yields

$$\|M_{\Theta'}^{\text{Rich}}\|_2 \leq \frac{1 - \xi}{1 + \xi} \quad \text{for } \Theta' = \frac{2}{\Lambda + \lambda} \left(1 - s \frac{1 - \xi}{1 + \xi}\right) \tag{3.31c}$$

with $s := \tau/\sqrt{\lambda\Lambda + \tau^2}$ and $\xi := \frac{1 - s}{1 + s} \frac{\lambda}{\Lambda}$.

Proof. Let $\vartheta \in (0, 1)$ be arbitrary. In analogy to (3.23), the first term in

$$\begin{aligned} \|M_{\Theta}^{\text{Rich}}\|_2 &= \|I - \Theta A\|_2 = \|[\vartheta I - \Theta A_0] + [(1 - \vartheta)I - i\Theta A_1]\|_2 \\ &\leq \|\vartheta I - \Theta A_0\|_2 + \|(1 - \vartheta)I - i\Theta A_1\|_2 \end{aligned}$$

is bounded by $\|\vartheta I - \Theta A_0\|_2 \leq \max\{|\vartheta - \Theta\lambda|, |\vartheta - \Theta\Lambda|\}$. Since the matrix $C := (1 - \vartheta)I - i\Theta A_1$ is normal, $\|C\|_2 = \rho(C)$ holds. From

$$\sigma(C) = \{1 - \vartheta - \vartheta\Theta\mu : \mu \in \sigma(A_1)\} \quad \text{and} \quad \sigma(A_1) \subset [-\tau, \tau] \quad (\text{cf. (C.3e)}),$$

we conclude that $\rho(C) \leq [(1 - \vartheta)^2 + \Theta^2\tau^2]^{1/2}$. Together, we obtain

$$\|M_\Theta^{\text{Rich}}\|_2 \leq \max \{|\vartheta - \Theta\lambda|, |\vartheta - \Theta\Lambda|\} + [(1 - \vartheta)^2 + \Theta^2\tau^2]^{1/2}.$$

The optimal choice $\vartheta = \frac{1}{2}\Theta(\Lambda + \lambda)$ yields (3.31b). Under condition (3.31a), one verifies that the bound in (3.31b) remains below one. \square

For the Hermitian case ($\tau = 0$), the estimate (3.31c) corresponds exactly to the convergence rate (3.26c). For $\tau \neq 0$, the convergence rate can be estimated even better than by the $\|M_\Theta^{\text{Rich}}\|_2$ bound in (3.31c).

Theorem 3.31. *Under the assumption (3.30a,b), the estimate*

$$\rho(M_\Theta^{\text{Rich}}) \leq r_\Theta := \sqrt{\Theta^2\tau^2 + \max \{|1 - \Theta\lambda|, |1 - \Theta\Lambda|\}}$$

holds with λ and Λ as in Theorem 3.30. Convergence is ensured in the form $r_\Theta < 1$ if

$$0 < \Theta < \overline{\Theta} \quad \text{with } \overline{\Theta} := \begin{cases} 2\Lambda / (\Lambda^2 + \tau^2) & \text{if } \tau^2 < \lambda\Lambda, \\ 2\lambda / (\Lambda^2 + \tau^2) & \text{if } \tau^2 \geq \lambda\Lambda. \end{cases} \quad (3.32a)$$

r_Θ is minimal for

$$\Theta' := \min \left\{ \frac{\lambda}{\lambda^2 + \tau^2}, \frac{2}{\lambda + \Lambda} \right\}.$$

Further, the norm estimate (3.32b) holds:

$$\|(M_\Theta^{\text{Rich}})^m\|_2 \leq 2r_\Theta^m \quad (m \geq 0). \quad (3.32b)$$

Proof. (B.33) shows that r_Θ is an upper bound of the numerical radius $r(M_\Theta^{\text{Rich}})$ of the iteration matrix. Analysing r_Θ as a function of Θ yields (3.32a) and the value Θ' . (3.32b) follows from (B.28d). \square

While (3.30b) represents the inequality $-\tau I \leq A_1 \leq \tau I$, it is also possible to require an estimate of A_1 in relation to A_0 by

$$A_1^2 \leq \tau A_0 \quad (3.33a)$$

or even only by

$$-\sqrt{\tau} A_0^{1/2} \leq A_1 \leq \sqrt{\tau} A_0^{1/2}. \quad (3.33b)$$

Inequality (3.33a) implies (3.33b). From (3.33b), using (3.30a), we arrive at the estimate (3.30b) with $\sqrt{\tau\Lambda}$ instead of τ . An estimate based directly on (3.30a) and (3.33a) can be found in Samarskii–Nikolaev [330, page 101]. Note that iA_1 in the next corollary is not necessarily the skew-Hermitian part of A .

Corollary 3.32. Assume $A = A_0 + iA_1$ with a positive definite A_0 . Let (3.30a) and

$$A_1^H A_1 \leq \tau A_0$$

be valid. Then

$$r(M_\Theta^{\text{Rich}}) \leq \begin{cases} (1 - \Theta\lambda)^2 + \Theta^2\lambda\tau & \text{if } 0 \leq \Theta \leq \Theta^*, \\ (1 - \Theta\Lambda)^2 + \Theta^2\Lambda\tau & \text{if } \Theta \geq \Theta^*, \end{cases}$$

holds for the numerical radius, where $\Theta^* := \frac{2}{\lambda + \Lambda + \tau}$. The optimal parameter Θ minimising the bound is

$$\Theta_{\text{opt}} := \min\{1, \kappa\}\Theta^*, \quad \text{where } \kappa := \frac{\lambda + \Lambda + \tau}{2(\lambda + \Lambda)}.$$

This value yields

$$r(M_{\Theta_{\text{opt}}}^{\text{Rich}})^2 \leq 1 - \frac{1 - \rho_0}{1 + \rho_0} \left(2 - \frac{1}{\kappa} \right) \min\{1, \kappa\}, \quad \rho_0 := \frac{1 - \lambda/\Lambda}{1 + \lambda/\Lambda}.$$

We conclude with a remark about monotone convergence.

Corollary 3.33. Let $K > 0$ be any positive definite matrix and $\|\cdot\|_K$ the related norm (C.5a,c). Monotone convergence of Richardson's iteration with respect to $\|\cdot\|_K$ can be obtained by replacing the assumption (3.28a,b) in Theorem 3.28 with

$$A^H K + KA \geq 2\lambda K, \quad A^H KA \leq \frac{1}{2}\Lambda(A^H K + KA). \quad (3.34)$$

These inequalities are equivalent to

$$\begin{aligned} \Re \langle Ax, Kx \rangle &\geq \lambda \langle Kx, x \rangle && \text{for all } x \in \mathbb{C}^I, \\ \langle Ax, KA x \rangle &\leq \Lambda \Re \langle Ax, Kx \rangle && \text{for all } x \in \mathbb{C}^I. \end{aligned}$$

Under assumption (3.34), the corresponding estimates (3.29a, b) hold with the K -norm $\|M_\Theta^{\text{Rich}}\|_K$ instead of $\|M_\Theta^{\text{Rich}}\|_2$.

Proof. Let $M := M_\Theta^{\text{Rich}} = I - \Theta A$. Using (3.34), we conclude from

$$\begin{aligned} M^H K M &= K - \Theta(A^H K + KA) + \Theta^2 A^H K A \\ &\leq K - \Theta(1 - \frac{1}{2}\Theta\Lambda)(A^H K + KA) \\ &\leq K - \Theta(2 - \Theta\Lambda)\lambda K \end{aligned}$$

that $I - \Theta\lambda(2 - \Theta\Lambda)I \geq K^{-1/2}M^H K M K^{-1/2} = \hat{M}^H \hat{M}$ holds with $\hat{M} := K^{1/2}MK^{-1/2}$. This is equivalent to

$$1 - \Theta\lambda(2 - \Theta\Lambda) \geq \|\hat{M}\|_2^2 = \|M\|_K^2 \geq \rho(M)^2 \quad (\text{cf. (C.3f)})$$

and shows that (3.29a) is valid with respect to the $\|\cdot\|_K$ norm. \square

3.5.2 Convergence Criterion for Positive Definite Iterations

In §6 we shall investigate the set \mathcal{L}_{pos} of *positive definite iterations* in more detail. For a system $Ax = b$, a positive definite iteration is associated with matrices N and W of the second and third normal forms that are also positive definite if A is so:

$$A > 0 \implies N > 0 \text{ and } W > 0.$$

Examples are Richardson's iteration for $\Theta > 0$ ($N = \Theta I > 0$) and the Jacobi iteration ($W = D$).

Theorem 3.34. *Let W be the matrix of the third normal form (2.12); i.e., the iteration matrix is $M = I - W^{-1}A$ (cf. (2.13')).*

(a) *Under the assumption*

$$2W > A > 0, \quad (3.35a)$$

the iteration $x^{m+1} = x^m - W^{-1}(Ax^m - b)$ converges. Furthermore, the convergence is monotone with respect to the energy norm $\|\cdot\|_A$ and the norm $\|\cdot\|_W$:

$$\rho(M) = \|M\|_A = \|M\|_W < 1. \quad (3.35b)$$

(b) *For real λ and Λ with $0 < \lambda \leq \Lambda$, assume*

$$0 < \lambda W \leq A \leq \Lambda W. \quad (3.35c)$$

Then the spectrum of M is real and is contained in

$$\sigma(M) \subset [1 - \Lambda, 1 - \lambda]. \quad (3.35d)$$

The convergence rate is

$$\rho(M) = \|M\|_A = \|M\|_W \leq \max\{1 - \lambda, \Lambda - 1\}, \quad (3.35e)$$

where equality holds instead of ' \leq ' if λ and Λ are the optimal bounds in (3.35c). These optimal bounds can be expressed as

$$\lambda = 1/\|W^{1/2}A^{-1}W^{1/2}\|_2, \quad \Lambda = \|W^{-1/2}AW^{-1/2}\|_2. \quad (3.35f)$$

Convergence is equivalent to $0 < \lambda \leq \Lambda < 2$ with λ, Λ in (3.35f).

(c) *Assume $W > 0$, $A = A^H$, and $0 < \lambda \leq \Lambda$. The conditions (3.35c) and (3.35d) are equivalent. In particular, (3.35a) is equivalent to $\sigma(M) \subset (-1, 1)$, and the equivalence (3.35g) holds:*

$$W \geq A > 0 \iff \sigma(M) \subset [0, 1]. \quad (3.35g)$$

The proof will be postponed to §6.2 (cf. Theorem 6.10). The analysis of the damped iteration can also be found in §6.2.1.

The optimal bounds λ and Λ in (3.35c) are the minimal and maximal eigenvalues of the generalised eigenvalue problem

$$Ae = \lambda We \quad (W > 0, e \neq 0).$$

The next statement generalises Theorem 3.34 by replacing W with its Hermitian part in (3.35a). A further generalisation with $W + W^H > \frac{1}{2}(A + A^H) > 0$ instead of (3.36) is given in Theorem 7.26. Although the next theorem follows from Theorem 7.26, a direct proof is given.

Theorem 3.35. *Let A be positive definite, while the matrix W of the third normal form satisfies*

$$W + W^H > A > 0. \quad (3.36)$$

Then W is regular and the iteration converges monotonically with respect to the energy norm $\|\cdot\|_A$:

$$\rho(M) \leq \|M\|_A < 1 \quad \text{for } M = I - W^{-1}A.$$

Proof. (a) To prove the regularity of the matrix W , assume that $Wx = 0$. From $0 = \langle Wx, x \rangle + \langle x, Wx \rangle = \langle (W + W^H)x, x \rangle > 0$, we conclude that $x = 0$, since $W + W^H > 0$. Hence, W is regular.

(b) As $\rho(M) \leq \|M\|_A$ by (B.20b), only $\|M\|_A < 1$ has to be shown. By (C.5d), we have $\|M\|_A = \|A^{1/2}MA^{-1/2}\|_2 = \|\hat{M}\|_2$ for $\hat{M} = I - A^{1/2}W^{-1}A^{1/2}$. One verifies that

$$\begin{aligned} \hat{M}^H \hat{M} &= (I - A^{1/2}W^{-H}A^{1/2})(I - A^{1/2}W^{-1}A^{1/2}) \\ &= I - A^{1/2}(W^{-H} + W^{-1})A^{1/2} + A^{1/2}W^{-H}AW^{-1}A^{1/2} \\ &= I - A^{1/2}W^{-H}(W + W^H)W^{-1}A^{1/2} + A^{1/2}W^{-H}AW^{-1}A^{1/2} \\ &< I - A^{1/2}W^{-H}AW^{-1}A^{1/2} + A^{1/2}W^{-H}AW^{-1}A^{1/2} = I. \end{aligned} \quad (3.37)$$

Hence, by Lemma C.5b, $\|M\|_A = \|\hat{M}\|_2 = \rho(\hat{M}^H \hat{M})^{1/2} < \rho(I)^{1/2} = 1$. \square

3.5.3 Jacobi Iteration

Theorem 3.36. *Sufficient for the convergence of the Jacobi iteration (3.7b) are the conditions (3.38), which also imply $\sigma(M^{\text{Jac}}) \subset (-1, 1)$:*

$$A \text{ and } 2D - A \text{ are positive definite, i.e., } 2D > A > 0. \quad (3.38)$$

The contraction numbers with respect to the norms $\|\cdot\|_A$ and $\|\cdot\|_D$ coincide with the convergence rate:

$$\rho(M^{\text{Jac}}) = \|M^{\text{Jac}}\|_A = \|M^{\text{Jac}}\|_D < 1. \quad (3.39)$$

Proof. Choose $W := D$ in Theorem 3.34a. \square

Remark 3.37. (a) The matrices A and $2D - A$ have identical diagonal entries, whereas their off-diagonal entries have opposite signs.

(b) The statements $A > 0$ and $2D - A > 0$ are identical for 2×2 matrices, but they may differ for matrices of size $n \times n$ with $n > 2$.

Remark 3.38 (block-Jacobi iteration). None of the proofs makes use of the diagonal form of the matrix D . We only used the fact that $D > 0$ follows from $A > 0$. Therefore, if $D = \text{diag}\{A\}$ is replaced with $D := \text{blockdiag}\{A\}$, all statements above remain valid for the block-Jacobi iteration.

A convergence criterion involving a different kind of positivity is the diagonal dominance as discussed in Theorem 7.20 and Proposition 7.23.

3.5.4 Gauss–Seidel and SOR Iterations

Theorem 3.39. *The Gauss–Seidel iteration converges for positive definite matrices⁴ A . The convergence is monotone with respect to the energy norm:*

$$\rho(M^{\text{GS}}) \leq \|M^{\text{GS}}\|_A < 1. \quad (3.40)$$

Proof. $A > 0$ implies $D > 0$ and $E^H = F$. The matrix $W = W^{\text{GS}}$ in (3.12) satisfies $W + W^H = D - E + (D - E)^H = 2D - E - F = D + A > A$. The matrices D , E , and F are defined in (3.11a–d). Hence, condition (3.36) is satisfied, and Theorem 3.35 proves (3.40). \square

Since the Gauss–Seidel iteration is the special case $\omega = 1$ of the SOR iteration, we formulate further convergence statements and quantitative estimates for the SOR method only.

Lemma 3.40 (Kahan [232]). *For any $A \in \mathfrak{D}(\Phi_\omega^{\text{SOR}})$ (cf. (3.6)), the following inequality holds:*

$$\rho(M_\omega^{\text{SOR}}) \geq |\omega - 1| \quad \text{for all } \omega \in \mathbb{C}. \quad (3.41)$$

Therefore, $|\omega - 1| < 1$ is necessary for convergence. For real ω , this condition is equivalent to $0 < \omega < 2$.

Proof. Let $n := \# I$ be the matrix size. Since $I - \omega L$ and $(1 - \omega)I + \omega U$ are triangular matrices, $\det(I - \omega L) = 1$ and $\det((1 - \omega)I + \omega U) = (1 - \omega)^n$ hold, the representation $M_\omega^{\text{SOR}} = (I - \omega L)^{-1} \{(1 - \omega)I + \omega U\}$ (cf. (3.15b)) yields

$$\det(M_\omega^{\text{SOR}}) = \frac{1}{\det(I - \omega L)} \det((1 - \omega)I + \omega U) = (1 - \omega)^n.$$

On the other hand, each determinant is the product of all eigenvalues of the matrix: $\det(M_\omega^{\text{SOR}}) = \prod_{\nu=1}^n \lambda_\nu$ (λ_ν are the eigenvalues of M_ω^{SOR}). Together, we obtain $\prod_{\nu=1}^n \lambda_\nu = (1 - \omega)^n$ or $\prod |\lambda_\nu| = |1 - \omega|^n$. Therefore, at least one factor (eigenvalue) λ_ν must exist with $|\lambda_\nu| \geq |1 - \omega|$. This eigenvalue proves (3.41). \square

⁴ For an early mentioning of this result see von Mises–Pollaczek–Geiringer [381, §3] from 1929.

Inequality (3.41) implies that $0 < \omega < 2$ is a necessary condition. The next theorem⁵ shows that $0 < \omega < 2$ is also sufficient for convergence.

Theorem 3.41 (Ostrowski [302]). *Assume that A is positive definite and can be split into*

$$A = D - E - E^H \quad (3.42a)$$

with the properties (3.42b,c):

$$E \text{ is a strictly lower triangular matrix,} \quad (3.42b)$$

$$D \text{ is the diagonal of } A \text{ and regular.} \quad (3.42c)$$

Furthermore, assume

$$0 < \omega < 2. \quad (3.42d)$$

Then the SOR iteration (3.15a–c) converges:

$$\rho(M_\omega^{\text{SOR}}) < 1. \quad (3.42e)$$

The convergence is monotone with respect to the energy norm:

$$\rho(M_\omega^{\text{SOR}}) \leq \|M_\omega^{\text{SOR}}\|_A < 1. \quad (3.42f)$$

Instead of this theorem we are going to prove a more general statement. The splitting (3.42a) does not differ from $A = D - E - F$ in (3.11a–d), since $F = E^H$ holds for any Hermitian matrix A . The assumptions of Theorem 3.41 can be weakened. The iteration defined by (3.42a,b',c') is more general than the standard SOR method.

Corollary 3.42. Let $A > 0$. The statements (3.42e,f) of Theorem 3.41 remain valid if instead of (3.42b) and (3.42c), we only assume (3.42a) with

$$E \text{ is arbitrary,} \quad (3.42b')$$

$$D \text{ is an arbitrary positive definite matrix.} \quad (3.42c')$$

Under the conditions (3.42a,c',d), the matrix $D - \omega E$ is always regular.

By Lemma C.4e, the diagonal D in (3.42c) is a positive definite matrix and hence also satisfies (3.42c'). We remark that the assumption ‘ A is positive definite’ in Theorem 3.41 is not only sufficient but also necessary (cf. Varga [375, p.77]).

Proof. The matrix of the third normal form is $W = W_\omega^{\text{SOR}} = \frac{1}{\omega}D - E$ (cf. (3.15e)). Inequality (3.36) of Theorem 3.35 is satisfied:

$$W + W^H = \frac{2}{\omega}D - E - E^H = A + \left(\frac{2}{\omega} - 1\right)D > A > 0 \quad (3.43)$$

because of (3.42c') and the equivalence $\frac{2}{\omega} - 1 > 0 \iff 0 < \omega < 2$. Concerning regularity, compare with Theorem 3.35. \square

⁵ Sometimes the theorem is named after Reich [316].

Theorem 3.41 does not state which ω is the most favourable one. This question will be answered in Theorem 4.27 for the spectral radius $\rho(M_\omega^{\text{SOR}})$. Instead, one can also analyse the contraction number $\|M_\omega^{\text{SOR}}\|_A$ or its upper bound as a function of ω and look for an optimal ω in this sense.

Lemma 3.43. *Under the assumptions $A > 0$ and (3.42a,b',c',d), we have*

$$\|M_\omega^{\text{SOR}}\|_A = \sqrt{1 - \left(\frac{2}{\omega} - 1\right) / \|A^{-1/2}W_\omega^{\text{SOR}}D^{-1/2}\|_2^2}. \quad (3.44)$$

The norm $\|A^{-1/2}W_\omega^{\text{SOR}}D^{-1/2}\|_2^2$ in (3.44) can be estimated by

$$\|A^{-1/2}W_\omega^{\text{SOR}}D^{-1/2}\|_2^2 \leq 1/c \quad (3.45a)$$

if and only if the equivalent inequalities (3.45b) and (3.45c) are valid:

$$WD^{-1}W^H \leq \frac{1}{c}A \quad (W = W_\omega^{\text{SOR}}), \quad (3.45b)$$

$$W^{-H}DW^{-1} \geq cA^{-1}. \quad (3.45c)$$

Proof. (i) The equivalence of (3.45b) and (3.45c) follows from (C.3g). The equivalence of (3.45b) and (3.45a) can be concluded from the fact that

$$\frac{1}{c}I \geq A^{-1/2}WD^{-1}W^H A^{-1/2} = \left[A^{-1/2}WD^{-1/2}\right] \left[A^{-1/2}WD^{-1/2}\right]^H \ (\geq 0)$$

(cf. (C.3b')) is equivalent to

$$\frac{1}{c} \geq \left\| \left[A^{-1/2}WD^{-1/2}\right] \left[A^{-1/2}WD^{-1/2}\right]^H \right\|_2 = \|A^{-1/2}WD^{-1/2}\|_2^2 \quad (\text{cf. (C.3f)}).$$

(ii) Define $\hat{M} = A^{1/2}M_\omega^{\text{SOR}}A^{-1/2}$. From inequality (3.37), the representation (3.43): $A - W - W^H = (1 - \frac{2}{\omega})D$, and (3.45c), we obtain

$$\begin{aligned} \hat{M}^H \hat{M} &= I + A^{1/2}W^{-H}[A - W - W^H]W^{-1}A^{1/2} \\ &= I - \left(\frac{2}{\omega} - 1\right) A^{1/2}W^{-H}DW^{-1}A^{1/2}. \end{aligned}$$

The largest eigenvalue of the product $\hat{M}^H \hat{M}$ is 1 minus the $(\frac{2}{\omega} - 1)$ -fold of the smallest eigenvalue of $A^{1/2}W^{-H}DW^{-1}A^{1/2} = X^{-H}X^{-1} = (XX^H)^{-1}$, where $X := A^{-1/2}WD^{-1/2}$. The latter eigenvalue is equal to $1/\rho(XX^H) = 1/\|X\|_2^2$. Equation (3.44) follows from

$$\|M_\omega^{\text{SOR}}\|_A^2 = \|\hat{M}\|_2 = \rho(\hat{M}^H \hat{M}) = 1 - (\frac{2}{\omega} - 1)/\|A^{-1/2}W_\omega^{\text{SOR}}D^{-1/2}\|_2^2. \quad \square$$

Via $\frac{2}{\omega} - 1$, the right-hand side in (3.44) depends explicitly on ω . However, $W_\omega^{\text{SOR}} = \frac{1}{\omega}D - E$ also contains the parameter ω . The minimisation of the norm $\|M_\omega^{\text{SOR}}\|_A$ is the subject of the following theorem (cf. Samarskii–Nikolaev [330] and Young [412, page 464]).

Theorem 3.44. *Let the constants $\gamma, \Gamma > 0$ fulfil*

$$0 < \gamma D \leq A, \quad (3.46a)$$

$$(\frac{1}{2}D - E)D^{-1}(\frac{1}{2}D - E^H) \leq \frac{1}{4}\Gamma A. \quad (3.46b)$$

Further, assume (3.42a,d). Then, (3.45a–c) holds with the value

$$c = 1/\left[\frac{\Omega^2}{\gamma} + \Omega + \frac{\Gamma}{4}\right] \quad \text{with} \quad \Omega := \frac{2-\omega}{2\omega} \in (0, \infty). \quad (3.46c)$$

The SOR contraction number can be estimated by

$$\|M_\omega^{\text{SOR}}\|_A \leq \sqrt{1 - 2\Omega/\left[\frac{\Omega^2}{\gamma} + \Omega + \frac{\Gamma}{4}\right]}. \quad (3.47a)$$

The right-hand side takes the following minimum (cf. Remark 3.29):

$$\|M_{\omega'}^{\text{SOR}}\|_A \leq \sqrt{\frac{\sqrt{\Gamma} - \sqrt{\gamma}}{\sqrt{\Gamma} + \sqrt{\gamma}}} \quad \text{for} \quad \omega' = \frac{2}{1 + \sqrt{\gamma\Gamma}}. \quad (3.47b)$$

Proof. We rewrite $W := W_\omega^{\text{SOR}} = \frac{1}{\omega}D - E$ as

$$W = \Omega D + \left(\frac{1}{2}D - E\right) \quad \text{with} \quad \Omega := \frac{2-\omega}{2\omega} = \frac{1}{\omega} - \frac{1}{2}$$

and estimate as follows:

$$\begin{aligned} WD^{-1}W^H &= [\Omega D + (\frac{1}{2}D - E)] D^{-1} [\Omega D + (\frac{1}{2}D - E^H)] \\ &= \Omega^2 D + \Omega(\frac{1}{2}D - E + \frac{1}{2}D - E^H) + (\frac{1}{2}D - E)D^{-1}(\frac{1}{2}D - E^H) \\ &\stackrel{(3.42a)}{=} \Omega^2 D + \Omega A + (\frac{1}{2}D - E)D^{-1}(\frac{1}{2}D - E^H) \\ &\stackrel{(3.46a,b)}{\leq} \left[\frac{\Omega^2}{\gamma} + \Omega + \frac{\Gamma}{4}\right] A. \end{aligned}$$

Hence, (3.45b) holds with $\frac{1}{c} = \frac{\Omega^2}{\gamma} + \Omega + \frac{\Gamma}{4}$. Inserting inequality (3.45a) into (3.44), we get (3.47a). The function $\Omega/[\frac{\Omega^2}{\gamma} + \Omega + \frac{\Gamma}{4}]$ attains its global maximum in $(0, \infty)$ at $\Omega = \frac{1}{2}\sqrt{\gamma\Gamma}$ corresponding to ω' in (3.47b). Evaluating this expression yields the bound in (3.47b). \square

Concerning the constants γ and Γ , we add the following comments.

Corollary 3.45. Assume (3.42a,b',c'). (a) Let the Jacobi iteration be defined by D in (3.42a): $M^{\text{Jac}} := D^{-1}(E + E^H)$. The optimal bound in (3.46a) is

$$\gamma = 1 - \rho(M^{\text{Jac}}). \quad (3.48a)$$

(b) Set $d := \rho(D^{-1}ED^{-1}E^H) = \|D^{-1/2}ED^{-1/2}\|_2^2$. Then, (3.46b) holds with

$$\Gamma = 2 + \frac{4d-1}{\gamma} \quad (\text{in particular, } \Gamma \leq 2 \text{ for } d \leq \frac{1}{4}). \quad (3.48b)$$

Proof. (a) The best bound in (3.46a) is the smallest eigenvalue of

$$D^{-1}A = I - D^{-1}(E + E^H) = I - M^{\text{Jac}}.$$

(b) Forming the products in (3.46b) and using $E + E^H = D - A$ and $D \leq \frac{1}{\gamma}A$ yield

$$\begin{aligned} \frac{1}{4}D - \frac{1}{2}(E + E^H) + ED^{-1}E^H &\leq \frac{1}{4}D - \frac{1}{2}(E + E^H) + dD \\ &= \frac{1}{4}[(4d+1)D - 2(E + E^H)] = \frac{1}{4}[(4d-1)D + 2A] \leq \frac{1}{4}\left[2 + \frac{4d-1}{\gamma}\right]A. \quad \square \end{aligned}$$

The notation M_ω^{SOR} in (3.47a,b) and M^{Jac} in (3.48a) are justified only if D is the diagonal or block diagonal of A . If D in Theorem 3.44 is another matrix, a new method is defined and the iteration matrix in (3.47a,b) should be named differently.

Conclusion 3.46 (order improvement). Assume (3.42a,b',c') and define

$$d := \rho(D^{-1}ED^{-1}E^H) \leq 1/4.$$

Let τ be the order of the Jacobi iteration: $\rho(M^{\text{Jac}}) = 1 - \gamma = 1 - C_{\text{Jac}}h^\tau + \mathcal{O}(h^{2\tau})$. In the case of the Gauss–Seidel iteration ($\omega = 1$), the bound (3.47a) has the same order:

$$\|M_1^{\text{SOR}}\|_A = \|M^{\text{GS}}\|_A \leq \sqrt{1 - \frac{4}{\Gamma+2+\frac{1}{\gamma}}} \leq \frac{1}{\sqrt{1+4\gamma}} = 1 - 2C_{\text{Jac}}h^\tau + \mathcal{O}(h^{2\tau}). \quad (3.49)$$

However, the order is improved (halved) for $\omega := \omega'$ in (3.47b):

$$\|M_{\omega'}^{\text{SOR}}\|_A \leq 1 - \sqrt{\gamma\Gamma} + \mathcal{O}(\gamma/\Gamma) = 1 - \sqrt{\frac{C}{2}}h^{\tau/2} + \mathcal{O}(h^\tau).$$

This estimate even holds (with another constant) if the condition $d \leq \frac{1}{4}$ is replaced with $d \leq 1/4 + \mathcal{O}(h^\tau)$. The size $d = \mathcal{O}(1)$ is sufficient for (3.49).

Proof. Insert the values (3.48a,b) into (3.47a,b). \square

The improvement of the order will become clearer and more transparent in §4.6.3.

The previous estimates are of the form $\rho(M_\omega^{\text{SOR}}) \leq 1 - \dots$. Concerning the characterisation of the term ‘…’, specific assumptions are required. Next, we cite a convergence result, which holds for general positive definite matrices. According to Exercise 3.20, the (block-)SOR iteration is invariant with respect to a scaling by $\Delta = D^{-1}$. The transformation $A \rightarrow D^{-1}A$ or $A \rightarrow D^{-1/2}AD^{-1/2}$ results in a new matrix A satisfying

$$D = \text{diag}\{A\} = I. \quad (3.50)$$

Following (3.50), we then have $L = E$ and $U = F$ (cf. (3.15d)). The next result is due to Oswald [305].

Theorem 3.47. *Let $A \in \mathbb{R}^{I \times I}$ be a positive definite matrix of size $n := \#I$ satisfying (3.50). Then for a suitable $\omega^* \in (0, 2)$ the estimate*

$$\rho(M_{\omega^*}^{\text{SOR}}) \leq 1 - \frac{2}{1 + \sqrt{1 + [(\text{cond}_2(A) - 1) \log_2(2n-2)]^2}} = 1 - \frac{2}{\text{cond}_2(A) \log_2 n} + \dots$$

holds, where the asymptotic statement refers to $\text{cond}_2(A) \log_2 n \rightarrow \infty$.

A discussion of the optimal choice of ω for *generalised SOR methods*, in which L and U possibly deviate from a strictly triangular form, can be found in Hanke–Neumann–Niethammer [213].

Theorem 3.48 (Niethammer [292]). *For a real matrix A , assume $A + A^\top > 0$ and (3.50). Then Λ and $\tilde{\Lambda}$ in*

$$\begin{aligned} \Lambda &:= \lambda_{\max}(\tfrac{1}{2}(L + L^\top + U + U^\top)), & \tilde{\Lambda} &:= \lambda_{\max}(\tfrac{1}{2}(L + L^\top - U - U^\top)), \\ \sigma &:= \rho(\tfrac{1}{2}(L - L^\top + U - U^\top)), & \tilde{\sigma} &:= \rho(\tfrac{1}{2}(L - L^\top - U + U^\top)) \end{aligned}$$

satisfy $0 \leq \Lambda < 1$ and $\tilde{\Lambda} \geq 0$. The SOR method converges for ω with

$$0 < \omega < 2 / \left[1 + \tilde{\Lambda} + \frac{\sigma \tilde{\sigma}}{1 - \Lambda} \right].$$

For $A > 0$ and $L = U^\top$, we obtain $\sigma = \tilde{\Lambda} = 0$. Hence the inequalities above become $0 < \omega < 2$ (cf. Theorem 3.41). If $A - I$ is skew-symmetric, i.e., $L = -U^\top$, one proves the following corollary using $\Lambda = \tilde{\sigma} = 0$ and $\tilde{\Lambda} = \rho(U - L)$.

Corollary 3.49. Assume that $A = I - L + L^\top$ (L lower triangular matrix). Then the SOR iteration converges for ω satisfying $0 < \omega < 2 / (1 + \rho(L + L^\top))$. If, in addition, L is componentwise nonnegative and $\rho(L + L^\top) < 1$, the SOR iteration diverges for all other real ω .

A similar divergence statement can also be shown for $L \neq -U^\top$ if $L - U$ is componentwise nonnegative. The optimal parameter $\omega_{\text{opt}} < 1$ yields an *under-relaxation* method.

Convergence results for complex matrices can be found in Niethammer [291].

3.5.5 Convergence of the Block Variants

Theorem 3.50. *Theorem 3.36 is also valid for the block-Jacobi iteration if D represents a block diagonal in (3.38) and (3.39).*

Proof. Not only the diagonal $D = \text{diag}\{A\}$ but also the block diagonal $D = \text{blockdiag}\{A\}$ is positive definite (cf. 3.17b); hence, the assertion follows from Remark 3.38. \square

Remark 3.37b corresponds to the following statement.

Exercise 3.51. Let A be a 2×2 block matrix, i.e., assume $\#B=2$. Prove that A and $2D - A$ with $D = \text{blockdiag}\{A\}$ have the same eigenvalues: $\sigma(A) = \sigma(2D - A)$.

From Exercise 3.51, one concludes that if A is positive definite, $2D - A$ is also. This proves the next corollary.

Corollary 3.52. The block-Jacobi iteration converges for a positive definite 2×2 block-matrix A .

In the case of the blockwise Gauss–Seidel and SOR methods, the block diagonal D of a positive definite matrix A satisfies the conditions (3.42b',c'). Therefore, the statement of Ostrowski's theorem (Theorem 3.41) also holds for the block-SOR method and covers the case of the block-Gauss–Seidel iteration for $\omega = 1$.

Theorem 3.53. *All statements in Theorem 3.39 to Conclusion 3.46 remain valid for the block-versions of the Gauss–Seidel and SOR iteration.*

The matrix

$$I_j := \text{blockdiag}\{\underbrace{I, I, \dots, I}_{j \text{ blocks}}, \underbrace{0, \dots, 0}_{\beta-j \text{ blocks}}\} \quad \text{for } 1 \leq j \leq \beta := \#B$$

is the identity matrix with respect to the first j blocks. The following convergence statement is proved by Bank–Dupont–Yserentant [30, Theorem 3.4, (3.42), (3.67)].

Theorem 3.54. *Let $A > 0$. The block-Gauss–Seidel iteration converges with the rate*

$$\rho(M_{\text{block}}^{\text{GS}}) \leq \sqrt{1 - 1/\sum_{j=1}^{\beta} \|I_j\|_A^2}.$$

3.6 Convergence Rates in the Case of the Model Problem

3.6.1 Richardson and Jacobi Iteration

The convergence rate $\rho(M_{\Theta}^{\text{Rich}}) = \max\{|1 - \Theta\lambda_{\min}|, |1 - \Theta\lambda_{\max}|\}$ of the Richardson method only depends on the extreme eigenvalues λ_{\min} and λ_{\max} of the matrix (cf. (3.23)). Inserting the values for λ_{\min} and λ_{\max} given in (3.1b,c) into (3.23)–(3.25), we obtain the following statement.

Theorem 3.55. *In the model case, the Richardson method has the rate*

$$\rho(M_{\Theta}^{\text{Rich}}) = \max \left\{ \left| 1 - 8\Theta^{-2} \sin^2 \frac{\pi h}{2} \right|, \left| 1 - 8\Theta^{-2} \cos^2 \frac{\pi h}{2} \right| \right\}.$$

Convergence holds for $0 < \Theta < h^2/[4 \cos^2(\pi h/2)] = h^2/4 + \mathcal{O}(h^4)$. The optimal convergence rate is attained for $\Theta = h^2/4$ and equals

$$\rho(M_{\Theta}^{\text{Rich}}) = 1 - 2 \sin^2(\pi h/2) = \cos(\pi h) \quad \text{for } \Theta = \Theta_{\text{opt}} = h^2/4. \quad (3.51)$$

In the Poisson model case, $D = 4h^{-2}I$ holds. Hence, the Jacobi iteration $x^{m+1} = x^m - D^{-1}(Ax^m - b)$ coincides with the Richardson iteration $x^{m+1} = x^m - \Theta(Ax^m - b)$ for $\Theta = h^2/4$. Statement (3.51) yields the next theorem.

Theorem 3.56. *In the model case, the Jacobi iteration leads to the convergence rate*

$$\rho(M^{\text{Jac}}) = 1 - 2 \sin^2(\pi h/2) = \cos(\pi h). \quad (3.52)$$

Replacing the square Ω of the model problem by a rectangle with N_x and N_y subintervals in the x and y direction, we obtain

$$\rho(M^{\text{Jac}}) = \frac{1}{2} [\cos(\pi/N_x) + \cos(\pi/N_y)]$$

instead of (3.51) and (3.52).

Remark 3.57. The convergence rate (3.52) of the Jacobi iteration has the form (2.32a): $\rho(M^{\text{Jac}}) = 1 - \eta^{\text{Jac}}$ with

$$\eta^{\text{Jac}} = 2 \sin^2(\pi h/2) = \pi^2 h^2/2 + \mathcal{O}(h^4),$$

i.e., the convergence is of order $\tau = 2$, and the constant in (2.32c) is equal to

$$C_{\eta}^{\text{Jac}} = \pi^2/2. \quad (3.53)$$

The same constant hold for Richardson's iteration with the optimal Θ_{opt} in (3.51).

3.6.2 Block-Jacobi Iteration

The eigenvector $e^{\alpha\beta}$ of A defined in (3.2) is also an eigenvector of the block-diagonal matrix D of A , where the rows of the grid form the blocks. For symmetry reasons, we obtain the same results if the blocks are defined by the columns of the grid.

Lemma 3.58. *Let A have a row-block structure and D be the corresponding block-diagonal matrix. Then $e^{\alpha\beta}$ is an eigenvector of D related to the eigenvalue $d_{\alpha\beta}$:*

$$De^{\alpha\beta} = d_{\alpha\beta}e^{\alpha\beta} \quad \text{with } d_{\alpha\beta} = h^{-2} [2 + 4 \sin^2 \frac{\alpha h \pi}{2}] \quad \text{for } 1 \leq \alpha, \beta < N.$$

Proof. For each grid point $(x, y) = (\nu h, \mu h) \in \Omega_h$, we have (cf. (3.3))

$$\begin{aligned} (De^{\alpha\beta})(x, y) &= h^{-2} 2h [4 \sin(\alpha x \pi) \sin(\beta y \pi) \\ &\quad - \sin(\alpha(x+h)\pi) \sin(\beta y \pi) - \sin(\alpha(x-h)\pi) \sin(\beta y \pi)] \\ &= h^{-2} [4 - 2 \cos(ih\pi)] e^{\alpha\beta}(x, y) = h^{-2} [2 + 4 \sin^2(ih\pi/2)] e^{\alpha\beta}(x, y). \quad \square \end{aligned}$$

The eigenvalues of $2D - A$ are $2d_{\alpha\beta} - \lambda_{\alpha\beta} = 4h^{-2} [\cos^2(\frac{\beta h \pi}{2}) + \sin^2(\frac{\alpha h \pi}{2})]$ with $\lambda_{\alpha\beta}$ in (3.1a). Their positivity proves $2D - A > 0$. Therefore, the block-Jacobi method converges (cf. Theorem 3.50). For determining the convergence speed, we have to study the eigenvalues of the iteration matrix $M = I - D^{-1}A$:

$$\sigma(M_{\text{block}}^{\text{Jac}}) = \{(d_{\alpha\beta} - \lambda_{\alpha\beta}) / d_{\alpha\beta} : 1 \leq \alpha, \beta \leq N-1\}.$$

Since $\left| \frac{d_{\alpha\beta} - \lambda_{\alpha\beta}}{d_{\alpha\beta}} \right| = \frac{|1 - 2 \sin^2(\beta h \pi/2)|}{|1 + 2 \sin^2(\alpha h \pi/2)|}$ ($1 \leq \alpha, \beta \leq N-1$), we may optimise the numerator and denominator separately. The numerator is maximal for $\beta = 1$, the denominator takes its minimum for $\alpha = 1$. This yields

$$\rho(M_{\text{block}}^{\text{Jac}}) = \frac{1 - 2 \sin^2(h\pi/2)}{1 + 2 \sin^2(h\pi/2)} = \frac{\cos(h\pi)}{1 + 2 \sin^2(h\pi/2)}. \quad (3.54)$$

In the case of the Poisson model problem, the size of rows and columns is identical and therefore row- and column-block-Jacobi iterations behave the same. The situation is different if the discretisation uses different step sizes h_x, h_y .

Exercise 3.59. Let A be the discretisation matrix of the problem in Exercise 3.4 with step sizes h_x, h_y . What are the values of $\rho(M_{\text{block}}^{\text{Jac}})$ for the row and column versions? Which iteration converges faster?

The asymptotic expansion of (3.54) in h yields

$$\rho(M_{\text{block}}^{\text{Jac}}) = 1 - 4 \sin^2(h\pi/2) + \mathcal{O}(h^4) = 1 - \pi^2 h^2 + \mathcal{O}(h^4) = 1 - \eta_{\text{block}}^{\text{Jac}}$$

with $\eta_{\text{block}}^{\text{Jac}} = \pi^2 h^2 + \mathcal{O}(h^4)$, i.e., the order of convergence is $\tau = 2$, and the constant in (2.32c) is $C_{\eta}^{\text{blockJac}} = \pi^2$.

The block version is more expensive than the pointwise Jacobi iteration. On the other hand, the blockwise iteration is faster. Using the cost factor C_{Φ} in (3.22a,b) and the quantities $C_{\eta}^{[\text{block}]\text{Jac}}$ (cf. (3.53)), we determine the coefficients C_{eff} of the effective amount of work (cf. (2.32d)):

$$\text{Eff}(\Phi^{\text{Jac}}) = \frac{2}{\pi^2} h^{-2} + \mathcal{O}(1), \quad \text{Eff}(\Phi_{\text{block}}^{\text{Jac}}) = \frac{7}{5\pi^2} h^{-2} + \mathcal{O}(1).$$

This proves the next remark.

Remark 3.60. For the Poisson model problem in §1.2, the block-Jacobi iteration is more effective by a factor of 0.7 than the pointwise Jacobi iteration.

3.6.3 Numerical Examples for the Jacobi Variants

Table 3.1 reports the results of the pointwise and blockwise Jacobi iterations. As in Table 1.1, the numbers refer to the Poisson model problem of step size $h = \frac{1}{32}$. For selected iteration numbers m , the table presents the value $u_{16,16}^m$ at the midpoint. Note that the iterates $u_{16,16}^m$ should converge to

$$u\left(\frac{1}{2}, \frac{1}{2}\right) = 0.5.$$

The table also contains the maximum norm

$$\varepsilon_m := \|e^m\|_\infty$$

of $e^m = u^m - u_h$ and the reduction factor

m	pointwise Jacobi iteration			blockwise Jacobi iteration			
	$u_{16,16}$	ε_m	$\rho_{m,m-1}$	m	$u_{16,16}$	ε_m	$\rho_{m,m-1}$
1	-0.0010	1.759		1	-0.0019	1.666	
2	-0.0019	1.644	0.93504	2	-0.0039	1.560	0.93621
3	-0.0029	1.588	0.96598	3	-0.0059	1.475	0.94605
.....
62	-0.0480	0.795	0.99321	37	-0.0449	0.734	0.98597
63	-0.0480	0.789	0.99311	38	-0.0426	0.727	0.98953
64	-0.0480	0.784	0.99313	39	-0.0429	0.715	0.98478
.....
100	-0.0230	0.629	0.99468	100	0.14077	0.374	0.98565
101	-0.0217	0.626	0.99462	101	0.14176	0.372	0.99433
102	-0.0205	0.623	0.99464	102	0.14713	0.367	0.98619
103	-0.0192	0.619	0.99458	103	0.14812	0.364	0.99376
.....
200	0.14011	0.374	0.99497	200	0.36033	0.141	0.99008
201	0.14173	0.372	0.99493	201	0.36077	0.139	0.99077
202	0.14333	0.370	0.99497	202	0.36299	0.138	0.98996
203	0.14493	0.368	0.99493	203	0.36342	0.137	0.99090
.....
297	0.27122	0.231	0.99508	297	0.44474	0.055	0.99411
298	0.27231	0.230	0.99512	298	0.44563	0.055	0.98671
299	0.27340	0.229	0.99508	299	0.44580	0.054	0.99414
300	0.27447	0.228	0.99512	300	0.44666	0.053	0.98668

Table 3.1 Jacobi iteration for $N = 32$ in the model case.

$$\rho_{m,m-1} = \varepsilon_m / \varepsilon_{m-1}.$$

We observe that $\rho_{m,m-1}$ converges to different limits for odd and even m . The explanation is that with $r := \rho(M_{\text{block}}^{\text{Jac}})$, $-r$ is also an eigenvalue of the iteration matrix (cf. Remark 4.9). Hence, the dominating error part has the form

$$r^m e_1 + (-r)^m e_2 = r^m [e_1 + (-1)^m e_2]$$

and oscillates with the period 2. The geometric mean of two successive factors approximates the spectral radius $\rho(M_{\text{block}}^{\text{Jac}})$. The mean values

$$\sqrt{\varepsilon_{300}/\varepsilon_{298}} = \begin{cases} 0.995099 & \text{for the pointwise method,} \\ 0.990401 & \text{for the blockwise method,} \end{cases}$$

are in a good agreement with the values $\rho(M^{\text{Jac}}) = \cos \frac{\pi}{32} = 0.99518$ and $\rho(M_{\text{block}}^{\text{Jac}}) = 0.990416$ that result from (3.52) and (3.54) for $h = 1/32$.

In the case of the Poisson model case, the pointwise Jacobi iteration is identical to the Richardson iteration (cf. Remark 3.8). Therefore the left part of Table 3.1 also refers to the Richardson iteration.

3.6.4 SOR and Block-SOR Iteration with Numerical Examples

For evaluating the SOR bounds in Theorem 3.44, the constants γ and Γ must be determined.

Lemma 3.61. *For the Poisson model problem, the pointwise SOR iteration with lexicographical ordering satisfies (3.46a,b) with $\gamma = 2 \sin^2(\pi h/2)$ and $\Gamma = 2$. The optimal ω' in (3.47b) is*

$$\omega' = 2 / [2 + 2 \sin(\pi h/2)] = 2 - 2\pi h + \mathcal{O}(h^4). \quad (3.55a)$$

The bounds for $\omega = 1$ and $\omega = \omega'$ are

$$\|M^{GS}\|_A \leq \sqrt{\frac{1}{1 + 8 \sin^2(\pi h/2)}} = 1 - \pi^2 h^2 + \mathcal{O}(h^4), \quad (3.55b)$$

$$\|M_{\omega'}^{\text{SOR}}\|_A \leq \frac{\cos(\pi h/2)}{1 + \sin(\pi h/2)} = 1 - \pi h/2 + \mathcal{O}(h^2). \quad (3.55c)$$

Proof. (i) γ in (3.46a) is the smallest eigenvalue of $D^{-1}A = \frac{1}{4}h^2 A$; hence, $\gamma = \frac{1}{4}h^2 \lambda_{\min} = 2 \sin^2(\pi h/2)$ (cf. (3.1b)).

(ii) For lexicographical ordering, the matrix E contains at most two entries $-h^{-2}$ per row and column; hence, $\|E\|_\infty \leq 2h^{-2}$ and $\|E^H\|_\infty \leq 2h^{-2}$ hold and imply $\rho(EE^H) \leq \|E\|_\infty \|E^H\|_\infty \leq 4h^{-4}$. The inequality

$$\begin{aligned} \left(\frac{1}{2}D - E\right)D^{-1}\left(\frac{1}{2}D - E^H\right) &= -\frac{1}{4}D - \frac{1}{2}(E + E^H) + ED^{-1}E^H \\ &= \frac{1}{4}D + \frac{1}{2}A + ED^{-1}E^H = -h^{-2}I + \frac{1}{2}A + \frac{1}{4}h^2EE^H \\ &\leq -h^{-2}I + \frac{1}{2}A + \frac{1}{4}h^24h^{-4} = \frac{1}{2}A \end{aligned}$$

shows (3.46b) with $\Gamma = 2$.

(iii) The statements (3.55a–c) follow by inserting this result into (3.47a,b). \square

The inequalities (3.55b,c) show that the order of convergence improves from $1 - \mathcal{O}(h^2)$ to $1 - \mathcal{O}(h)$. However, the bound in the right-hand side of (3.55c) is distinctly less favourable than the convergence rates $\rho(M_{\omega'}^{\text{SOR}})$; i.e., the estimate is too pessimistic. On the other side, the bound in (3.55b) and the convergence rate $\rho(M^{GS})$ coincide up to $\mathcal{O}(h^4)$. Table 3.2 contrasts the bounds (3.55b,c) with the spectral radii determined in Theorem 4.27. Since the respective optimal parameters ω' in (3.55a) and ω_{opt} in (4.28b) differ slightly, the results for both of the values are reported.

In the case of the block-SOR method, γ is the smallest eigenvalue of $D^{-1}A$ with the block diagonal D of A . Similar considerations as in §3.6.2 lead to

$$\gamma = \frac{1 - 2 \sin^2(\pi h/2)}{1 + 2 \sin^2(\pi h/2)}.$$

Lemma 3.58 shows that $d_{\alpha\beta} \geq 2h^{-2}$. This implies the inequalities

$$D \geq 2h^{-2}I$$

and

$$\|D^{-1}\|_2 = \rho(D^{-1}) \leq \frac{h^2}{2}.$$

The matrix E from

$$A = D - E - E^H$$

contains only one nonzero entry $-h^{-2}$ per row and column; hence, $\|E\|_\infty = \|E^H\|_\infty = h^{-2}$ and $\rho(EE^H) \leq \|EE^H\|_\infty \leq h^{-4}$ hold. As above, we conclude $\Gamma = 2$ from

$$\left(\frac{1}{2}D - E\right) D^{-1} \left(\frac{1}{2}D - E^H\right) = \frac{1}{4}D + \frac{1}{2}A + ED^{-1}E^H \leq \frac{1}{2}A$$

because of $ED^{-1}E^H \leq \frac{1}{2}h^2EE^H \leq \frac{1}{2}h^{-2}I \leq \frac{1}{4}D$. This proves the next lemma.

Lemma 3.62. *For the model problem, the block-SOR method with lexicographical block ordering satisfies (4.32a,b) with*

$$\Gamma = 2 \quad \text{and} \quad \gamma = [1 - 2 \sin^2(\pi h/2)] / [1 + 2 \sin^2(\pi h/2)].$$

The optimal ω' in (3.47b) is

$$\omega' = 2 / \left[1 + \sqrt{8} \sin \frac{\pi h}{2} / \sqrt{1 + 2 \sin^2 \frac{\pi h}{2}} \right] = 2 - 2\sqrt{2}\pi h + \mathcal{O}(h^4).$$

The bounds for the particular cases $\omega = 1$ and $\omega = \omega'$ are

$$\begin{aligned} \|M_{\text{block}}^{\text{GS}}\|_A &\leq 1 - 2\pi^2 h^2 + \mathcal{O}(h^4), \\ \|M_{\omega'}^{\text{blockSOR}}\|_A &\leq 1 - \frac{\pi h}{\sqrt{2}} + \mathcal{O}(h^2). \end{aligned}$$

h	1/8	1/16	1/32	1/64	1/128
bound (3.55b)	0.8756	0.9637	0.9905	0.9975996	0.9993982
$\rho(M^{\text{GS}})$	0.8536	0.9619	0.9904	0.9975924	0.9993977
ω'	1.4387	1.6722	1.8213	1.9064278	1.9520897
ω_{opt}	1.4465	1.6735	1.8215	1.9064547	1.9520932
bound (3.55c)	0.8207	0.9063	0.9521	0.9757526	0.9878028
$\rho(M_{\omega'}^{\text{SOR}})$	0.5174	0.6991	0.8293	0.9086167	0.9526634
bound	0.8207	0.9063	0.9521	0.9757527	0.9878028
$\rho(M_{\omega_{\text{opt}}}^{\text{SOR}})$	0.4465	0.6735	0.8215	0.9064547	0.9520932

Table 3.2 Contraction numbers and convergence rates in the model case.

Chapter 4

Analysis of Classical Iterations Under Special Structural Conditions

Abstract The central part of this chapter is the theorem of Young about SOR convergence. It requires ‘consistently ordered matrices’. This is a more involved structural matrix property. In Section 4.1, we consider the simpler structure of 2-cyclic matrices. Sections 4.3–4.5 investigate the Richardson, Jacobi, and Gauss–Seidel iteration in this case. Section 4.6 contains the analysis of the SOR iteration. Finally, Section 4.7 presents numerical results for the model problem.

4.1 2-Cyclic Matrices

First, we define the term ‘weakly 2-cyclic’ for matrices and for matrix pairs (A, D) , where D is the diagonal or block-diagonal part of A .

Definition 4.1. A matrix $A \in \mathbb{K}^{I \times I}$ is called *weakly 2-cyclic* (or weakly cyclic of index 2) if a block structure $\{I_1, I_2\}$ with nonempty index subsets $I_1, I_2 \subset I$ exists such that

$$a_{\alpha\beta} = 0 \quad \text{if } \alpha, \beta \in I_1 \text{ or if } \alpha, \beta \in I_2. \quad (4.1)$$

Condition (4.1) states that the diagonal blocks vanish:

$$A^{11} = 0, \quad A^{22} = 0, \quad \text{i.e.,} \quad A = \begin{bmatrix} 0 & A^{12} \\ A^{21} & 0 \end{bmatrix}$$

Often, not A but $A - D$ has the form required in (4.1). In this case, we introduce the same term for the pair (A, D) .

Definition 4.2. The pair (A, D) ($A, D \in \mathbb{K}^{I \times I}$) is called *weakly 2-cyclic* if $A - D$ is weakly 2-cyclic in the sense of Definition 4.1. An equivalent statement is that a block structure $\{I_1, I_2\}$ with nonempty index subsets $I_1, I_2 \subset I$ exists such that

$$D = \text{blockdiag}\{A^{11}, A^{22}\}. \quad (4.2)$$

Let B be the block structure $\{I_1, I_2\}$ in Definition 4.1. Denote the block-diagonal part of a matrix (with respect to B) by $\text{blockdiag}_B\{\cdot\}$. Then A is weakly 2-cyclic if and only if

$$\text{blockdiag}_B\{A\} = 0.$$

The pair (A, D) is weakly 2-cyclic if and only if

$$\text{blockdiag}_B\{A\} = D. \quad (4.2')$$

The additional term ‘weakly’ in front of ‘2-cyclic’ indicates that the ordering of the indices is irrelevant. This is different in the next definition.

Definition 4.3. A or (A, D) are called *2-cyclic* if the index set I is ordered and the matrix A or respectively the pair (A, D) is weakly 2-cyclic with respect to the blocks $I_1 = \{1, \dots, n_1\}$ and $I_2 = \{n_1 + 1, \dots, n\}$ for a suitable n_1 with $1 \leq n_1 \leq n - 1$.

The condition $1 \leq n_1 \leq n - 1$ ensures that both I_1 and I_2 be nonempty. The property ‘2-cyclic’ is different from the property ‘cyclic of index 2’ as, e.g., introduced by Varga [375, page 35]. A 2-cyclic matrix has the form

$$A = \begin{array}{|c|c|} \hline & 0 & A_1 \\ \hline A_2 & & 0 \\ \hline \end{array} \left. \begin{array}{l} \left. \begin{array}{c} 0 \\ A_1 \end{array} \right\} I_1 \\ \left. \begin{array}{c} A_2 \\ 0 \end{array} \right\} I_2 \end{array} \right. .$$

Note that, in general, $A_1 = A^{12} \in \mathbb{K}^{I_1 \times I_2}$ and $A_2 = A^{21} \in \mathbb{K}^{I_2 \times I_1}$ are not square block matrices. The pair (A, D) is 2-cyclic if

$$A = \begin{bmatrix} D_1 & A_1 \\ A_2 & D_2 \end{bmatrix}, \quad D = \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix}, \quad A - D = \begin{bmatrix} 0 & A_1 \\ A_2 & 0 \end{bmatrix}. \quad (4.3)$$

The definitions immediately imply the following remark.

- Remark 4.4.** (a) The property *2-cyclic* for a special ordering of the indices implies *weakly 2-cyclic* for any ordering or no ordering of the indices.
- (b) The property *weakly 2-cyclic* is independent of the ordering of the indices, whereas in the case of the term *2-cyclic* the indices may be permuted only inside of the respective blocks I_1 and I_2 .
- (c) Let A or (A, D) be weakly 2-cyclic. If I is not ordered, then there is an ordering of the indices, so that A or (A, D) are 2-cyclic with respect to this ordering. If I is already ordered, there is a permutation of the indices with a corresponding permutation matrix P , so that $\hat{A} = PAP^T$ or $(\hat{A}, \hat{D}) = PDP^T$ are 2-cyclic.

Examples of (weakly) 2-cyclic matrices are given below for the Poisson model problem.

Example 4.5. Let A be the matrix of the Poisson model problem in §1.2.

(a) If $D = \text{diag}\{a_{\alpha\alpha} : \alpha \in I\}$ is the (pointwise) diagonal of A , then (A, D) is weakly 2-cyclic. If, as in Figure 1.2, the chequer-board ordering is used, (A, D) is even 2-cyclic. The exact definition of the *chequer-board ordering* (also called *red-black ordering*) reads as follows:

$$\begin{aligned}I_1 &= I_{\text{black}} = \{(x, y) = (ih, jh) \in \Omega_h : i + j \text{ even}\}, \\I_2 &= I_{\text{red}} = \{(x, y) = (ih, jh) \in \Omega_h : i + j \text{ odd}\}.\end{aligned}$$

(b) Let the rows (or columns) of the grid Ω_h form the block structure B and choose D as $\text{blockdiag}\{A_{\alpha\alpha} : \alpha \in B\} = \text{blockdiag}_B\{A\}$. Then (A, D) is weakly 2-cyclic. If the rows (or columns) are ordered according to the zebra pattern mentioned in Remark 3.16c, (A, D) is even 2-cyclic. The exact definition of the *zebra-row-block structure* is

$$I_1 = I_{\text{black}} = \{(x, y) = (ih, jh) \in \Omega_h : j \text{ even}\}, \quad (4.4a)$$

$$I_2 = I_{\text{white}} = \{(x, y) = (ih, jh) \in \Omega_h : j \text{ odd}\}. \quad (4.4b)$$

In the case of the zebra-column-block structure, one has to replace ‘ j even [odd]’ in (4.4a,b) with ‘ i even [odd]’.

Proof. (i) In the case of the chequer-board ordering, A has the block structure (1.9) with diagonal submatrices $4h^{-2}I \in \mathbb{R}^{(N-1) \times (N-1)}$ in the diagonal blocks. Hence the diagonal and block-diagonal parts of A coincide: $D = 4h^{-2}I \in \mathbb{R}^{n \times n}$. Therefore (4.2) holds: (A, D) is also 2-cyclic. For the chequer-board ordering, we have $I_1 = \{1, \dots, n_1\}$ and $I_2 = \{n_1 + 1, \dots, n\}$ with $n_1 := \#I_1$ being the number of *black* grid points. For all $n > 1$ (i.e., $h < \frac{1}{2}$), $n_1 \in [1, n - 1]$ holds. According to Remark 4.4a, A is weakly 2-cyclic for an arbitrary ordering or even no ordering of the indices.

(ii) In the case of the row-block structure, the diagonal blocks of A are given by the matrices $h^{-2}T$ in (1.8): $T = \text{tridiag}\{-1, 4, -1\}$. The block structure of A illustrated in (1.8) corresponds to the lexicographical ordering of the rows. The zebra-ordering (4.4a) leads to

$$A = h^{-2} \begin{array}{|c|c|} \hline T & -I \\ T & -I -I \\ \ddots & \ddots \ddots \\ & T & -I -I \\ \hline -I -I & T & & \\ -I \ddots & & T & \\ \ddots -I & & & \ddots \\ -I & & & T \\ \hline \end{array} . \quad (4.5)$$

Replacing the row-block structure by the coarser zebra-block structure, we obtain two diagonal blocks $A^{ii} = h^{-2} \text{blockdiag}\{T, \dots, T\}$ for $i = 1, 2$, where the number of the diagonal blocks T is determined by the number of the respective ‘black’ ($i = 1$) or ‘white’ ($i = 2$) rows (cf. (4.5)). The block-diagonal parts D of A with respect to the row-block structure and to the zebra-block structure coincide. This proves (4.2): (A, D) is 2-cyclic. As in (i), we see that, independently of the ordering of the indices, (A, D) is weakly 2-cyclic. \square

The model equation (1.4a) is called a five-point formula, since the equation at the point (ih, jh) contains only the five unknowns $u_{ij}, u_{i+1,j}, u_{i-1,j}, u_{i,j+1}$, and $u_{i,j-1}$. For more general problems than the Poisson equation (1.1a), one needs other formulae, e.g., nine-point formulae. In the latter case, the equation at (ih, jh) contains the nine unknowns

$$\{u_{k\ell} : k = i - 1, i, i + 1, \ell = j - 1, j, j + 1\}.$$

Since we do not exclude vanishing matrix coefficients, the five-point formulae are a subset of the nine-point formulae.

Exercise 4.6. Prove: (a) If A represents a nine-point formula, then, in general, (A, D) with $D := \text{diag}\{A\}$ is not weakly 2-cyclic. In particular, there is no ordering of the indices for which (A, D) is 2-cyclic.

(b) If A represents a nine-point formula and D is the row- or column-block diagonal of A , then (A, D) is weakly 2-cyclic as in Example 4.5b.

Statement (b) of the exercise can be generalised as follows.

Lemma 4.7. *Let A be either a tridiagonal matrix with the diagonal D or a block-tridiagonal matrix with respect to a block structure $\{I_1, I_2, \dots, I_\beta\}$ with the block diagonal D . Then (A, D) is weakly 2-cyclic.*

Proof. It is sufficient to prove the block case. The sets $J_1 := I_1 \cup I_3 \cup \dots$ and $J_2 := I_2 \cup I_4 \cup \dots$ define a coarser block structure. It is easy to see that the block diagonal of A with respect to the block structure $\{J_1, J_2\}$ coincides with D . Hence, (4.2') implies the assertion. \square

4.2 Preparatory Lemmata

Here we study the properties of a weakly 2-cyclic matrix B . For a suitable ordering of the indices, B takes the form

$$B = \begin{bmatrix} 0 & B_1 \\ B_2 & 0 \end{bmatrix}. \quad (4.6)$$

Note that the spectral properties discussed below do not depend on the ordering.

Lemma 4.8. *The spectrum of a weakly 2-cyclic matrix B with the off-diagonal blocks $B_1 = B^{12}$ and $B_2 = B^{21}$ is given by*

$$\sigma(B) = \sqrt[±]{\sigma(B_1B_2)} \cup \sqrt[±]{\sigma(B_2B_1)}. \quad (4.7a)$$

Here, the notation $\sqrt[±]{\sigma(C)} := \{\lambda \in \mathbb{C} : \lambda^2 \in \sigma(C)\}$ is used. The spectra $\sigma(B_1B_2)$ and $\sigma(B_2B_1)$ coincide up to vanishing eigenvalues:

$$\sigma(B_1B_2) \setminus \{0\} = \sigma(B_2B_1) \setminus \{0\}. \quad (4.7b)$$

Proof. (i) Let $e = \begin{bmatrix} e^1 \\ e^2 \end{bmatrix}$ be an eigenvector of B corresponding to an eigenvalue $\lambda \in \sigma(B)$. Then the following equivalence holds:

$$Be = \lambda e \iff \begin{cases} B_1e^2 = \lambda e^1, \\ B_2e^1 = \lambda e^2. \end{cases} \quad (4.8)$$

Inserting one of the equations on the right-hand side into the other, we get

$$\lambda^2 e^1 = B_1 B_2 e^1, \quad \lambda^2 e^2 = B_2 B_1 e^2.$$

By $e \neq 0$, either $e^1 \neq 0$ or $e^2 \neq 0$ must hold and therefore $\lambda^2 \in \sigma(B_1B_2)$ or $\lambda^2 \in \sigma(B_2B_1)$. In any case, $\lambda \in \sqrt[±]{\sigma(B_1B_2)} \cup \sqrt[±]{\sigma(B_2B_1)}$ is valid. Since $\lambda \in \sigma(B)$ is arbitrary, $\sigma(B) \subset \sqrt[±]{\sigma(B_1B_2)} \cup \sqrt[±]{\sigma(B_2B_1)}$ is proved.

(ii) Assume that $0 \neq \lambda \in \sqrt[±]{\sigma(B_1B_2)}$, i.e., $0 \neq \lambda^2 \in \sigma(B_1B_2)$. Let $e^1 \neq 0$ be the corresponding eigenvector: $\lambda^2 e^1 = B_1 B_2 e^1$. Set $e^2 := \frac{1}{\lambda} B_2 e^1$. We observe that

$$B_1 e^2 = \frac{1}{\lambda} B_1 B_2 e^1 = \frac{1}{\lambda} \lambda^2 e^1 = \lambda e^1.$$

By definition of e^2 , $B_2 e^1 = \lambda e^2$ holds. Hence, $e = \begin{bmatrix} e^1 \\ e^2 \end{bmatrix}$ satisfies the equations (4.8), i.e., $\lambda \in \sigma(B)$.

(iii) If $0 = \lambda^2 \in \sigma(B_1B_2) \cup \sigma(B_2B_1)$, one of the matrices B_1, B_2 has a nontrivial kernel. Without loss of generality, this might be B_1 : $B_1 e^2 = 0$ for some $e^2 \neq 0$. Since $e^1 := 0$ leads to $B_2 e^1 = 0$, $e = \begin{bmatrix} e^1 \\ e^2 \end{bmatrix}$ is the eigenvector corresponding to the eigenvalue $0 = \lambda \in \sigma(B)$.

(iv) The parts (ii) and (iii) prove $\sigma(B) \supset \sqrt[±]{\sigma(B_1B_2)} \cup \sqrt[±]{\sigma(B_2B_1)}$. Together with (i), we obtain the assertion (4.7a). (4.7b) follows from Theorem A.10. \square

The definition of $\sqrt[±]{\sigma(C)}$ leads to the next remark.

Remark 4.9. If λ is an eigenvalue of a weakly 2-cyclic matrix, then $-\lambda$ is also an eigenvalue.

Lemma 4.10. *Under the assumptions of Lemma 4.8, the following identity holds for the spectral radii:*

$$\rho(B) = \sqrt{\rho(B_1B_2)} = \sqrt{\rho(B_2B_1)}. \quad (4.9)$$

Proof. By Lemma A.20, $\rho(B_1B_2) = \rho(B_2B_1)$ holds. Together with (4.7a), we arrive at the assertion. \square

Remark 4.11. In the Hermitian case $B = B^H$, the blocks in Lemma 4.8 satisfy $B_1 = B_2^H$. By Theorem B.25, $\rho(B_1 B_2) = \rho(B_2 B_1) = \rho(B_1^H B_1) = \rho(B_2^H B_2)$ coincides with $\|B_1\|_2^2 = \|B_2\|_2^2$, so that

$$\rho(B) = \|B_1\|_2 = \|B_2\|_2.$$

Exercise 4.12. For a general matrix of the form (4.6), prove that

$$\|B\|_2 = \max \{\|B_1\|_2, \|B_2\|_2\}.$$

4.3 Analysis of the Richardson Iteration

First, we study the case of the parameter $\Theta = 1$. Furthermore, we assume that the block-diagonal part of A is the identity matrix I . For a suitable ordering of the indices, A takes the form

$$A = \begin{bmatrix} I & A_1 \\ A_2 & I \end{bmatrix}. \quad (4.10)$$

Theorem 4.13. Let (A, I) be weakly 2-cyclic with the off-diagonal blocks

$$A_1 = A^{12}, \quad A_2 = A^{21} \quad (\text{cf. (4.10)}).$$

Then the Richardson iteration $x^{m+1} = x^m - \Theta(Ax^m - b)$ with $\Theta = 1$ has the convergence rate

$$\rho(M_1^{\text{Rich}}) = \sqrt{\rho(A_1 A_2)} = \sqrt{\rho(A_2 A_1)}. \quad (4.11)$$

Proof. The iteration matrix $M_1^{\text{Rich}} = I - A$ coincides with the matrix B in §4.2 if $B_i := -A_i$. Hence, (4.9) in Lemma 4.10 implies the result (4.11). \square

For $\Theta \neq 1$, $M_\Theta^{\text{Rich}} = I - \Theta A$ leads to

$$\sigma(M_\Theta^{\text{Rich}}) = \{\lambda = 1 - \Theta(1 - \mu) : \mu \in \sigma(B)\} \quad \text{with } B = I - A \quad (4.12)$$

and $\sigma(B)$ as in (4.7a), where $B_1 := -A_1$ and $B_2 := -A_2$. For an arbitrary complex spectrum $\sigma(B)$, a simple characterisation of the spectral radius M_Θ^{Rich} is not so easy. Therefore, we assume

$$\beta := \rho(B) \in \sigma(B) \quad \text{for } B = I - A = - \begin{bmatrix} 0 & A_1 \\ A_2 & 0 \end{bmatrix}. \quad (4.13)$$

Condition (4.13) states that $\rho(B)$ is not only the absolute value $|\lambda|$ of some eigenvalue $\lambda \in \sigma(B)$ but even an eigenvalue of B . Sufficient conditions for (4.13) will follow after Theorem 4.14.

Theorem 4.14. Let (A, I) be weakly 2-cyclic satisfying (4.13). Then the Richardson iteration $x^{m+1} = x^m - \Theta(Ax^m - b)$ has the convergence rate

$$\rho(M_\Theta^{\text{Rich}}) = \begin{cases} 1 - \Theta(1 - \rho(B)) & \text{for } 0 \leq \Theta \leq 1, \\ \Theta(1 + \rho(B)) - 1 & \text{for } \Theta \geq 1, \\ 1 + |\Theta|(1 + \rho(B)) & \text{for } \Theta \leq 0 \end{cases} \quad (4.14)$$

with $\rho(B) = \sqrt{\rho(A_1 A_2)} = \sqrt{\rho(A_2 A_1)}$.

If $\rho(B) \geq 1$, the iteration is divergent for all $\Theta \in \mathbb{R}$. If $\rho(B) < 1$, the iteration converges for

$$0 < \Theta < \frac{2}{1 + \rho(B)}.$$

$\Theta = 1$ yields the optimal convergence rate (4.11).

Proof. (i) Let $\Theta \in [0, 1]$, $\mu \in \sigma(B)$, and $\beta := \rho(B)$. According to (4.12), we have to estimate $\lambda = 1 - \Theta(1 - \mu)$. Since $|\mu| \leq \beta$ by assumption (4.13),

$$\begin{aligned} |\lambda| &= |1 - \Theta(1 - \mu)| \\ &= |(1 - \Theta) + \Theta\mu| \leq 1 - \Theta + \Theta|\mu| \\ &\leq 1 - \Theta + \Theta\beta = 1 - \Theta(1 - \beta) \end{aligned} \quad (4.15)$$

holds. For $\mu = \beta \in \sigma(B)$, the equal sign holds in (4.15), so that $1 - \Theta(1 - \beta)$ is the smallest bound for $\rho(M_\Theta^{\text{Rich}})$. Hence, the first case in (4.14) is proved.

(ii) The cases $\Theta \geq 1$ and $\Theta \leq 0$ in (4.14) are treated analogously.

(iii) The further statements are a direct consequence of (4.14). \square

Condition (4.13) required in Theorem 4.14 is the subject of the following two criteria.

Criterion 4.15. If B has only real eigenvalues, condition (4.13) is satisfied. In particular, symmetry of B is sufficient: $A_1 = A_2^H$.

Proof. Let β_{\min} and β_{\max} be the minimal and maximal eigenvalues of B . Since, by Remark 4.9, the spectrum $\sigma(B)$ is symmetric,

$$\beta_{\min} = -\beta_{\max} \leq 0 \leq \beta_{\max}$$

proves $\rho(B) = \max\{|\beta_{\min}|, |\beta_{\max}|\} = \beta_{\max} \in \sigma(B)$. \square

Criterion 4.16. If all matrix entries of A_1 and A_2 are nonnegative (or all nonpositive), condition (4.13) is satisfied.

Proof. Theorem C.34 shows that $\beta := \rho(-B) \in \sigma(-B)$. Since $\sigma(-B) = \sigma(B)$ (cf. Remark 4.9), $\beta = \rho(B) \in \sigma(B)$ is also valid. \square

4.4 Analysis of the Jacobi Iteration

In the following, D is assumed to be the (pointwise) diagonal or block-diagonal part of A . Let (A, D) be weakly 2-cyclic with respect to a block structure $\{I_1, I_2\}$ with off-diagonal blocks $A_1 = A^{12}$ and $A_2 = A^{21}$. For a suitable ordering, A and D take the form

$$A = \begin{bmatrix} D_1 & A_1 \\ A_2 & D_2 \end{bmatrix} \quad \text{and} \quad D = \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix}. \quad (4.16)$$

The matrix $A' := D^{-1}A$ has the representation

$$D^{-1}A = \begin{bmatrix} I & A'_1 \\ A'_2 & I \end{bmatrix} = I - B \quad \text{with} \quad B = -\begin{bmatrix} 0 & A'_1 \\ A'_2 & 0 \end{bmatrix}, \quad (4.17)$$

where $A'_1 := D_1^{-1}A_1$, $A'_2 := D_2^{-1}A_2$.

Depending on whether D is the (pointwise) diagonal or block diagonal of A , the following theorem describes the pointwise or blockwise Jacobi method.

Theorem 4.17. *Let (A, D) be weakly 2-cyclic. Then the Jacobi method $x^{m+1} = x^m - D^{-1}(Ax^m - b)$ has the convergence rate*

$$\rho(M^{\text{Jac}}) = \rho(B) = \sqrt{\rho(A'_1 A'_2)} = \sqrt{\rho(D_1^{-1} A_1 D_2^{-1} A_2)}. \quad (4.18)$$

Proof. The Jacobi method is identical to the Richardson iteration applied to

$$A'x = b' \quad \text{with} \quad A' := D^{-1}A, \quad b' := D^{-1}b, \quad \text{and} \quad \Theta = 1$$

(cf. Proposition 5.44). Hence, (4.18) follows from Theorem 4.13. \square

To obtain statements about the damped Jacobi method, one has to satisfy condition (4.13) for B defined in (4.17). Besides Criteria 4.15 and 4.16, the following one can be applied.

Criterion 4.18. *Let (A, D) be weakly 2-cyclic. (a) If D is positive definite and A is Hermitian, then $B = M^{\text{Jac}} = I - D^{-1}A$ has only real eigenvalues with*

$$\rho(B) \in \sigma(B). \quad (4.19)$$

(b) If A is positive definite, $\rho(B) < 1$ is also valid, i.e., the Jacobi iteration converges.

Proof. (a) Since $\hat{B} = D^{-1/2}(D - A)D^{-1/2}$ is Hermitian, it has only real eigenvalues. Because \hat{B} is similar to $B = D^{-1}(D - A)$, $\sigma(\hat{B}) = \sigma(B)$ holds. Therefore, all eigenvalues of B are real and Criterion 4.15 is applicable.

(b) By Corollary 3.52, the Jacobi iteration converges. Since $B = M^{\text{Jac}}$, $\rho(B) < 1$ follows. \square

Criterion 4.16 yields the following sufficient condition for (4.19). If D^{-1} , A_1 , D_2^{-1} , A_2 have only nonnegative matrix entries, (4.19) holds.

Conclusion 4.19. Assume that (A, D) is weakly 2-cyclic and satisfies (4.19). If $\rho(B) \geq 1$, the damped Jacobi method $x^{m+1} = x^m - \vartheta D^{-1}(Ax^m - b)$ diverges for all $\vartheta \in \mathbb{R}$. If $\rho(B) < 1$, it converges for

$$0 < \vartheta < 2/(1 + \rho(B));$$

and the optimal convergence rate is attained for $\vartheta = 1$, i.e., the undamped Jacobi method is already optimal.

Proof. Apply Theorem 4.14. \square

4.5 Analysis of the Gauss–Seidel Iteration

Let the index set I be ordered and assume that the matrix A has the form:

$$A = \begin{bmatrix} D_1 & A_1 \\ A_2 & D_2 \end{bmatrix} \quad (\text{cf. (4.16)})$$

with respect to the block structure $\{I_1, I_2\}$. A allows the splitting $A = D - E - F$ with

$$D = \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix}, \quad E = \begin{bmatrix} 0 & 0 \\ -A_2 & 0 \end{bmatrix}, \quad F = \begin{bmatrix} 0 & -A_1 \\ 0 & 0 \end{bmatrix} \quad (4.20)$$

(cf. (3.19a–d)). Furthermore, (A, D) is 2-cyclic. The pointwise or blockwise diagonal D of A leads to the pointwise or blockwise Gauss–Seidel iteration. The Gauss–Seidel iteration matrix is $M^{\text{GS}} = (D - E)^{-1}F$. By

$$(D - E)^{-1} = \begin{bmatrix} D_1 & 0 \\ A_2 & D_2 \end{bmatrix}^{-1} = \begin{bmatrix} D_1^{-1} & 0 \\ -D_2^{-1}A_2D_1^{-1} & D_2^{-1} \end{bmatrix},$$

we obtain

$$M^{\text{GS}} = (D - E)^{-1}F = \begin{bmatrix} 0 & -D_1^{-1}A_1 \\ 0 & D_2^{-1}A_2D_1^{-1}A_1 \end{bmatrix}. \quad (4.21)$$

Theorem 4.20. The Gauss–Seidel iteration $x^{m+1} = (D - E)^{-1}(Fx^m + b)$ defined according to (4.20) by the matrices in $A = D - E - F$ has the convergence rate

$$\rho(M^{\text{GS}}) = \rho(D_1^{-1}A_1 D_2^{-1}A_2). \quad (4.22)$$

Proof. Inspection of (4.21) yields $\rho(M^{\text{GS}}) = \rho(D_2^{-1}A_2D_1^{-1}A_1)$ (cf. (A.10)). By Lemma A.20, $\rho(D_2^{-1}A_2D_1^{-1}A_1) = \rho(D_1^{-1}A_1D_2^{-1}A_2)$ holds. \square

Conclusion 4.21. Let (A, D) be 2-cyclic. (a) Then the Jacobi method converges if and only if the Gauss–Seidel method converges.

(b) The convergence of the Gauss–Seidel iteration is exactly twice as fast as that of the Jacobi iteration:

$$\rho(M^{\text{GS}}) = \rho(M^{\text{Jac}})^2. \quad (4.23)$$

(c) M^{GS} has at least an n_1 -fold eigenvalue $\lambda = 0$ where $n_1 \times n_1$ is the size of the first block D_1 of A .

Proof. A comparison of (4.22) with (4.18) yields Eq. (4.23), proving the parts (a) and (b). Part (c) follows because of the zero blocks in (4.21). \square

Since the Jacobi and Gauss–Seidel iterations require the same amount of computational work, we conclude from (4.23) the following remark.

Conclusion 4.22. Let (A, D) be 2-cyclic. For the Jacobi method, the effective amount of work defined in (2.31a) is twice as large as for the Gauss–Seidel method:

$$\text{Eff}(\varPhi^{\text{Jac}}) = 2\text{Eff}(\varPhi^{\text{GS}}),$$

but the order of linear convergence defined in §2.3.3 coincides for both methods.

The results of Conclusions 4.21 and 4.22 will again be confirmed in §4.6 under somewhat weaker assumptions.

4.6 Analysis of the SOR Iteration

The analysis of SOR has started in §3.5.4 and is continued below for consistently ordered matrices. The closely related symmetric SOR iteration (SSOR) will be investigated in §5.4.3 and §6.3.

4.6.1 Consistently Ordered Matrices

In §4.5, (A, D) was assumed to be 2-cyclic. This assumption is close to Young’s ‘property A’ (cf. Young [412]). ‘Property A’ can be generalised by the notion of the ‘consistent ordering’ due to Varga [375].

Definition 4.23. Let the index set I be ordered. Split A into $A = D - E - F$ according to (3.11a–d) or (3.19a–d). Consequently, the matrices $L := D^{-1}E$ and $U := D^{-1}F$ are strictly triangular matrices. A is called *consistently ordered* if the eigenvalues of the matrix $zL + \frac{1}{z}U$ do not depend on $z \in \mathbb{C} \setminus \{0\}$.

In the following, we avoid the term ‘consistent ordering of A ’ for two reasons. First, it is somewhat inexact, since L and U depend not only on A but also on D if matrices D of diagonal as well as block-diagonal form are admitted. Hence, the consistent ordering is a property of A and the block structure B . Second, contrary to its name, the addressed property is independent of the ordering of the indices if we admit other L and U than triangular matrices.

Instead, we require the pair (L, U) to satisfy the following property:

$$\text{The eigenvalues of } zL + \frac{1}{z}U \text{ do not depend on } z \in \mathbb{C} \setminus \{0\}. \quad (4.24)$$

Criterion 4.24. (L, U) satisfies (4.24) and, moreover, A is consistently ordered in the sense of Definition 4.23 if L and U are strictly lower and upper triangular matrices, respectively, satisfying one of the following four conditions:

$$L + U \text{ is 2-cyclic}, \quad (4.25a)$$

$$L + U \text{ is tridiagonal}, \quad (4.25b)$$

$$L + U \text{ is block-tridiagonal with vanishing diagonal blocks} \quad (4.25c)$$

$$L + U \text{ is block-tridiagonal, where the diagonal blocks } (L + U)^{ii} \quad (4.25d)$$

are tridiagonal and the (possibly rectangular) off-diagonal blocks

$(L + U)^{i,i \pm 1}$ are diagonal.

Here, a rectangular matrix A is called diagonal if at most the diagonal entries A_{ii} are different from zero.

Proof. (i) Set $B := L + U$. Since L and U are strictly triangular matrices, $b_{ii} = 0$ holds for the diagonal entries.

(ii) Using Lemma 4.8, we conclude the assertion from (4.25a). Furthermore, (4.25a) is a special case of (4.25c).

(iii) Assume (4.25b). Let $z \in \mathbb{C} \setminus \{0\}$ and construct the diagonal matrix

$$\Delta := \text{diag}\{1, z, z^2, \dots, z^{n-1}\}.$$

$B' := \Delta B \Delta^{-1}$ has the entries $b'_{ij} = b_{ij} z^{i-j}$. Since $b'_{ij} \neq 0$ only for $|i - j| = 1$, B' is of the form $zL + \frac{1}{z}U$. Because B and B' are similar, (L, U) satisfies property (4.24).

(iv) Let $\{I_1, \dots, I_b\}$ be the block structure in the case (4.25c). We define

$$\Delta_b := \text{blockdiag}\{I^0, zI^1, z^2I^2, \dots, z^{b-1}I^{b-1}\},$$

where I^k is the identity matrix of the respective block-size $I^k \in \mathbb{R}^{I_k \times I_k}$. The transformed matrix $B' := \Delta_b B \Delta_b^{-1}$ contains the blocks $(B')^{ij} = z^{i-j} B^{ij}$, so that the assertion follows as in (iii).

(v) In the case of (4.25d), apply first the similarity transformation Δ_b from (iv): B' contains the factor z (or $\frac{1}{z}$) in the blocks below (or above) the diagonal; however, the diagonal blocks $(B')^{ii} = B^{ii}$ of tridiagonal structure are still unchanged. We define the diagonal matrix

$$\Delta_B := \text{blockdiag}\{(\Delta_B)^{ii} : i = 1, \dots, b\}, \quad (\Delta_B)^{ii} := \text{diag}\{1, z, \dots, z^{\#I_i-1}\}.$$

The transformation $C' := \Delta_B C \Delta_B^{-1}$ does not change blocks of diagonal shape. Hence,

$$B'' := \Delta_B B' \Delta_B^{-1} = \Delta_B \Delta_b B \Delta_b^{-1} \Delta_B^{-1}$$

has the off-diagonal blocks

$$(B'')^{i,i-1} = z B^{i,i-1} \quad \text{and} \quad (B'')^{i,i+1} = \frac{1}{z} B^{i,i+1},$$

whereas, as in (iii), the diagonal blocks have the entries zb_{ij} in the lower and $\frac{1}{z}b_{ij}$ in the upper triangular part, i.e., $B'' = zL + \frac{1}{z}U$. Since B'' is similar to B , the assertion follows. \square

Remark 4.25. The properties (4.25a–d) imply that (A, D) is weakly 2-cyclic if D is the (pointwise) diagonal as in the cases (4.25b,d) and, otherwise, the block diagonal of A .

Proof. For (4.25a–c), apply Remark 4.4a or Lemma 4.7a,b, respectively. In the case of (4.25d), $L + U$ has a five-point structure admitting a chequer-board-like ordering, for which $A - D = -L - U$ is 2-cyclic. \square

Lemma 4.26. Let (L, U) satisfy condition (4.24). (a) Then

$$\sigma(\alpha L + \beta U) = \sigma(\sqrt[±]{\alpha\beta}(L + U)) \quad \text{for all } \alpha, \beta \in \mathbb{C}.$$

(b) In particular, $L + U$ and $-(L + U)$ have identical spectra.

(c) If all eigenvalues of $L + U$ are real, $\rho(L + U) \in \sigma(L + U)$ holds.

Proof. (i) For $\alpha\beta \neq 0$, choose $z := \sqrt[±]{\alpha/\beta}$. Since $zL + \frac{1}{z}U$ and $L + U$ have the same eigenvalues, this is also true for the matrices

$$\alpha L + \beta U = \sqrt[±]{\alpha\beta}(zL + \frac{1}{z}U) \quad \text{and} \quad \sqrt[±]{\alpha\beta}(L + U).$$

(ii) Since $\alpha = \beta = 0$ represents a trivial case, assume that $\alpha = 0$ and $\beta \neq 0$. Then the assertion becomes $\sigma(\beta U) = \sigma(0(L + U)) = \sigma(0) = \{0\}$ and is satisfied because U is a strictly triangular matrix (cf. Exercise A.19b). The case $\beta = 0$ is analogous.

(iii) For $\alpha = \beta = 1$, the expression $\sqrt[±]{\alpha\beta}$ also takes the value -1 , so that $\sigma(L + U) = \sigma(-(L + U))$ proves part (b). Part (c) is demonstrated as in Criterion 4.15. \square

4.6.2 Theorem of Young

The following theorem, which in its basic form is due to Young [411], describes the pointwise as well as blockwise SOR iteration (3.15a):

$$x^{m+1} = M_{\omega}^{\text{SOR}} x^m + N_{\omega}^{\text{SOR}} b \quad \text{with} \quad (4.26a)$$

$$A = D - E - F, \quad L := D^{-1}E, \quad U := D^{-1}F, \quad (4.26b)$$

$$M^{\text{SOR}} = (I - \omega L)^{-1} \{(1 - \omega)I + \omega U\}, \quad N_{\omega}^{\text{SOR}} = \omega(I - \omega L)^{-1}D^{-1}, \quad (4.26c)$$

where $A = D - E - F$ is split according to (3.11a–d) or (3.19a–d). The theorem is valid for any iteration of the form (4.26a–c); i.e., D may be different from the diagonal part and E, F different from the triangular parts of A . The matrix

$$M^{\text{Jac}} := L + U = I - D^{-1}A$$

represents the iteration matrix of the (pointwise or blockwise) Jacobi method, provided that D coincides with the diagonal or block diagonal of A .

Theorem 4.27. *For the iteration (4.26a–c) we assume:*

$$0 < \omega < 2, \quad (4.27a)$$

$$M^{\text{Jac}} \text{ has only real eigenvalues,} \quad (4.27b)$$

$$\beta := \rho(M^{\text{Jac}}) < 1, \quad (4.27c)$$

$$D \text{ and } I - \omega L \text{ are regular, } (L, U) \text{ satisfies condition (4.24).} \quad (4.27d)$$

Then the following statements hold: (a) Iteration (4.26a–c) converges.

(b) The convergence rate is equal to

$$\rho(M_{\omega}^{\text{SOR}}) = \begin{cases} 1 - \omega + \frac{1}{2}\omega^2\beta^2 + \omega\beta\sqrt{1 - \omega + \frac{\omega^2\beta^2}{4}} & \text{if } 0 < \omega \leq \omega_{\text{opt}}, \\ \omega - 1 & \text{if } \omega_{\text{opt}} \leq \omega < 2, \end{cases} \quad (4.28a)$$

$$\text{where } \omega_{\text{opt}} := \frac{2}{1 + \sqrt{1 - \beta^2}}. \quad (4.28b)$$

(c) The convergence rate $\rho(M_{\omega}^{\text{SOR}})$ is minimal for $\omega = \omega_{\text{opt}}$.

(d) For $\omega \leq \omega_{\text{opt}}$, the spectral radius $\rho(M_{\omega}^{\text{SOR}}) \in \sigma(M_{\omega}^{\text{SOR}})$ is an eigenvalue.

(e) For $\omega \geq \omega_{\text{opt}}$, all eigenvalues $\lambda \in \sigma(M_{\omega}^{\text{SOR}})$ satisfy $|\lambda| = \omega - 1$.

Before proving the theorem, we discuss its assumptions and results.

Concerning (4.27a). The assumption $0 < \omega < 2$ is necessary for convergence as we know from Lemma 3.40.

Concerning (4.27b). Because of Criterion 4.18a, M^{Jac} has the required real eigenvalues if A is Hermitian and D is positive definite. Criterion 4.18b and Corollary 3.42 even provide the following sufficient criterion for (4.27b–d).

Criterion 4.28. Let D be the (block) diagonal of A . If A is positive definite, conditions (4.27b,c) and the first part of (4.27d) are satisfied.

Concerning (4.27c). $\beta < 1$ is equivalent to the convergence of the Jacobi method. The condition $\beta < 1$ is necessary because of the next statement.

Exercise 4.29. If $\beta \geq 1$, the SOR iteration diverges for all $\omega \in \mathbb{R}$.

Concerning (4.27d). If (4.26a–c) represents a true SOR iteration, L must be a strictly triangular matrix. Then the regularity of $I - \omega L$ is trivial.

Concerning (4.28a,b): Except for the trivial case $\beta = 0$, we have

$$1 < \omega_{\text{opt}} < 2,$$

so that the optimal convergence speed always leads to a true overrelaxation method. Underrelaxation ($0 < \omega < 1$) is always slower than the Gauss–Seidel iteration, which is regained for $\omega = 1$. For $\omega = 1$, we again obtain the result (4.23). Figure 4.1 shows $\rho(M_\omega^{\text{SOR}})$ as a function of ω for $\beta = \cos(\pi/8) = 0.92388$ with $\omega_{\text{opt}} = 2/(1 + \sin(\pi/8)) = 1.44646$. The latter value corresponds to the model problem with $h = 1/8$.

Proof of Theorem 4.27. (i) Let $\lambda \in \sigma(M_\omega^{\text{SOR}})$. The corresponding eigenvector e satisfies $\{(1 - \omega)I + \omega U\}e = \lambda(I - \omega L)e$, i.e.,

$$(\omega U + \lambda \omega L)e = (\lambda + \omega - 1)e,$$

so that $\lambda + \omega - 1 \in \sigma(\omega U + \lambda \omega L)$. Since $\sigma(\omega U + \lambda \omega L) = \sigma(\sqrt[+]{\lambda} \omega(L + U))$ holds by Lemma 4.26a, there is an eigenvalue $\mu \in \sigma(M^{\text{Jac}}) = \sigma(L + U)$ with

$$\lambda + \omega - 1 = \sqrt[+]{\lambda} \omega \mu, \quad \text{i.e.,} \tag{4.29a}$$

$$(\lambda + \omega - 1)^2 = \omega^2 \lambda \mu^2. \tag{4.29b}$$

Since, for any eigenvalue μ of $M^{\text{Jac}} = L + U$, $-\mu$ is also an eigenvalue (cf. Lemma 4.26b), the two solutions of (4.29b) belong to $\sigma(M^{\text{Jac}})$. Vice versa, we conclude that for any $\mu \in \sigma(M^{\text{Jac}})$ both solutions

$$\lambda = 1 - \omega + \frac{1}{2}\omega^2 \mu^2 \pm \omega \mu \sqrt{1 - \omega + \frac{1}{4}\omega^2 \mu^2} \tag{4.29c}$$

fulfil Eq. (4.29a) with suitable sign in $\sqrt[+]{\lambda}$. Since $\sqrt[+]{\lambda} \omega \mu$ is an eigenvalue of $\sqrt[+]{\lambda} \omega(L + U)$, it is also an eigenvalue of $\omega U + \lambda \omega L$; hence, we arrive at $\lambda \in \sigma(M_\omega^{\text{SOR}})$ and obtain

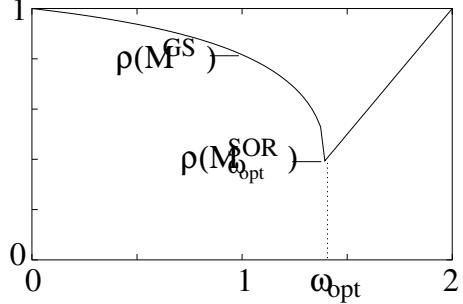


Fig. 4.1 The convergence rate $\rho(M_\omega^{\text{SOR}})$ as a function of ω .

$$\lambda \in \sigma(M_{\omega}^{\text{SOR}}) \iff \mu \in \sigma(M^{\text{Jac}}) \quad (\lambda, \mu \text{ satisfy (4.29b)}). \quad (4.29d)$$

(ii) Let $\omega_{\text{opt}} \leq \omega < 2$. This inequality is equivalent to $1 - \omega + \frac{1}{4}\omega^2\beta^2 \leq 0$. By $-\beta \leq \mu \leq \beta$, we obtain the inequality

$$1 - \omega + \frac{1}{4}\omega^2\mu^2 \leq 0$$

for all $\mu \in \sigma(M^{\text{Jac}})$, implying that Eq. (4.29b) has two complex conjugate roots:

$$\lambda = \lambda_{\Re} - i\lambda_{\Im}, \quad \lambda_{\Re} = 1 - \omega + \frac{1}{2}\omega^2\mu^2.$$

Since the product of the roots of a quadratic equation coincides with the absolute term of the equation, we obtain

$$|\lambda|^2 = (\omega - 1)^2, \quad \text{i.e., } |\lambda| = |\omega - 1|.$$

Hence, M_{ω}^{SOR} has only eigenvalues λ with an absolute value $|\omega - 1|$. This proves the second case in (4.28a), as well as the statements (a) and (e).

(iii) Assume the second case $0 < \omega < \omega_{\text{opt}}$. If $\omega \in (1, \omega_{\text{opt}})$, there may be eigenvalues $\mu \in \sigma(M^{\text{Jac}})$ with $\mu^2 < 4(\omega - 1)/\omega^2$, for which the radicand in (4.29c) is negative. As before, these μ generate eigenvalues $\lambda \in \sigma(M_{\omega}^{\text{SOR}})$ with $|\lambda| = |\omega - 1|$. This value, however, is smaller than the right-hand side in (4.28a). The latter proves to be an eigenvalue of M_{ω}^{SOR} by choosing $\mu := \beta \in \sigma(M^{\text{Jac}})$ in (4.29d) [concerning $\beta \in \sigma(M^{\text{Jac}})$ compare with Lemma 4.26c]. Since the discussion can be reduced to the case of the real solutions of (4.29c), it is easy to see that $|\lambda|$ attains its maximum at $\mu = \beta$. \square

Because of $\omega_{\text{opt}} > 1$ (cf. (4.28b)), $\omega = 1$ lies in the interval $(0, \omega_{\text{opt}}]$. Theorem 4.27 yields the following results for the Gauss–Seidel method which is the special case of $\omega = 1$.

Conclusion 4.30 (Gauss–Seidel iteration). *Under the assumptions (4.27b–d), the (block-)Gauss–Seidel iteration converges and has exactly the squared convergence speed of the (block-)Jacobi iteration:*

$$\rho(M^{[\text{block}]GS}) = \rho(M_1^{[\text{block}]SOR}) = \beta^2 = \rho(M^{[\text{block}]Jac})^2,$$

as already mentioned in (4.23) for the 2-cyclic case. Furthermore, $\rho(M^{[\text{block}]GS})$ belongs to the spectrum $\sigma(M^{[\text{block}]GS})$. The statement of Remark 4.22 is still valid.

For the case of complex relaxation parameter ω with $|\omega - 1| < 1$, a convergence result is given by Niethammer–Varga [294, Theorem 12]. Complex parameters ω make sense if the matrix $M^{[\text{block}]Jac}$ is nonsymmetric and has complex eigenvalues. For this case, we refer to Young–Huang [414].

4.6.3 Order Improvement by SOR

We recall the term ‘order of an iterative method’ as defined in §2.3.3. Considering a family of systems corresponding to different step sizes h (and therefore to different dimensions), the Jacobi iteration has the order τ if

$$\rho(M^{\text{Jac}}) = 1 - C_\eta^{\text{Jac}} h^\tau + \mathcal{O}(h^{2\tau}) \quad \text{for } h \rightarrow 0. \quad (4.30)$$

The Poisson model problem in §1.2 leads to the order $\tau = 2$. A comparison of Jacobi versus Gauss–Seidel using (4.23): $\rho(M^{\text{GS}}) = \rho(M^{\text{Jac}})^2$, shows that

$$\rho(M^{\text{Jac}})^2 = (1 - C_\eta^{\text{Jac}} h^\tau + \dots)^2 = 1 - 2C_\eta^{\text{Jac}} h^\tau + \dots = 1 - C_\eta^{\text{GS}} h^\tau + \dots.$$

Hence, only the coefficient $C_\eta^{\text{GS}} = 2C_\eta^{\text{Jac}}$ is improved, whereas the order remains unchanged.

In the weakly 2-cyclic case, the variation of the parameter ω of the damped (extrapolated) Jacobi method (5.9) is without success. By Conclusion 4.19, the choice $\omega = 1$ and therefore the standard Jacobi method are optimal. The more notable is the possibility of improving the SOR convergence rate by the proper choice $\omega = \omega_{\text{opt}}$. The next theorem shows that in this way the order is improved (halved).

Theorem 4.31. *Let $\tau > 0$ be the order of the Jacobi method (cf. (4.30)). Under the assumptions of Theorem 4.27, the SOR method with $\omega = \omega_{\text{opt}}$ has the order $\frac{\tau}{2}$:*

$$\rho(M_{\omega_{\text{opt}}}^{\text{SOR}}) = 1 - C_\eta^{\text{SOR}} h^{\tau/2} + \mathcal{O}(h^\tau) \quad \text{with} \quad (4.31a)$$

$$C_\eta^{\text{SOR}} = 2\sqrt{2C_\eta^{\text{Jac}}}. \quad (4.31b)$$

Proof. Following (4.28b), we have

$$\rho(M_{\omega_{\text{opt}}}^{\text{SOR}}) = \omega_{\text{opt}} - 1 = \frac{2}{1 + \sqrt{1 - \beta^2}} - 1 = \frac{1 - \sqrt{1 - \beta^2}}{1 + \sqrt{1 - \beta^2}}. \quad (4.31c)$$

By

$$1 - \beta^2 = 1 - \rho(M^{\text{Jac}})^2 = 1 - [1 - C_\eta^{\text{Jac}} h^\tau + \mathcal{O}(h^{2\tau})]^2 = 2C_\eta^{\text{Jac}} h^\tau + \mathcal{O}(h^{2\tau}),$$

the square root $\sqrt{1 - \beta^2}$ has the expansion $\sqrt{2C_\eta^{\text{Jac}}} h^{\tau/2} + \mathcal{O}(h^\tau)$. Inserting this expression into (4.31c), we obtain

$$\rho(M_{\omega_{\text{opt}}}^{\text{SOR}}) = 1 - 2\sqrt{2C_\eta^{\text{Jac}}} h^{\tau/2} + \mathcal{O}(h^\tau),$$

proving (4.31a,b). \square

4.6.4 Practical Handling of the SOR Method

According to Remark 3.6, choosing ω properly causes a practical problem. In general, the value of $\beta = \rho(M^{\text{Jac}})$ is unknown. Thus, the optimal relaxation parameter ω_{opt} is also not available. Then, one may proceed as follows (see also Young [412, §6.6]).

Initially, choose some $\omega \leq \omega_{\text{opt}}$, e.g., $\omega = 1$. Perform a few SOR steps with this parameter ω and determine an approximation $\tilde{\lambda}$ to $\rho(M_{\omega}^{\text{SOR}})$ from the ratios of $\|x^{m+1} - x^m\|_2$ (see the final part in §2.4). Using $\tilde{\lambda}$, we can produce β via Eq. (4.28a) (case $\omega \leq \omega_{\text{opt}}$):

$$\beta \approx \tilde{\beta} := |\tilde{\lambda} + \omega - 1| / (\omega \sqrt{\tilde{\lambda}})$$

(cf. (4.29a)). Using $\tilde{\beta}$, one determines an approximation ω to ω_{opt} by (4.28b). As long as $\omega \leq \omega_{\text{opt}}$, it is possible to iterate the described approximation of ω_{opt} . Since the function $\rho(M_{\omega}^{\text{SOR}})$ has a vertical tangent at $\omega = \omega_{\text{opt}}$ from the left, any deviation $\omega = \omega_{\text{opt}} - \varepsilon$ ($\varepsilon > 0$) to the left deteriorates the convergence considerably. Therefore, one should better choose $\omega \approx \omega_{\text{opt}}$ too large: $\omega > \omega_{\text{opt}}$. A program following this strategy can be found in Meis–Marcowitz [281, 282, Appendix A.4]. See also Reid [317].

4.6.5 p -Cyclic Matrices

The property ‘weakly 2-cyclic’ can be generalised. A is called *weakly p -cyclic* if a $p \times p$ block structure exists, so that only the blocks $A^{1,p}, A^{2,1}, A^{3,2}, \dots, A^{p,p-1}$ are nonvanishing. This case is discussed in detail by Varga [375, §4.2]. Under suitable further assumptions, the SOR method converges for

$$0 < \omega < \frac{p}{p-1}.$$

Optimal convergence holds for the unique positive root

$$\omega = \omega_{\text{opt}} < \frac{p}{p-1}$$

of the polynomial

$$(p-1)^{p-1} \rho(M^{\text{Jac}})^p \omega^p = p^p (\omega - 1).$$

The corresponding rate is

$$\rho(M_{\omega_{\text{opt}}}^{\text{SOR}}) = (\omega_{\text{opt}} - 1)(p-1) < 1$$

(cf. Eiermann–Niethammer–Ruttan [118]). Also in this case, the order of linear convergence is halved (cf. Theorem 4.31).

4.7 Application to the Model Problem

4.7.1 Analysis in the Model Case

For the five-point formula of the model problem in §1.2, Criterion 4.24 is always applicable, since one of the following cases applies:

- Pointwise variants (i.e., D is diagonal):
 - For the lexicographical ordering of the indices, A has the form (1.8). The sum $L + U$ satisfies the condition (4.25d) of Criterion 4.24.
 - For chequer-board ordering, A takes the 2-cyclic form (1.9), so that $L + U$ fulfils condition (4.25a).
- Blockwise variants (i.e., D is block-diagonal):
 - Assume that the rows or columns of the grid constitute the block structure. If these blocks are ordered *lexicographically*, $L + U$ shows the block-tridiagonal structure required in (4.25c).
 - In the case of the zebra ordering of the blocks as in Example 4.5, $L + U$ has a 2-cyclic form and satisfies (4.25a).

Besides property (4.24), Criterion 4.24 also proves that (A, D) is weakly 2-cyclic in all the cases mentioned above.

Theorem 4.32. *Assume the Poisson model problem in §1.2 with step size h .*

(a) For both the lexicographical and the chequer-board ordering, the pointwise Gauss–Seidel iteration has the convergence rate

$$\rho(M^{\text{GS}}) = \cos^2(\pi h) = 1 - \sin^2(\pi h) = 1 - \pi^2 h^2 + \mathcal{O}(h^4). \quad (4.32a)$$

(b) The row- and column-block-Gauss–Seidel iteration with lexicographical or zebra ordering has the convergence rate

$$\rho(M^{\text{blockGS}}) = 1 - 8 \sin^2 \frac{\pi h}{2} / \left(1 + 2 \sin^2 \frac{\pi h}{2} \right). \quad (4.32b)$$

(c) For the pointwise SOR methods with lexicographical or chequer-board ordering, the optimal relaxation parameter is $\omega_{\text{opt}} = 2/(1 + \sin(\pi h)) = 2 - 2\pi h + \mathcal{O}(h^2)$, leading to the convergence rate

$$\rho(M_{\omega_{\text{opt}}}^{\text{SOR}}) = \omega_{\text{opt}} - 1 = 1 - \frac{2 \sin(\pi h)}{1 + \sin(\pi h)} = \frac{1 - \sin(\pi h)}{1 + \sin(\pi h)}. \quad (4.33)$$

(d) For the block-SOR versions corresponding to case (b), the following values apply: $\omega_{\text{opt}} = 2/[1 + 2\sqrt{2} \sin(\pi h/2)/\cos(\pi h)]$ and

$$\rho(M_{\omega_{\text{opt}}}^{\text{blockSOR}}) = \omega_{\text{opt}} - 1 = 1 - \frac{4\sqrt{2} \sin(\pi h/2)}{\cos(\pi h) + 2\sqrt{2} \sin(\pi h/2)}. \quad (4.34)$$

Proof. By Conclusion 4.30, $\rho(M^{\text{[block]GS}})$ is the square of $\rho(M^{\text{[block]Jac}})$, described in (3.52) and (3.54) for the model problem. Parts (c) and (d) result from (4.28b,a). \square

Remark 4.33. The point- and blockwise Gauss–Seidel and SOR iterations described in Theorem 4.32 require the following effective amount of work:

$$\begin{aligned}\text{Eff}(\varPhi^{\text{GS}}) &= \pi^{-2} h^{-2} + \mathcal{O}(1) = 0.101 h^{-2} + \mathcal{O}(1), \\ \text{Eff}(\varPhi^{\text{blockGS}}) &= 0.7\pi^{-2} h^{-2} + \mathcal{O}(1) = 0.0709 h^{-2} + \mathcal{O}(1), \\ \text{Eff}(\varPhi^{\text{SOR}}) &= 0.7\pi^{-1} h^{-1} + \mathcal{O}(1) = 0.2228 h^{-1} + \mathcal{O}(1), \\ \text{Eff}(\varPhi^{\text{blockSOR}}) &= 0.9h^{-1}/(\sqrt{2}\pi) + \mathcal{O}(1) = 0.2026 h^{-1} + \mathcal{O}(1).\end{aligned}\quad (4.35)$$

Proof. The cost factors C_\varPhi are already represented in (3.22a–c): $C_\varPhi^{\text{GS}} = 1$, $C_\varPhi^{\text{blockGS}} = C_\varPhi^{\text{SOR}} = 7/5$, $C_\varPhi^{\text{blockSOR}} = 9/5$. The convergence rates (4.32a,b), (4.33), and (4.34) have the form $1 - C_\eta h^{-\tau}$ with the constants

$$\begin{aligned}C_\varPhi^{\text{GS}} &= \pi^2, & C_\varPhi^{\text{blockGS}} &= 2\pi^2, & \tau^{\text{[block]GS}} &= 2, \\ C_\varPhi^{\text{SOR}} &= \pi^2, & C_\varPhi^{\text{blockSOR}} &= 2\pi^2, & \tau^{\text{[block]SOR}} &= 1\end{aligned}$$

(cf. also (4.31b)). The assertion follows from the representation (2.32d). \square

The numbers in (4.35) indicate, e.g., that the block variants are more effective than the corresponding pointwise iterations. Although the SOR method is somewhat more expensive than the Gauss–Seidel iteration, the SOR method is already more effective than the Gauss–Seidel method if $h \leq 0.7/\pi \approx 1/5$.

4.7.2 Gauss–Seidel Iteration: Numerical Examples

Table 1.1 contains the results of the lexicographical and chequer-board Gauss–Seidel method. After showing more favourable values in the beginning, the error reduction factors $\varepsilon_{m-1}/\varepsilon_m$ converge for both orderings to $\rho(M^{\text{GS}}) = \cos^2(\frac{\pi}{32}) = 0.99039264$ (cf. (4.32a)).

The next test is concerned with the column-block structure. Table 4.1 contains the value of the iterates at the midpoint, the maximum norm $\varepsilon_m = \|u^m - u_h\|_\infty$, and the reduction factors $\rho_{m,m-1} = \varepsilon_{m-1}/\varepsilon_m$, which in the examples approximate (almost monotonically increasing) the limit

$$\rho(M^{\text{blockGS}}) = 1 - 8 \sin^2(\frac{\pi}{64}) / (1 + 2 \sin^2(\frac{\pi}{64}))^2 = 0.980923.$$

Although the lexicographical and zebra orderings yield different iterates x^m , they lead to the same convergence rate.

m	lexicographical ordering			zebra ordering of the blocks		
	$u_{16,16}$	ε_m	$\rho_{m,m-1}$	$u_{16,16}$	ε_m	$\rho_{m,m-1}$
5	-0.01926	1.23834	0.939842	-0.01950	1.17160	0.958731
10	-0.03592	1.01501	0.965208	-0.03752	0.95064	0.968133
20	-0.04928	0.76180	0.974912	-0.04015	0.71340	0.976522
100	0.34781	0.15219	0.980968	0.36033	0.14097	0.980690
200	0.47781	0.02229	0.980934	0.47964	0.02046	0.980690
300	0.49677	0.00325	0.980924	0.49703	0.00298	0.980623

Table 4.1 Results of the block-Gauss–Seidel method for $N=32$.

4.7.3 SOR Iteration: Numerical Examples

Table 1.2 contains the results of the SOR iteration for the relaxation parameter $\omega_{\text{opt}} = 2/(1 + \sin \pi h) = 1.821465$ which is optimal for the grid size $h = 1/32$. The reduction factor that should converge to $\omega_{\text{opt}} - 1 = 0.821465$ behaves very irregularly. In particular, one observes the tendency that in the beginning, the reduction factors are distinctly worse than the asymptotic convergence rate. The same observation holds for the block variants reported in Table 4.2 using the optimal relaxation parameter $\omega_{\text{opt}} = 1.7572848$. Because of the irregular behaviour of the factors $\rho_{m-1,m} = \varepsilon_{m-1}/\varepsilon_m$, an additional column with the factors

$$\bar{\rho}_m := (\varepsilon_{m-10}/\varepsilon_m)^{1/10} = (\rho_m \cdot \rho_{m-1} \cdot \rho_{m-2} \cdot \dots \cdot \rho_{m-9})^{1/10}$$

averaged over 10 values is presented in Table 4.2.

Since, initially, the convergence speed of the SOR method is slower, it is no contradiction when, as recommended in Meis–Marcowitz [281, 282] and verified by examples, ω is chosen somewhat larger than ω_{opt} in order to reach a given error bound as soon as possible.

m	lexicographic ordering			zebra ordering of the blocks		
	ε_m	$\rho_{m-1,m}$	$\bar{\rho}_m$	ε_m	$\rho_{m-1,m}$	$\bar{\rho}_m$
10	0.6217327	0.9109	0.8954	0.2978516	0.7280	0.8318
20	0.2146420	0.8841	0.7142	0.0279097	0.7847	0.7892
30	0.0146717	0.4890	0.7647	0.0023936	0.7675	0.7822
40	0.0017416	0.8516	0.8081	0.0002034	0.7723	0.7815
50	0.0001095	0.7375	0.7583	0.0000144	0.8078	0.7672
60	0.0000119	0.7585	0.8007	9.6527 ₁₀ -7	0.7777	0.7632
70	6.4684 ₁₀ -7	0.7814	0.7477	6.8937 ₁₀ -8	0.7684	0.7680
80	5.6020 ₁₀ -8	0.8006	0.7830	4.8121 ₁₀ -9	0.7545	0.7663
90	3.5398 ₁₀ -9	0.7565	0.7587	3.092 ₁₀ -10	0.7407	0.7600
100	2.269 ₁₀ -10	0.7139	0.7598	4.184 ₁₀ -11		

Table 4.2 Results of the block-SOR iteration for $N = 32$ with $\omega = \omega_{\text{opt}}$.

Chapter 5

Algebra of Linear Iterations

Abstract Most of the interesting iterative schemes are built from simpler units. The background is the fact that the set \mathcal{L} of consistent linear iterations form an algebra, i.e., there are several operations defined on \mathcal{L} . In this chapter we define the following operations: *Transposition* of a linear iteration Φ produces the adjoint iteration Φ^* . This gives rise to the definition of symmetric and positive definite iterations in Section 5.1. *Damping* of a linear iteration Φ by a scalar factor is often used to enforce convergence (cf. Section 5.2). *Addition* $\Phi + \Psi$ of two linear iterations is defined in §5.3. *Multiplication* of two linear iterations leads to the important construction of the product iteration: $\Phi, \Psi \mapsto \Phi \circ \Psi$ (cf. Section 5.4). For instance, iterations can be symmetrised (cf. §5.4.2). *Secondary iterations* are needed for the solution of auxiliary problems (cf. Section 5.5). Multiplication of Φ by a *left, right, or two-sided transformation* is described in Section 5.6. Using these operations, we can construct new linear iterations.

The following links refer to all interactions of the indicated operations with other operations.

Adjoint iteration Φ^* . (5.2a–c): iteration matrix and Φ^{**} , Remark 5.19b: damping, Remark 5.22: addition, Lemma 5.28: product, (5.48b): transformation.

Damping $\vartheta \cdot \Phi$. Remark 5.19: adjoint iteration, Remark 5.22: addition, (5.48a): transformation.

Addition $\Phi + \Psi$. Remark 5.22: adjoint iteration and damping, Exercise 5.27c: product, (5.48c): transformation.

Multiplication $\Phi \circ \Psi$. Exercise 5.27: addition, Lemma 5.28: adjoint iteration, (5.48d): transformation.

Transformations $\Phi \circ T$, $T \circ \Phi$. (5.38a–g), (5.43a–g), and (5.47a–g): domain, normal forms, etc., (5.48a–d): damping, adjoint iteration, addition, product.

5.1 Adjoint, Symmetric, and Positive Definite Iterations

In the following, it will be important to express dependence of the matrices $M = M_\Phi$, $N = N_\Phi$, $W = W_\Phi$ on the underlying matrix $A \in \mathfrak{D}(\Phi)$ of the system. Therefore we use the explicit notation $M_\Phi[A]$, $N_\Phi[A]$, $W_\Phi[A]$ introduced in §2.2.2.

5.1.1 Adjoint Iteration

5.1.1.1 Definition

Let $\Phi \in \mathcal{L}$ be any linear and consistent iteration with the domain $\mathfrak{D}(\Phi) \subset \mathbb{K}^{I \times I}$:

$$\Phi(x, b, A) = x - N_\Phi[A](Ax - b)$$

(note that the mapping $A \mapsto N[A]$ is defined for all matrices $A \in \mathfrak{D}(\Phi)$). The *adjoint iteration* $\Phi^* \in \mathcal{L}$ is defined on

$$\mathfrak{D}(\Phi^*) := \{A \in \mathbb{K}^{I \times I} : A^\mathsf{H} \in \mathfrak{D}(\Phi)\}$$

by

$$\Phi^*(x, b, A) := x - (N_\Phi[A^\mathsf{H}])^\mathsf{H}(Ax - b). \quad (5.1a)$$

Obviously, the matrix of the second normal form of Φ^* is

$$N_{\Phi^*}[A] = (N_\Phi[A^\mathsf{H}])^\mathsf{H}. \quad (5.1b)$$

The iteration matrix is

$$M_{\Phi^*}[A] = I - N_{\Phi^*}[A]A = (I - A^\mathsf{H}N_\Phi[A^\mathsf{H}])^\mathsf{H}.$$

If $N_{\Phi^*}[A]$ is regular, the corresponding matrix of the third normal form is

$$W_{\Phi^*}[A] = (W_\Phi[A^\mathsf{H}])^\mathsf{H}.$$

The definition of Φ^* implies the following more or less trivial statements:

$$\Phi^{**} = \Phi, \quad (5.2a)$$

$$M_{\Phi^*}[A] = A^{-1}(M_\Phi[A^\mathsf{H}])^\mathsf{H}A \quad \text{for regular } A, \quad (5.2b)$$

$$\rho(M_{\Phi^*}[A]) = \rho(M_\Phi[A^\mathsf{H}]). \quad (5.2c)$$

Because of (5.2c), the convergence of $\Phi^*(\cdot, \cdot, A)$ need not be analysed again if the convergence behaviour of $\Phi(\cdot, \cdot, A^\mathsf{H})$ is known.

5.1.1.2 Application to the Gauss–Seidel and SOR Iterations

We recall the splitting $A = D - E - F$ explained in (1.17) and (3.11a–d). To express dependence on the matrix A , we write

$$A = D[A] - E[A] - F[A].$$

The matrix A^H has a corresponding splitting

$$A^H = D[A^H] - E[A^H] - F[A^H].$$

The comparison with the diagonal, strictly upper and strictly lower triangular parts of

$$A^H = (D[A] - E[A] - F[A])^H = D[A]^H - E[A]^H - F[A]^H$$

proves that

$$D[A^H] = D[A]^H, \quad E[A^H] = F[A]^H, \quad F[A^H] = E[A]^H.$$

Provided that the diagonal part $D[A]$ is real, the adjoint Gauss–Seidel iteration is

$$\begin{aligned} (\Phi^{GS})^*(x, b, A) &= x - (D[A^H] - E[A^H])^{-H}(Ax - b) \\ &= x - (D[A] - F[A])^{-1}(Ax - b); \end{aligned}$$

i.e., instead of $N^{GS} = (D - E)^{-1}$ involving the lower triangular matrix, the upper triangular matrix is used in $N^{GS*} = (D - F)^{-1}$. The iteration matrix is

$$M^{GS*}[A] = I - (D - F)^{-1}A = (D - F)^{-1}E.$$

On the other hand, $(\Phi^{GS})^*$ can be generated differently. The Gauss–Seidel iteration requires some ordering of the indices in I . Using the *reverse ordering*, the roles of E and F are interchanged. This proves the following result.

Proposition 5.1. *Let Φ^{GS} correspond to a certain ordering of the index set I . The reverse ordering defines the backward Gauss–Seidel iteration $\Phi_{\text{backw}}^{\text{GS}}$. If $D = \text{diag}(A) \in \mathbb{R}^{I \times I}$, the iterations are related by*

$$(\Phi^{GS})^* = \Phi_{\text{backw}}^{\text{GS}}. \quad (5.3)$$

The algorithmic description of $\Phi_{\text{backw}}^{\text{GS}}$ uses (3.9) with the second line replaced with

$$\textbf{for } i := n \textbf{ downto } 1 \textbf{ do } x[i] := \left(b[i] - \sum_{j \in I \setminus \{i\}} a[i, j]x[j] \right) / a[i, i];$$

Remark 5.2. A statement analogous to (5.3) holds for the block-Gauss–Seidel iteration and the point- and blockwise SOR iterations. In particular, the iteration matrix of $(\Phi_{\omega}^{\text{SOR}})^*$ is $M_{\omega}^{\text{SOR}*} = (D - \omega F)^{-1}\{(1 - \omega)D + \omega E\}$.

5.1.2 Symmetric Iterations

Definition 5.3 (\mathcal{L}_{sym}). $\Phi \in \mathcal{L}$ is called *symmetric* if

$$\Phi = \Phi^*.$$

The corresponding set of symmetric iterations is denoted by \mathcal{L}_{sym} .

Using the definition of Φ^* , we obtain the characterisation

$$\Phi \in \mathcal{L}_{\text{sym}} \iff N[A] = N[A^H]^H \quad \text{for all } A \in \mathfrak{D}(\Phi).$$

Conclusion 5.4. If $\Phi \in \mathcal{L}_{\text{sym}}$ and $A \in \mathfrak{D}(\Phi)$, then $N[A]$ is Hermitian if A is so:

$$A = A^H \implies N[A] = N[A]^H. \quad (5.4)$$

Since a consistent linear iteration satisfies $M = I - NA$, we obtain the following criterion.

Criterion 5.5. Assume that $\mathfrak{D}(\Phi)$ contains only regular matrices; otherwise redefine $\mathfrak{D}(\Phi)$ by $\{A \in \mathfrak{D}(\Phi) : A \text{ regular}\}$. $\Phi \in \mathcal{L}_{\text{sym}}$ holds if and only if

$$(M[A] A^{-1})^H = M[A^H] A^{-H}.$$

A particular consequence is

$$A = A^H \implies M[A] A^{-1} = (M[A] A^{-1})^H.$$

Examples of symmetric iterations are the Richardson iteration with real Θ , since $N^{\text{Rich}}[A] = \Theta I$, and the (block-)Jacobi iteration, since

$$N^{\text{Jac}}[A] = \text{diag}\{A\}.$$

The Gauss–Seidel iteration is not symmetric.

Lemma 5.6. If $\Phi \in \mathcal{L}_{\text{sym}}$ and $A > 0$ (i.e., A positive definite), then $N > 0$ is a necessary condition for convergence.

Proof. $M = I - NA$ is similar to $I - \hat{A}$ with $\hat{A} := A^{1/2} N A^{1/2}$. If $N = N^H$ does not satisfy $N > 0$, \hat{A} has a nonpositive eigenvalue so that $\rho(M) \geq 1$. \square

Exercise 5.7. Let $A > 0$. $\Phi \in \mathcal{L}_{\text{sym}}$ implies that M is A -selfadjoint; i.e., $\langle Mx, y \rangle_A = \langle x, My \rangle_A$ for all x, y , where $\langle \cdot, \cdot \rangle_A$ is the A -scalar product (C.5b).

The construction of symmetric iterations will be discussed in §5.4.2.

5.1.3 Positive Definite Iterations

The positive definiteness of a linear iteration strengthens the symmetry in (5.4).

Definition 5.8 (\mathcal{L}_{pos}). $\Phi \in \mathcal{L}$ is called *positive definite* if it is symmetric and satisfies the implication

$$A > 0 \implies N[A] > 0 \quad \text{for all } A \in \mathfrak{D}(\Phi).$$

The positive definite iterations form the set $\mathcal{L}_{\text{pos}} \subset \mathcal{L}_{\text{sym}}$.

Since $N > 0$ implies $W = N^{-1} > 0$ for the matrix of the third normal form, the characterisation above is equivalent to

$$A > 0 \implies W[A] > 0 \quad \text{for all } A \in \mathfrak{D}(\Phi).$$

The examples of symmetric iterations above can be repeated.

Example 5.9. (a) The Richardson iteration with $\Theta > 0$ is positive definite.
(b) The (block-)Jacobi iteration is positive definite.

Lemma 5.6 yields a simple criterion for the positive definiteness of a symmetric iteration.

Criterion 5.10. If $\Phi \in \mathcal{L}_{\text{sym}}$ converges for all positive definite matrices $A \in \mathfrak{D}(\Phi)$, it belongs to \mathcal{L}_{pos} .

Convergence properties in the case of $A > 0$ and $N > 0$ are already discussed in §3.5.2. Theorem 6.11 will state that after suitable damping, convergence can be guaranteed.

There will be cases requiring semidefiniteness as defined below.

Definition 5.11 ($\mathcal{L}_{\text{semi}}$). $\Phi \in \mathcal{L}_{\text{sym}}$ is called *positive semidefinite* if it satisfies the implication

$$A \geq 0 \implies N[A] \geq 0 \quad \text{for } A \in \mathfrak{D}(\Phi).$$

The set of positive semidefinite iterations is denoted by $\mathcal{L}_{\text{semi}}$.

Note that $\mathcal{L}_{\text{pos}} \subset \mathcal{L}_{\text{semi}} \subset \mathcal{L}_{\text{sym}} \subset \mathcal{L}$. Examples of positive semidefinite iterations are A -orthogonal projections. The term of an (orthogonal) projection (cf. Definition A.29) is generalised to linear iterations.

Definition 5.12. (a) $\Phi \in \mathcal{L}$ is called a *projection* if $\Phi = \Phi \circ \Phi$, using the product defined in §5.4.

(b) If, in addition, $\Phi \in \mathcal{L}_{\text{sym}}$, Φ is called an A -orthogonal projection.

Exercise 5.13. Prove: (a) $\Phi \in \mathcal{L}$ is a projection if and only if $MN = 0$ or, equivalently, $NAN = N$ hold.

(b) The iteration matrix M of a projection Φ has a spectrum contained in $\{0, 1\}$. $\sigma(M) = \{1\}$ holds for the A -orthogonal projection $\Phi = \mathcal{Z}$ defined in Remark 5.22. $\sigma(M) = \{0\}$ holds for the direct solution $\Phi(x, b, A) = A^{-1}b$. All other projections Φ satisfy $\sigma(M) = \sigma(N) = \{0, 1\}$.

(c) Let $A > 0$. If $\Phi \in \mathcal{L}$ is an A -orthogonal projection, the matrices $A^{1/2}MA^{-1/2}$ and $A^{1/2}NA^{1/2}$ are orthogonal projections.

Finally, we mention another case that also leads to a positive definite matrix.

Definition 5.14 ($\mathcal{L}_{>0}$). $\Phi(\cdot, \cdot, A)$ for $A \in \mathfrak{D}(\Phi)$ is called *directly positive definite* if

$$A \text{ regular} \Rightarrow N[A]A > 0. \quad (5.5)$$

We denote the set of directly positive definite iterations by $\mathcal{L}_{>0}$.

Remark 5.15. (a) Let $\Phi \in \mathcal{L}_{\text{pos}}$ and $A > 0$. Then $N[A]A > 0$ holds if the matrices $N[A]$ and A commute.

(b) Examples of directly positive definite iterations are the Richardson iteration for $\Theta > 0$ and $A > 0$, and the Jacobi iteration if D and A commute (even the block-Jacobi iteration satisfies the latter condition for the Poisson model problem).

$\Phi(\cdot, \cdot, A)$ may be directly positive definite for general regular matrices A as the next example shows.

Example 5.16. An example of an iteration satisfying (5.5) is the choice $N[A] := A^H$, i.e., the iteration

$$\Phi(x, b, A) := x - A^H(Ax - b) \quad \text{for } A \in \mathfrak{D}(\Phi) := \{A \text{ regular}\}. \quad (5.6)$$

Regularity of A implies that $NA = A^HA$ is positive definite.

Remark 5.17. The damped version of iteration (5.6) is called the *Landweber iteration* (cf. Landweber [256]).¹ It can be viewed as the Richardson iteration applied to the system $A^H Ax = A^H b$. The solution to the minimisation problem (least squares problem)

$$\min_x \|Ax - b\|_2 \quad (A \in \mathbb{K}^{J \times I}, x \in \mathbb{K}^I, b \in \mathbb{K}^J, \#J \geq \#I),$$

is determined by the normal equations $A^H Ax = A^H b$, provided that the matrix A has maximal rank (cf. Björck [47]).

Concerning convergence see Remark 5.46.

¹ In tomography applications, this iteration is called the *simultaneous iterative reconstruction technique* (SIRT).

5.1.4 Positive Spectrum of NA

Assume that the matrix A of the system $Ax = b$ and the corresponding matrix $N = N[A]$ of the second normal form satisfy

$$\sigma(NA) \subset \mathbb{R}_+ = \{x > 0 : x \in \mathbb{R}\}. \quad (5.7)$$

Lemma 5.18. *Each of the following conditions is sufficient for property (5.7):*

- (a) $\Phi(\cdot, \cdot, A)$ is directly positive definite, i.e., $NA > 0$ (cf. (5.5)).
- (b) $A > 0$ and $N > 0$ (cf. Remark C.7).
- (c) Φ is a positive definite iteration applied to $A > 0$.

Convergence results based on (5.7) will follow in §6.2.1.

5.2 Damping of Linear Iterations

5.2.1 Definition

Here we discuss the operation

$$(\vartheta, \Phi) \in \mathbb{K} \times \mathcal{L} \quad \mapsto \quad \Phi_\vartheta = \vartheta \cdot \Phi \in \mathcal{L}.$$

The second normal form of $\Phi \in \mathcal{L}$ is

$$x^{m+1} = x^m - N(Ax^m - b).$$

Multiplying N by $\vartheta \in \mathbb{K}$, we obtain the corresponding *damped iteration*

$$x^{m+1} = x^m - \vartheta N(Ax^m - b) \quad (\vartheta \in \mathbb{K}). \quad (5.8)$$

Its usual notation is Φ_ϑ . More explicitly, the matrices $M_\vartheta = I - \vartheta NA$, $N_\vartheta = \vartheta N$, $W_\vartheta = \frac{1}{\vartheta} W$ of the normal forms are denoted by

$$M_{\Phi_\vartheta}[A] = I - \vartheta N_\Phi[A] A, \quad N_{\Phi_\vartheta}[A] = \vartheta N_\Phi[A], \quad W_{\Phi_\vartheta}[A] = \frac{1}{\vartheta} W_\Phi[A].$$

The term ‘damped’ holds in a proper sense only for $0 < \vartheta < 1$. For $\vartheta = 1$, we regain the original method, whereas for $\vartheta > 1$ the iteration Φ_ϑ is the ‘extrapolated’ version. For simplicity, we use the term ‘damped’ for all ϑ .

- Remark 5.19.** (a) Damping is associative: $\vartheta_1 \cdot (\vartheta_2 \cdot \Phi) = (\vartheta_1 \vartheta_2) \cdot \Phi$ for $\vartheta_1, \vartheta_2 \in \mathbb{K}$.
 (b) $(\vartheta \cdot \Phi)^* = \bar{\vartheta} \cdot \Phi^*$ holds for all $\vartheta \in \mathbb{K}$.
 (c) Symmetry of Φ implies symmetry of $\vartheta \cdot \Phi$ if and only if $\vartheta \in \mathbb{R}$.
 (d) If $\Phi \in \mathcal{L}_{\text{pos}}$ and $\vartheta > 0$, then $\vartheta \cdot \Phi \in \mathcal{L}_{\text{pos}}$.
 (e) If $\Phi \in \mathcal{L}_{\text{semi}}$ and $\vartheta \geq 0$, then $\vartheta \cdot \Phi \in \mathcal{L}_{\text{semi}}$.

The damped iteration can be implemented in two ways.

- (i) If the correction $x^m \mapsto \delta := N(Ax^m - b)$ is available (cf. (2.12')), the standard iteration $x^m \mapsto x^m - \delta$ can be replaced with $x^m \mapsto x^m - \vartheta \cdot \delta$.
- (ii) If only the implementation of the complete map $x \mapsto \Phi(x, b)$ is available, the result of the damped iteration can be obtained from

$$\Phi_\vartheta(x, b) := x + \vartheta [\Phi(x, b) - x] = (1 - \vartheta)x + \vartheta \Phi(x, b).$$

A natural question concerns the optimal damping parameter. Here optimality might be connected

- with the spectral radius: ϑ_{opt} is the minimiser of $\rho(M_\vartheta)$,
- with a certain norm: ϑ_{opt} is the minimiser of $\|M_\vartheta\|$.
- If we have some bound $\varphi(\vartheta)$ of these quantities, ϑ_{opt} may also be the minimiser of $\varphi(\vartheta)$.

An answer will be given in Theorem 6.7.

5.2.2 Damped Jacobi Iteration

In the case of the Jacobi iteration with $N^{\text{Jac}} = D^{-1}$ (cf. (3.7b)), the *damped Jacobi iteration* is

$$x^{m+1} = x^m - \vartheta D^{-1}(Ax^m - b) \quad (\vartheta \in \mathbb{K}). \quad (5.9)$$

Here $W[A] = D := \text{diag}\{A\}$ (pointwise Jacobi) or $W[A] = D := \text{blockdiag}\{A\}$ (blockwise Jacobi) hold.

The assumption $2D - A > 0$ required in Theorem 3.36 can be omitted, provided that a suitable damping is used. Theorem 6.11 will state that the damped Jacobi iteration $\Phi_\vartheta^{\text{Jac}}$ has the rate

$$\rho(M_\vartheta) = \|M_\vartheta\|_A = \|M_\vartheta\|_W = \max\{|1 - \vartheta\lambda_{\min}|, |1 - \vartheta\lambda_{\max}|\},$$

and that the choice $0 < \vartheta < 2/\lambda_{\max}(D^{-1}A)$ ensures convergence, while the optimal parameter $\vartheta_{\text{opt}} = \frac{2}{A+\lambda}$ yields $\rho(M_\vartheta) = \frac{A-\lambda}{A+\lambda}$, where $\lambda := \lambda_{\min}(D^{-1}A)$ and $A := \lambda_{\max}(D^{-1}A)$ (cf. Theorem 6.7).

Exercise 5.20. Prove that $\lambda_{\min}(D^{-1}A) \geq \frac{1}{\kappa(A)} = \frac{1}{\text{cond}_2(A)}$. Hint: $\|D\|_2 \leq \|A\|_2$.

Theorems 6.23 and 6.24 will involve a real matrix

$$A = A_0 + iA_1 \in \mathbb{R}^{I \times I} \quad (A_0 > 0, A_1^\mathsf{H} = A_1).$$

In this context, we remark that

$$D := \text{diag}\{A\} = \text{diag}\{A_0\} > 0$$

holds since the skew-Hermitian part iA_1 of a real matrix has vanishing diagonal entries.

In the case of the model problem, damping does not improve convergence. The values λ_{\min} and λ_{\max} defined in (3.1b,c) lead to $\lambda_{\min}(D^{-1}A) = 4h^2\lambda_{\min}$ and $\lambda_{\max}(D^{-1}A) = 4h^2\lambda_{\max}$. Since $\lambda_{\min}(D^{-1}A) + \lambda_{\max}(D^{-1}A) = 2$, the optimal ϑ_{opt} of Theorem 6.7 is $\vartheta_{\text{opt}} = 1$; i.e., the undamped Jacobi iteration is optimal. The fundamental reason is given by Conclusion 4.19: in the weakly 2-cyclic case, $\vartheta = 1$ is optimal.

The damped Jacobi iteration will become important in the context of multigrid methods (cf. Example 11.32b).

5.2.3 Accelerated SOR

One has tried to dampen (extrapolate) the SOR iteration. The resulting method $\vartheta \cdot \Phi_{\omega}^{\text{SOR}}$ is called the *accelerated overrelaxation* (AOR).

Analogously, the modifications SSOR and MSOR defined later in §5.4.3 and §6.3.3 yield accelerated versions SAOR and MAOR.

Concerning the questionability of iterations with multiple parameters, we recall Remark 3.6. For more details, we refer to Hadjidimos [211, §5].

5.3 Addition of Linear Iterations

Addition of linear iterations will play an important role for additive Schwarz and additive multigrid methods (cf. §12.5.3)

Definition 5.21. The sum of $\Phi, \Psi \in \mathcal{L}$ with corresponding matrices N_{Φ} and N_{Ψ} of the second normal form is defined by

$$(\Phi + \Psi)(x, b) := x - (N_{\Phi} + N_{\Psi})(Ax - b).$$

Note that $(\Phi + \Psi)(x, b) \neq \Phi(x, b) + \Psi(x, b)$, but

$$(\Phi + \Psi)(x, b) = \Phi(x, b) + \Psi(x, b) - x.$$

The underlying idea is that the corrections $N_{\Phi}(Ax - b)$ and $N_{\Psi}(Ax - b)$ may have different but complementary properties so that the (weighted) sum of the corrections is better than each single term.

Remark 5.22. Addition and damping of linear iterations satisfy the distributive properties

$$\begin{aligned} (\vartheta_1 \cdot \Phi) + (\vartheta_2 \cdot \Phi) &= (\vartheta_1 + \vartheta_2) \cdot \Phi, \\ \vartheta \cdot (\Phi + \Psi) &= (\vartheta \cdot \Phi) + (\vartheta \cdot \Psi). \end{aligned}$$

The adjoint operation yields

$$(\Phi + \Psi)^* = \Phi^* + \Psi^*.$$

The zero element of the addition is the *identical iteration* $\mathcal{Z}(x, b, A) := x$ corresponding to $N_{\mathcal{Z}} = 0$.

Convergence of Φ and Ψ is neither sufficient nor necessary for convergence of $\Phi + \Psi$. The sum

$$(\vartheta \cdot \Phi) + (\vartheta \cdot \Psi) = 2\vartheta \cdot \Phi$$

of convergent iterations can become divergent, since the scaling parameter leaves the interval of convergence (see, e.g., (6.5)). On the other hand, sums of divergent iterations may be convergent. We recall that $\Phi_i(x, b) := x - N_i(Ax - b)$ is divergent if the kernel of N_i is nontrivial (e.g., if N_i is only positive semidefinite). The next proposition shows that certain sums are convergent.

Proposition 5.23. *Let $A > 0$ and $N_i \geq 0$ for $1 \leq i \leq k$ with $\Gamma_i := \rho(N_i A)$. For $\omega_i \in (0, 2/\Gamma_i)$, define the weighted sum*

$$\Phi := \frac{1}{k} \sum_{i=1}^k \omega_i \cdot \Phi_i,$$

i.e.,

$$\Phi(x, b) = x - \frac{1}{k} \sum_{i=1}^k \omega_i N_i (Ax - b).$$

The iteration Φ is convergent if and only if²

$$\bigcap_{i=1}^k \ker(N_{\Phi_i}) = \{0\}. \quad (5.10)$$

Proof. $N_i \geq 0$ and the definition of Γ_i imply that $0 \leq N_i \leq \Gamma_i A^{-1}$. Summation yields

$$0 \leq N_{\Phi} = \frac{1}{k} \sum_{i=1}^k \omega_i N_i \leq \left(\frac{1}{k} \sum_{i=1}^k \omega_i \Gamma_i \right) A^{-1} < 2A^{-1}.$$

Convergence holds if and only if $0 \leq N_{\Phi}$ can be replaced by $0 < N_{\Phi}$. If (5.10) is not valid, there is some $x \neq 0$ with $x \in \ker(N_{\Phi_i})$ for all $1 \leq i \leq k$. The definition of N_{Φ} shows that $N_{\Phi}x = 0$; hence Φ is divergent. If (5.10) holds, for any $x \neq 0$ there exist an index i_x with $N_{i_x}x \neq 0$ and we obtain

$$\langle N_{\Phi}x, x \rangle = \frac{1}{k} \sum_{i=1}^k \omega_i \underbrace{\langle N_i x, x \rangle}_{\geq 0} \geq \frac{\omega_{i_x}}{k} \langle N_{i_x} x, x \rangle > 0.$$

$N_{\Phi} > 0$ proves convergence of Φ . □

² $\ker(A) = \{x : Ax = 0\}$ is the kernel of the matrix A .

5.4 Product Iterations

The following product is the sequential application of two mappings. It is not related to the product of N_Φ and N_Ψ , but leads to the product of the iteration matrices.

5.4.1 Definition and Properties

Definition 5.24. For $\Phi, \Psi \in \mathcal{L}$, the product iteration $\Phi \circ \Psi$ is defined by

$$x^{m+1} = (\Phi \circ \Psi)(x^m, b) := \Phi(\Psi(x^m, b), b). \quad (5.11)$$

Special product iterations will be studied, e.g., in §6.

Proposition 5.25. (a) If Φ and Ψ are consistent, then $\Phi \circ \Psi$ is also.³

(b) The iteration matrices of Φ , Ψ , and $\Phi \circ \Psi$ are related by

$$M_{\Phi \circ \Psi} = M_\Phi M_\Psi. \quad (5.12a)$$

The convergence rates of $\Phi \circ \Psi$ and $\Psi \circ \Phi$ are identical.

(c) Let N_Φ , N_Ψ , and $N_{\Phi \circ \Psi}$ be the respective matrices of the second normal form of Φ , Ψ , $\Phi \circ \Psi$. Then $\Phi \circ \Psi$ is characterised by (5.12b), where the last expression requires consistency of Φ :

$$N_{\Phi \circ \Psi} = M_\Phi N_\Psi + N_\Phi = N_\Phi + N_\Psi - N_\Phi A N_\Psi. \quad (5.12b)$$

(d) Let W_Φ , W_Ψ and $W_{\Phi \circ \Psi}$ be the respective matrices of the third normal form of Φ , Ψ , $\Phi \circ \Psi$. Assume that N_Φ and N_Ψ are regular. If $W_\Phi - W_\Psi - A$ is singular, $\Phi \circ \Psi$ diverges. Otherwise,

$$W_{\Phi \circ \Psi} = W_\Psi (W_\Phi + W_\Psi - A)^{-1} W_\Phi. \quad (5.12c)$$

Proof. (a) Inserting the solution $x^m := x^* := A^{-1}b$ into (5.11), we obtain that $(\Phi \circ \Psi)(x^*, b) := \Phi(\Psi(x^*, b), b) = \Phi(x^*, b) = x^*$.

(b) Use $\Phi(\Psi(x, b), b) = \Phi(M_\Psi x + N_\Psi b, b) = M_\Phi(M_\Psi x + N_\Psi b) + N_\Phi b = M_\Phi M_\Psi x + (M_\Phi N_\Psi + N_\Phi)b$. The equality $\rho(M_\Phi M_\Psi) = \rho(M_\Psi M_\Phi)$ follows from Lemma A.20.

(c) The previous part (b) proves that $N_{\Phi \circ \Psi} = M_\Phi N_\Psi + N_\Phi$. Consistency yields $M_\Phi = I - N_\Phi A$ (cf. Theorem 2.11). This proves the last statement in (5.12b).

(d) Regularity of N_Φ and N_Ψ ensures the existence of $W_\Phi = N_\Phi^{-1}$ and $W_\Psi = N_\Psi^{-1}$. From (5.12b), we conclude that $N_{\Phi \circ \Psi} = W_\Phi^{-1} + W_\Psi^{-1} - W_\Phi^{-1} A W_\Psi^{-1} =$

³ The background of this statement is the fact that definition (5.11) also applies to inconsistent iterations. The definition of \mathcal{L} includes the consistency.

$W_\Phi^{-1} (W_\Phi + W_\Psi - A) W_\Psi^{-1}$. Singularity of $W_\Phi - W_\Psi - A$ implies the singularity of $N_{\Phi \circ \Psi}$. According to Corollary 2.17b, $\Phi \circ \Psi$ cannot be convergent. Otherwise the inversion of $N_{\Phi \circ \Psi} = W_\Phi^{-1} (W_\Phi + W_\Psi - A) W_\Psi^{-1}$ yields (5.12c). \square

Remark 5.26. (a) Convergence of the factors Φ and Ψ is neither sufficient nor necessary for the convergence of $\Phi \circ \Psi$.

(b) However, a sufficient condition for convergence of $\Phi \circ \Psi$ is that both Φ and Ψ satisfy the condition (2.20) of Theorem 2.19 with respect to the *same* norm:

$$\|M_\Phi\| < 1 \quad \text{and} \quad \|M_\Psi\| < 1,$$

where one of the strict inequalities ‘ < 1 ’ may be replaced with ‘ ≤ 1 ’.

The last statement is trivial because $\|M_\Phi M_\Psi\| \leq \|M_\Phi\| \|M_\Psi\|$. Part (a) is illustrated by two examples.

The first example shows that $\Phi \circ \Psi$ may diverge, although Φ and Ψ converge. Let $A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$. The first iteration Φ is defined via $W_\Phi = \begin{bmatrix} 1 & 2 \\ 0 & -2 \end{bmatrix}$. The iteration matrix is $M_\Phi = I - W_\Phi^{-1} A = \begin{bmatrix} -1/2 & -1 \\ 1/2 & 1 \end{bmatrix}$. Since its eigenvalues are 0 and $\frac{1}{2}$, $\rho(M_\Phi) = \frac{1}{2}$ implies convergence. The second iteration Ψ uses $W_\Psi = \begin{bmatrix} -2 & 4 \\ 0 & 1 \end{bmatrix}$. The iteration matrix is $M_\Psi = \begin{bmatrix} -1/2 & 1/2 \\ -1 & 1 \end{bmatrix}$. Since $M_\Psi = M_\Phi^\top$, also $\rho(M_\Psi) = \frac{1}{2}$ holds; i.e., both Φ and Ψ are convergent. The product iteration $\Phi \circ \Psi$ has the iteration matrix $M_\Phi M_\Psi = \frac{5}{4} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$ with the eigenvalues 0 and $\frac{5}{2}$. Hence the product iteration is divergent.

On the other hand, products of divergent methods may converge. For instance, the Kaczmarz iteration in §5.6.3 is convergent, although it is the product of divergent projections.

The algebraic properties of the product operation are the subject of the next exercise. The adjoint $(\Phi \circ \Psi)^*$ is investigated in Lemma 5.28. The interaction with the damping of an iteration will be discussed in §5.4.2.2.

Exercise 5.27. Prove: (a) The unit element of the product operation is the identical iteration \mathcal{Z} defined in Remark 5.22: $\Phi = \Phi \circ \mathcal{Z} = \mathcal{Z} \circ \Phi$.

(b) The product is associative:

$$(\Phi \circ \Psi) \circ \Omega = \Phi \circ (\Psi \circ \Omega) \quad \text{for } \Phi, \Psi, \Omega \in \mathcal{L}.$$

(c) The interaction with addition is not distributive. Instead we have

$$\begin{aligned} (\Phi' + \Phi'') \circ \Psi &= (\Phi' \circ \Psi) + (\Phi'' \circ \Psi) - \Psi, \\ \Phi \circ (\Psi' + \Psi'') &= (\Phi \circ \Psi') + (\Phi \circ \Psi'') - \Phi. \end{aligned}$$

5.4.2 Constructing Symmetric Iterations

5.4.2.1 Definition of Φ^{sym}

We shall see that symmetric and, in particular, positive definite iterations offer computational advantages. On the other hand, important iterations as the Gauss–Seidel iteration are not symmetric. The product operation \circ enables a simple construction of a related symmetric iteration. We recall that the matrices related to Φ applied to the system $Ax = b$ are denoted by $M_\Phi[A]$, $N_\Phi[A]$, and $W_\Phi[A]$.

A first combination of the mapping $\Phi \mapsto \Phi^*$ and \circ is the subject of the next lemma.

Lemma 5.28. *The product iteration $\Phi \circ \Psi$ satisfies $(\Phi \circ \Psi)^* = \Psi^* \circ \Phi^*$.*

Proof. Let A be any underlying matrix. Definition (5.1a,b) states that $(\Phi \circ \Psi)^*$ is characterised by

$$N_{(\Phi \circ \Psi)^*}[A] = (N_{\Phi \circ \Psi}[A^H])^H.$$

The definition of $N_{\Phi \circ \Psi}[A]$ in (5.12b) shows that

$$\begin{aligned} (N_{\Phi \circ \Psi}[A^H])^H &= (N_\Phi[A^H] + N_\Psi[A^H] - N_\Phi[A^H]A^HN_\Psi[A^H])^H \\ &= (N_\Phi[A^H])^H + (N_\Psi[A^H])^H - N_\Psi[A^H]^HA(N_\Phi[A^H])^H. \end{aligned}$$

Using again (5.1a,b) and (5.12b), we can continue with

$$(N_{\Phi \circ \Psi}[A^H])^H = (N_{\Phi^*}[A]) + (N_{\Psi^*}[A]) - N_{\Psi^*}[A]A N_{\Phi^*}[A] = N_{\Psi^* \circ \Phi^*}[A],$$

and the equations above yield the identity $N_{(\Phi \circ \Psi)^*}[A] = N_{\Psi^* \circ \Phi^*}[A]$ implying $(\Phi \circ \Psi)^* = \Psi^* \circ \Phi^*$. \square

For each iteration $\Phi \in \mathcal{L}$, we define the corresponding *symmetrised iteration*

$$\Phi^{\text{sym}} := \Phi^* \circ \Phi. \quad (5.13)$$

Theorem 5.29. (a) Φ^{sym} is a symmetric iteration: $\Phi^{\text{sym}} \in \mathcal{L}_{\text{sym}}$.

(b) Let $\Phi \in \mathcal{L}$ be associated with $N[A]$ and $W[A]$. We use the abbreviations

$$N = N[A], \quad N' := (N[A^H])^H, \quad \text{and} \quad W = W[A], \quad W' := (W[A^H])^H.$$

Then the matrices associated with Φ^{sym} read as follows, provided that the inverses $W = N^{-1}$, $W' = N'^{-1}$, and $(W + W' - A)^{-1}$ exist:

$$M^{\text{sym}} = (I - N'A)(I - NA) = I - N^{\text{sym}}A,$$

$$N^{\text{sym}} = N + N' - N'AN,$$

$$W^{\text{sym}} = W(W + W' - A)^{-1}W'.$$

(c) If $A = A^H$, then

$$N' = N^H, \quad W' = W^H, \quad N^{\text{sym}} = (N^{\text{sym}})^H, \quad W^{\text{sym}} = (W^{\text{sym}})^H,$$

while

$$A M^{\text{sym}} = (M^{\text{sym}})^H A.$$

Proof. According to Lemma 5.28, $(\Phi^{\text{sym}})^* = (\Phi^* \circ \Phi)^* = \Phi^* \circ \Phi^{**}$. Property (5.2a) yields $\Phi^* \circ \Phi^{**} = \Phi^* \circ \Phi = \Phi^{\text{sym}}$. The equality $(\Phi^{\text{sym}})^* = \Phi^{\text{sym}}$ states that Φ^{sym} is symmetric (cf. Definition 5.3). The parts (b) and (c) are elementary. \square

The same proof applies to the following statement.

Corollary 5.30. Let $\Psi \in \mathcal{L}_{\text{sym}}$, while $\Phi \in \mathcal{L}$ is arbitrary. Then also

$$\Phi^* \circ \Psi \circ \Phi \in \mathcal{L}_{\text{sym}}.$$

5.4.2.2 Combination with Damping

There are three possibilities to use damping in connection with the product iteration.

- We may construct Φ^{sym} as above and afterward apply damping to Φ^{sym} . The result is denoted by $(\Phi^{\text{sym}})_\vartheta := \vartheta \cdot \Phi^{\text{sym}}$.
- Instead to $\Phi^* \circ \Phi$, we apply the product to the damped iteration Φ_ϑ . The product is denoted by $(\Phi_\vartheta)^{\text{sym}} := \Phi_\vartheta^* \circ \Phi_\vartheta$.
- We may even combine both approaches by damping the second product resulting in $(\Phi_{\vartheta_1})_{\vartheta_2}^{\text{sym}} = \vartheta_2 \cdot (\vartheta_1 \cdot \Phi)^{\text{sym}}$.

Note that in general $(\Phi^{\text{sym}})_\vartheta \neq (\Phi_\vartheta)^{\text{sym}}$. In the case of $A = A^H$ (cf. Theorem 5.29c), the corresponding matrices of the second and third normal forms are

$$(N^{\text{sym}})_\vartheta = \vartheta (N + N^H - N^H A N), \quad (W^{\text{sym}})_\vartheta = \frac{1}{\vartheta} W (W + W^H - A)^{-1} W^H, \\ (N_\vartheta)^{\text{sym}} = \vartheta (N + N^H - \vartheta N^H A N), \quad (W_\vartheta)^{\text{sym}} = \frac{1}{\vartheta} W (W + W^H - \vartheta A)^{-1} W^H.$$

Here, N and $W = N^{-1}$ refer to Φ without damping.

Remark 5.31. Let $A > 0$ and $W + W^H > 0$ be valid. If $W + W^H > A$ (cf. (6.10)) does not hold, one can choose a positive factor ϑ bounded by

$$\vartheta < \|A^{-1/2}(W + W^H)A^{-1/2}\|_2.$$

Then $W + W^H > \vartheta A$ holds and implies that $(W_\vartheta)^{\text{sym}} > 0$. Hence, $(\Phi_\vartheta)^{\text{sym}}$ is a positive definite iteration.

5.4.2.3 Practical Implementation

The description of N^{sym} does *not* mean that the iteration Φ^{sym} should be implemented via $x^{m+1} = x^m + N^{\text{sym}}(Ax^m - b)$. In general, it is cheaper to follow the definition of the product:

$$x^m \mapsto x^{m+1/2} := \Phi(x^m, b) \mapsto x^{m+1} := \Phi^*(x^{m+1/2}, b). \quad (5.14)$$

Nevertheless, there may be more efficient implementations of $x^m \mapsto x^{m+1}$ in special cases (cf. Remark 6.27).

In particular, it makes no sense to apply the symmetrisation to symmetric iterations. In these cases $\Phi^{\text{sym}} = \Phi \circ \Phi =: \Phi^2$ is the twofold application of Φ . Using (5.14), the only difference is that we ignore every second iterate. The convergence rate squares, $\rho(M_{\Phi^2}) = \rho(M_{\Phi}^2) = \rho(M_{\Phi})^2$, while the cost doubles. The effective work is invariant: $\text{Eff}(\Phi^2) = \text{Eff}(\Phi)$ (cf. (2.31a)).

5.4.3 Symmetric Gauss–Seidel and SSOR

Since the Gauss–Seidel and SOR are nonsymmetric iterations, their symmetric versions are of interest:

$$\Phi^{\text{symGS}} := \Phi_{\text{backw}}^{\text{GS}} \circ \Phi^{\text{GS}}, \quad \Phi_{\omega}^{\text{SSOR}} := \Phi_{\omega}^{\text{backwSOR}} \circ \Phi_{\omega}^{\text{SOR}} \in \mathcal{L}_{\text{sym}}. \quad (5.15)$$

In these definitions we use that $\Phi_{\text{backw}}^{\text{GS}} = (\Phi^{\text{GS}})^*$ and $\Phi_{\omega}^{\text{backwSOR}} = (\Phi_{\omega}^{\text{SOR}})^*$ (cf. Proposition 5.1 and Remark 5.2). The extension of (5.15) to block versions is obvious, but note that (5.15) holds only for A with $D[A] = D[A^H]$; i.e., the diagonal blocks of A must be Hermitian.

The symmetric SOR method defined in (5.15) is abbreviated as SSOR.⁴ Note that the SSOR method again depends on the parameter ω .

Convergence statements on the symmetric Gauss–Seidel and SOR iterations will be given in §6.3.

5.5 Combination with Secondary Iterations

Writing the matrix N of the second normal form as W^{-1} , we see that possibly an auxiliary problem $W\delta = d$ has to be solved. In the example of §3, the solution is trivial since either W is diagonal (Jacobi) or a triangular matrix (Gauss–Seidel). But in the case of the blockwise Jacobi iteration, one has already to solve smaller systems of the form $A^{ii}\delta^i = d^i$, where $A^{ii} = D^{ii}$ are the diagonal blocks of A . In the Poisson model case, D^{ii} is tridiagonal and the exact solution via LU decomposition is easy, but below we will discuss a more involved problem.

⁴ In the Soviet literature, the term *alternate-triangular method* is used (cf. Samarskii–Nikolaev [330, Chapter 10]).

5.5.1 First Example for Secondary Iterations

The differential equation

$$-\Delta u + u_{xy} + au_x = f \quad \text{in } \Omega \quad (5.16a)$$

with the boundary condition (1.1b): $u = 0$ on $\Gamma = \partial\Omega$ can be discretised, e.g., by the seven-point formula

$$\frac{1}{2}h^{-2} \begin{bmatrix} -1 & -1 & 0 \\ 1+ah & 6 & 1-ah \\ 0 & -1 & -1 \end{bmatrix} u = f \quad (5.16b)$$

abbreviating the equations

$$\begin{aligned} \frac{1}{2}h^{-2} [6u(x, y) - u(x-h, y+h) - u(x, y+h) - u(x, y-h) \\ - u(x+h, y-h) - (1+ah)u(x-h, y) - (1-ah)u(x+h, y)] = f \end{aligned}$$

for $(x, y) \in \Omega_h$ (cf. §1.3.2 and [193, §5.1.4]). As long as $|ah| \leq 1$, the matrix A is an M-matrix. However, note that A is not symmetric, unless $a = 0$.

Assume that we have already a good solver for the simpler Poisson model problem related to the five-point formula $B = h^{-2} \begin{bmatrix} -1 & -1 \\ -1 & 4 & -1 \\ -1 & \end{bmatrix}$. Using B , we define the iteration Φ by

$$x^{m+1} = x^m - B^{-1}(Ax^m - b). \quad (5.17)$$

The associated matrices are $M = I - B^{-1}A$, $N = B^{-1}$, $W = B$.

We shall prove in Proposition 7.60 that the iteration Φ converges perfectly (at least together with an appropriate damping). The rate of convergence as well as the contraction numbers with respect to $\|\cdot\|_A$ do not deteriorate for $h \rightarrow 0$. However, the good convergence properties are offset by the difficulty in performing the mapping $d \mapsto B^{-1}d$ required for solving the system $B\delta = d$. In the given case, this would in principle be possible since there are direct solvers for the Poisson model problem (see the last paragraph in §1.5). These solvers, however, do not work for domains different from rectangles.

One remedy is the *approximate* solution of the mapping $d \mapsto B^{-1}d$ (i.e., of the equation $B\delta = d$) by some iterative technique. The iteration Φ^{\sec} for solving the auxiliary problem $B\delta = d$ is called the *secondary iteration* and leads to the following *composed iteration*:

composed iteration $\Phi_k(\cdot, \cdot, A)$: (5.18)

..... (5.18a)

$x^m \mapsto d := Ax^m - b$; (5.18b)

secondary iteration for solving $B\delta = d$: (5.18b)

set the starting iterate $\delta^0 := 0$; (5.18b₁)

perform k iteration steps $\delta^{i-1} \mapsto \delta^i = \Phi^{\sec}(\delta^{i-1}, d; B)$; (5.18b₂)

$x^{m+1} := x^m - \delta^k$. (5.18c)

The iteration Φ^{\sec} in (5.18b₂) may be replaced with semi-iterations (cf. §8).

The number k of inner iteration steps may depend on m or be constant. The larger k is, the better the sequences x^m from (5.17) and (5.18a–c) coincide. On the other hand, one would like to choose k as small as possible, since the amount of work for one (outer) iteration step Φ_k increases with k . This leads us to the natural question about the optimal choice of k .

5.5.2 Second Example for Secondary Iterations

Above both iterations Φ and Φ^{sec} belong to the same space \mathbb{K}^I . The next example uses more than one Φ_i^{sec} belonging to smaller vector spaces \mathbb{K}^{I_i} with $I_i \subsetneq I$.

In the case of the block variants of the Jacobi, Gauss–Seidel, or SOR methods, we have to solve systems for each block. Since in the model case the blocks have tridiagonal structure, the exact solution is easily and cheaply computable. This is different for three-dimensional boundary value problems; e.g., for the Poisson equation $-u_{xx} - u_{yy} - u_{zz} = f$ in the cube $\Omega = (0, 1)^3$.

In the two-dimensional case, the rows and columns are the natural block structures. In the three-dimensional case, we have several possibilities. We may gather all variables belonging to the varying x - or y - or z -value, while the other coordinates are fixed. Then we obtain $(N-1)^2$ one-dimensional blocks of size $N-1$, where $h = 1/N$. Instead, the blocks can be formed plane-wise: all variables corresponding to the point set $I^i := \{(ih, jh, kh) : 1 \leq j, k \leq N-1\}$ form one block. This yields $N-1$ blocks of size $(N-1)^2$ and defines the *yz-plane block structure*. Alternatively, we can use the xy - or xz -planes. Since larger blocks yield better convergence properties (cf. Theorem 7.13), we may be interested in the plane-block-SOR variant. Then we have to solve the auxiliary systems $D^{ii}\delta^i = d^i$, where the matrix D^{ii} corresponding to the diagonal block is a five-point formula. Again we need a secondary iteration for solving these subsystems.

At first glance, the secondary iteration seems to be expensive, but the submatrices D^{ii} have an advantageous property. Exercise 5.32 shows that the blocks are strongly diagonally dominant and have an h -independent condition.

Exercise 5.32. Prove: (a) Discretisation of the three-dimensional Poisson problem, corresponding to the five-point formula (1.4a) in two dimensions, reads as

$$h^{-2} [6u(x, y, z) - u(x-h, y, z) - u(x+h, y, z) - u(x, y-h, z) \\ - u(x, y+h, z) - u(x, y, z-h) - u(x, y, z+h)] = f(x, y, z).$$

The plane-wise block structure is defined by blocks of those grid points (x, y, z) with constant z . The diagonal blocks correspond to the five-point formula

$$D = h^{-2} \begin{bmatrix} -1 & & \\ -1 & 6 & -1 \\ & -1 & \end{bmatrix}.$$

- (b) $\sigma(D) \subset [2h^{-2}, 10h^{-2}]$, $\|D\|_\infty \leq 10h^{-2}$, $\|D^{-1}\|_\infty \leq \frac{1}{2}h^2$, $\text{cond}_\infty(D) \leq 5$.
- (c) Apply Proposition 7.23 to prove the h -independent rate 2/3 of the Jacobi iteration applied to the matrix D .

5.5.3 Convergence Analysis in the General Case

In the following, we denote the matrix of the third normal form of $\Phi = \Phi_A$ by B (cf. (5.17)) instead of W . The iteration matrix of $\Phi_A(x, b) = x - B^{-1}(Ax - b)$ is

$$M_A = I - B^{-1}A. \quad (5.19)$$

For solving the auxiliary equation $B\delta = d$, we apply the secondary iteration⁵ Φ_B :

$$\delta^{m+1} = \delta^m - C^{-1}(B\delta^m - d) = M_B\delta^m + N_Bd \quad (5.20)$$

with the iteration matrix $M_B = I - C^{-1}B$.

In the following, we always apply the secondary iteration in (5.18b) with a constant value k (otherwise we have to prescribe a suitable stopping criterion).

Lemma 5.33. *Let $\Phi_A \in \mathcal{L}$ be the iteration for solving $Ax = b$ by (5.17), while $\Phi_B \in \mathcal{L}$ belongs to $B\delta = d$. The composed iteration Φ_k defined for fixed $k > 0$ by (5.18) is a linear and consistent iteration for solving $Ax = b$. Its iteration matrix is*

$$M_k = I - \sum_{\mu=0}^{k-1} M_B^\mu N_B A \quad (M_B, N_B \text{ in (5.20)}). \quad (5.21a)$$

Consistency of Φ_B yields

$$M_k = M_A + M_B^k B^{-1}A \quad (M_A \text{ in (5.19)}). \quad (5.21b)$$

The matrix of the second normal form is

$$N_k = (I - M_B^k)B^{-1}. \quad (5.21c)$$

If M_B has an eigenvalue λ with $\lambda^k = 1$, the iteration Φ_k diverges; otherwise, the matrix of the third normal form (2.12) can be written as

$$W_k = B(I - M_B^k)^{-1}. \quad (5.21d)$$

Proof. According to Theorem 2.14, the iterate δ^k in (5.18b₂) has the representation $\delta^k = \sum_{\mu=0}^{k-1} M_B^\mu N_B d$ (note that $\delta^0 = 0$ and $d = Ax^m - b$). This proves (5.21a). The consistency of Φ_B implies that $N_B = (I - M_B)B^{-1}$ (cf. (2.9'')). The statements (5.21b–c) can be concluded from $\sum_{\mu=0}^{k-1} M_B^\mu (I - M_B) = I - M_B^k$ and (5.19). $W_k = N_k^{-1}$ holds for invertible N_k and proves (5.21d). \square

The representation (5.21b) permits an interpretation of the iteration matrix M_k as a perturbation of the iteration matrix M_A . The contraction number of Φ_k can be estimated as follows.

⁵ For simplicity assume $\mathfrak{D}(\Phi_A) = \{A\}$ and $\mathfrak{D}(\Phi_B) = \{B\}$.

Lemma 5.34. Let $\Phi_A, \Phi_B \in \mathcal{L}$ be the iterations in Lemma 5.33. The contraction numbers of Φ_k with respect to the spectral norm and, if B or A are also positive definite, with respect to the norms $\|x\|_B = \|B^{1/2}x\|_2$ and $\|x\|_A = \|A^{1/2}x\|_2$ are

$$\|M_k\|_2 \leq \|M_A\|_2 + \|M_B\|_2^k \|B^{-1}A\|_2, \quad (5.22a)$$

$$\|M_k\|_B \leq \|M_A\|_B + \|M_B\|_B^k \|B^{-1/2}AB^{-1/2}\|_2 \quad (\text{if } B > 0), \quad (5.22b)$$

$$\|M_k\|_A \leq \|M_A\|_A + \|M_B\|_A^k \|A^{1/2}B^{-1}A^{1/2}\|_2 \quad (\text{if } A > 0). \quad (5.22c)$$

Knowing the spectral radius $\rho(M_B)$ is not sufficient for analysing the secondary iteration because the spectral radius only describes the asymptotic convergence, whereas here we need precise upper bounds after a fixed number of k iteration steps. As a remedy, the contraction number of Φ_B may be replaced by the numerical radius $r(M_B)$.

Exercise 5.35. Prove: (a) Let Φ_A and Φ_B be linear and consistent. Then

$$r(M_k) \leq r(M_A) + 2r(M_B)^k \|B^{-1}A\|_2. \quad (5.22d)$$

(b) The factor $\|B^{-1}A\|_2$ in (5.22a,d) is bounded by

$$1 - \|M_A\|_2 \leq \|B^{-1}A\|_2 \leq 1 + \|M_A\|_2. \quad (5.22e)$$

The conclusions that we can draw from (5.22a–d) are the subject of the next statement.

Conclusion 5.36. (a) Let one of the quantities $\|M_A\|_2$, $\|M_A\|_B$, $\|M_A\|_A$, $r(M_A)$ and the corresponding quantities $\|M_B\|_2$, $\|M_B\|_B$, $\|M_B\|_A$, $r(M_B)$ be smaller than 1. Then the composed iteration Φ_k converges for sufficiently large k .

(b) One should choose k sufficiently large, so that the right-hand side of (5.22a) is of a size comparable with $\|M_A\|_2$, e.g., $\frac{1}{2}(1 + \|M_A\|_2)$ (similarly for (5.22b–d)). If $\|M_A\|_2 \leq \zeta < 1$ (ζ independent of h) and $\|M_B\|_2 = 1 - \mathcal{O}(h^\beta)$ ($\beta > 0$), the inequality $\|M_k\|_2 \leq (1 + \zeta)/2$ can be achieved with $k = \mathcal{O}(h^{-\beta})$. In this case, the effective amount of work for Φ_k is also of the order $\text{Eff}(\Phi_k) = \mathcal{O}(h^{-\beta})$. If, however, $\|M_A\|_2 = 1 - \mathcal{O}(h^\alpha)$ ($\alpha > 0$), inequality (5.22a) admits only the unfavourable estimate $\text{Eff}(\Phi_k) = \mathcal{O}(h^{-\alpha-\beta})$.

In particular, (5.22a–d) yields no statement⁶ ensuring the convergence of Φ_k for small k . Since, according to (5.22e), the factor $\|B^{-1}A\|_2$ attains at best the value ≈ 1 , one needs at least $k = \mathcal{O}(h^{-\beta})$ iterations to make the right-hand side of (5.22a) smaller than 1.

Since Φ_k is again a linear and consistent iteration, Φ_k may be used as the basic iteration of a semi-iteration (cf. §8). Another situation arises if k is not fixed but determined by some stopping criterion, or if a semi-iteration is applied as a secondary process. In these cases, Φ_k is nonlinear; hence, the suitability of Φ_k as the basic iteration of a semi-iterative method is questionable. For a discussion of this problem, we refer to Golub–Overton [156] and Axelsson–Vassilevski [18, 19].

⁶ Under additional conditions (properties similar to M- or H-matrices) Frommer–Szyld [143] prove convergence for all k as in the symmetric case discussed in §5.5.4.

5.5.4 Analysis in the Positive Definite Case

In the following, let Φ_A and Φ_B be positive definite iterations. Assume that the matrices in (5.19) and (5.20) satisfy

$$A > 0, \quad B > 0, \quad C > 0. \quad (5.23a)$$

Lemma 5.37. *Assume (5.23a). The iteration Φ_B converges if and only if*

$$0 < B < 2C. \quad (5.23b)$$

Under this assumption, the composed iteration Φ_k defined in (5.18) is also positive definite for all $k \in \mathbb{N}$.

Proof. According to (5.21c), Φ_k has the first normal form

$$x^{m+1} = M_k x^m + N_k b \quad \text{with } N_k = (I - M_B^k)B^{-1}.$$

We have to show $W_k > 0$ for the matrix of the third normal form of Φ_k :

$$W_k(x^m - x^{m+1}) = Ax^m - b. \quad (5.24a)$$

By Remark 3.34a, (5.23b) is equivalent to the convergence of Φ_B . $\rho(M_B) < 1$ implies that $I - M_B^k$ is regular; hence, the matrix $W_k = N_k^{-1} = B(I - M_B^k)^{-1}$ exists. The representation

$$M_B = I - C^{-1}B = B^{-1/2}(I - B^{1/2}C^{-1}B^{1/2})B^{1/2}$$

proves the symmetry of

$$W_k = B^{1/2}[I - (I - B^{1/2}C^{-1}B^{1/2})^k]^{-1}B^{1/2} = W_k^H. \quad (5.24b)$$

Since $\rho(M_B) < 1$ implies $-I < I - B^{1/2}C^{-1}B^{1/2} < 1$, this inequality, together with $[I - (I - B^{1/2}C^{-1}B^{1/2})^k] > 0$ (because of $k > 0$), yields positive definiteness of W_k ,

$$W_k > 0$$

and of Φ_k . □

It is not true that convergence of Φ_A and Φ_B implies convergence of Φ_k , but convergence can always be achieved by a suitable damping. In the following, we assume the inequalities

$$\gamma B \leq A \leq \Gamma B \quad \text{with } 0 < \gamma \leq \Gamma, \quad (5.25a)$$

$$\delta C \leq B \leq \Delta C \quad \text{with } 0 < \delta \leq \Delta \quad (5.25b)$$

The spectrum of $B^{1/2}C^{-1}B^{1/2}$ lies in $[\delta, \Delta]$ (cf. (C.3b,e)), and therefore we obtain $\sigma(I - B^{1/2}C^{-1}B^{1/2}) \subset [1 - \Delta, 1 - \delta]$. The spectrum of $I - (I - B^{1/2}C^{-1}B^{1/2})^k$

is contained in the interval $[\underline{\beta}, \bar{\beta}]$ with

$$\begin{aligned}\underline{\beta} &:= \begin{cases} 1 - (1 - \delta)^k & \text{for odd } k, \\ 1 - \max\{(1 - \Delta)^k, (1 - \delta)^k\} & \text{for even } k, \end{cases} \\ \bar{\beta} &:= \begin{cases} 1 - (1 - \Delta)^k & \text{for odd } k \text{ or } \Delta < 1, \\ 1 & \text{for even } k \text{ or } \Delta \geq 1. \end{cases}\end{aligned}\quad (5.25c)$$

Equation (5.24b) proves that

$$\underline{\beta} W_k \leq B \leq \bar{\beta} W_k.$$

By (5.25a) we obtain the following lemma.

Lemma 5.38. *The inclusions (5.25a,b) imply (5.26) for W_k in (5.24a):*

$$\gamma_k W_k \leq A \leq \Gamma_k W_k \quad \text{with } \gamma_k := \gamma \underline{\beta}, \quad \Gamma_k := \Gamma \bar{\beta}. \quad (5.26)$$

Let $\delta, \Delta, \gamma, \Gamma$ be the optimal bounds in (5.25a,b). Then (5.27) holds:

$$\kappa(W_k^{-1} A) = \frac{\Gamma_k}{\gamma_k} = \frac{\Gamma}{\gamma} \frac{\bar{\beta}}{\underline{\beta}} = \frac{\bar{\beta}}{\underline{\beta}} \kappa(B^{-1} A). \quad (5.27)$$

Analysing the iteration Φ_B separately, we obtain the optimal damping parameter

$$\Theta_B = 2/(\delta + \Delta) \quad (\text{cf. (3.25)}). \quad (5.28)$$

The matrix of the third normal form of the damped iteration Φ_{B,Θ_B} is $\Theta_B^{-1} C$ instead of C and leads to the bounds $\delta\Theta_B$ and $\Delta\Theta_B$ instead of δ and Δ . This scaling changes the ratio $\bar{\beta}/\underline{\beta}$. The next exercise discusses the factor Θ_B minimising the spectral condition number (5.27).

Exercise 5.39. Prove that (a) for even k , the parameter Θ_B in (5.28) yields the optimal spectral condition number (5.27). For odd k , however, the minimum of $\kappa(W_k^{-1} A)$ is attained at a value of Θ_B in the open interval

$$1/\Delta < \Theta_B < 2/(\delta + \Delta).$$

- (b) For $k = 1$, $\kappa(W_1^{-1} A) = \kappa(B^{-1} A)\kappa(C^{-1} B)$ holds independently of Θ_B .
- (c) For $k = 3$, the optimal value is

$$\Theta_B = 3/[\delta + \Delta + \sqrt{\Delta(\Delta - \delta) + \delta^2}].$$

Since, in the case of an odd $k \neq 3$, the optimal Θ_B is not explicitly described, in the following we always use (5.28).

An analysis of multiple secondary iterations similar to the example in §5.5.2 will be given in Remark 12.27.

5.5.5 Estimate of the Amount of Work

An important question concerns the number k of secondary iterations: for which k is the effective amount of work of Φ_k as favourable as possible. A trivial statement is given next.

Remark 5.40. The effective amount of work $\text{Eff}(\Phi_k)$ is minimal for some finite k because $\text{Eff}(\Phi_k) = \mathcal{O}(k)$ for $k \rightarrow \infty$.

Proof. $2C_A + 1$ operations are needed for (5.18a) and (5.18c) (concerning C_A , see §2.3.1). Let C_B be the amount of work for one secondary iteration step. Then the cost factor for Φ_k is

$$C_k = C' + kC'' \quad \text{with } C' := 2 + 1/C_A, \quad C'' := C_B/C_A. \quad (5.29)$$

C_k increases for $k \rightarrow \infty$ as $\mathcal{O}(k)$, while at best the convergence rate of Φ_k tends to that of Φ_A . \square

We assume that $\kappa := \kappa(C^{-1}B) = \Delta/\delta \gg 1$ holds for the spectral condition number corresponding to the method Φ_B (i.e., Φ_B is not a very fast iteration). Furthermore, let Φ_B be already optimally damped, i.e., $\Theta_B = 2/(\delta + \Delta) = 1$ (cf. (5.28)). Then

$$-(1 - \Delta) = 1 - \delta = \frac{\kappa - 1}{\kappa + 1} = 1 - \frac{1}{\kappa} + \mathcal{O}(\kappa^{-2})$$

proves that

$$(1 - \delta)^k = 1 - \frac{k}{\kappa} + \mathcal{O}\left(\left(\frac{k}{\kappa}\right)^2\right), \quad (1 - \Delta)^k = (-1)^k(1 - \delta)^k.$$

From (5.25c) we obtain the following expansion for $k \leq \kappa$:

$$\frac{\bar{\beta}}{\underline{\beta}} = \begin{cases} \kappa/k + \mathcal{O}(1) & \text{for odd } k, \\ \kappa/(2k) + \mathcal{O}(1) & \text{for even } k. \end{cases} \quad (5.30a)$$

First, we consider the case in which Φ_k serves as a (stationary) iterative method. Then the convergence rate

$$\rho(\Phi_k) = \frac{\kappa(W_k^{-1}A) - 1}{\kappa(W_k^{-1}A) + 1} \approx 1 - \frac{2}{\kappa(W_k^{-1}A)} = 1 - 2 \frac{\gamma\bar{\beta}}{\Gamma\bar{\beta}} \approx 1 - 2\alpha k \quad (5.30b)$$

holds for optimal damping (cf. (5.28)) with $\alpha = \frac{\gamma}{k\Gamma}$ for odd and $\alpha = \frac{2\gamma}{k\Gamma}$ for even k . From $-\log \rho(\Phi_k) \approx 2\alpha k$ and (5.29), we obtain

$$\text{Eff}(\Phi_k) \approx \frac{C' + kC''}{2\alpha k} = \frac{\frac{1}{k}C' + C''}{2\alpha}.$$

Conclusion 5.41. Initially, the effective amount of work of the iterative method Φ_k decreases with k until, for $k \approx \kappa$, the asymptotic representations (5.30a,b) lose their validity. Because of the better value $\bar{\beta}/\beta$ one should prefer even numbers k .

A different situation arises when the (positive definite) iteration Φ_k is used as the basic iteration of the Chebyshev method (see §8.3.4), since then the asymptotic rate is given by (8.32a) instead of (5.30b). According to (8.36), the expansion

$$\text{Eff}_{\text{semi-it}}(\Phi_k) \approx \left[\frac{1}{2}(C' + kC'') + \frac{3}{C_A} \right] \sqrt{\frac{\Gamma}{\gamma} \frac{\bar{\beta}}{\beta}} \approx \frac{\frac{1}{2}(C' + kC'') + \frac{3}{C_A}}{\sqrt{\alpha k}} \quad (5.30c)$$

holds with the same α as in (5.30b).

Conclusion 5.42. The semi-iterative effective amount of work (5.30c) becomes minimal for the even number k next to the value k_0 in (5.30d):

$$k_0 = \left(\frac{C'}{2} + \frac{3}{C_A} \right) / \frac{C''}{2} = \left(C' + \frac{6}{C_A} \right) / C'' = \left(2 + \frac{7}{C_A} \right) / C''. \quad (5.30d)$$

Since $k_0 < 3$ is realistic, $k = 2$ is the optimum.

5.5.6 Numerical Examples

We solve the introductory example (5.16b) for $a = 1$ by choosing B as the matrix of the Poisson model problem. The solution of the auxiliary equation $B\delta = d$ is approximated by the ADI method (see §8.5) with the cycle length $4 = 2^p$ ($p = 2$), where $k = 4$ is also chosen in (5.18b₂). Note that the ADI method is not directly applicable to the original problem since A is not a five-point matrix and not symmetric. The composed method Φ_4 is performed without damping for $h = 1/32$ and $1/64$. The exact solution is again

$$u = x^2 + y^2 \quad \text{for } f = 2x - 4 \quad \text{in (5.16a).}$$

The results for the different step widths show a rate of ≈ 0.46 . Note, however, that one Φ_4 step consists of $k = 4$ ADI steps. Therefore, $0.46^{1/4} = 0.82$ gives a better idea of the rate. The reader may determine the effective amount of work. Table 5.1 contains the Euclidean norm of the errors $\|x^m - x\|_2$ (m : number of outer iterations Φ_k for $k = 4$) and their ratios

$$\rho_{m,m-1} = \|x^m - x\|_2 / \|x^{m-1} - x\|_2.$$

m	$\ x^m - x\ _2$	$\rho_{m,m-1}$	m	$\ x^m - x\ _2$	$\rho_{m,m-1}$
1	$3.06_{10^{-2}}$	$4.093_{10^{-2}}$	1	$2.78_{10^{-2}}$	$3.628_{10^{-2}}$
2	$3.79_{10^{-3}}$	$1.239_{10^{-1}}$	2	$4.36_{10^{-3}}$	$1.565_{10^{-1}}$
3	$9.80_{10^{-4}}$	$2.582_{10^{-1}}$	3	$1.57_{10^{-4}}$	$3.611_{10^{-1}}$
4	$3.40_{10^{-4}}$	$3.473_{10^{-1}}$	4	$6.68_{10^{-4}}$	$4.243_{10^{-1}}$
5	$1.33_{10^{-4}}$	$3.927_{10^{-1}}$	5	$2.90_{10^{-4}}$	$4.340_{10^{-1}}$
6	$5.90_{10^{-5}}$	$4.415_{10^{-1}}$	6	$1.32_{10^{-5}}$	$4.576_{10^{-1}}$
7	$2.61_{10^{-5}}$	$4.427_{10^{-1}}$	7	$5.99_{10^{-5}}$	$4.512_{10^{-1}}$
8	$1.19_{10^{-5}}$	$4.557_{10^{-1}}$	8	$2.73_{10^{-5}}$	$4.568_{10^{-1}}$
9	$5.53_{10^{-6}}$	$4.640_{10^{-1}}$	9	$1.25_{10^{-6}}$	$4.566_{10^{-1}}$
10	$2.56_{10^{-6}}$	$4.640_{10^{-1}}$	10	$5.73_{10^{-6}}$	$4.581_{10^{-1}}$

Table 5.1 Φ_4 for $h = 1/32$ (left) and $h = 1/64$ (right).

5.6 Transformations

A general method producing linear iterations are transformations. In particular, any consistent linear iteration can be obtained by a suitable transformation applied to the Richardson iteration. More precisely, the left-transformation, the right-transformation, and the both-sided transformation can be distinguished.

Transformations are also called preconditioners when they improve the iteration.

5.6.1 Left Transformation

5.6.1.1 Definitions

The system of equations

$$Ax = b \quad (5.31)$$

can be transferred by multiplying from the left by a regular matrix T_ℓ into the equivalent system

$$T_\ell A x = T_\ell b. \quad (5.32)$$

We may regard (5.32) as a new system of equations:

$$\hat{A}x = \hat{b} \quad \text{with} \quad \hat{A} := T_\ell A, \quad \hat{b} := T_\ell b,$$

to which we now apply iterative methods. Here, two approaches are imaginable.

The naive approach computes \hat{A} and \hat{b} by $T_\ell A$ and $T_\ell b$ and uses these quantities instead of A and b . If T_ℓ is a diagonal matrix, the transition to (5.32) means that the equations in (5.31) are suitably scaled. Otherwise, this approach may be unfavourable compared with the next approach.

Again, dependence on the system matrix A is denoted explicitly by

$$\varPhi(x, b, A), \quad N[A], \quad W[A], \quad \text{etc.}$$

Applying the iteration

$$x^{m+1} = \varPhi(x^m, b, A) = x^m - N[A](Ax^m - b) = x^m - W[A]^{-1}(Ax^m - b)$$

to (\hat{A}, \hat{b}) instead of (A, b) , we obtain

$$x^{m+1} = x^m - N[\hat{A}](\hat{A}x^m - \hat{b}) = x^m - W[\hat{A}]^{-1}(\hat{A}x^m - \hat{b}). \quad (5.33)$$

The definition of \hat{A} and \hat{b} yields the representation $\hat{A}x^m - \hat{b} = T_\ell(Ax^m - b)$. Hence, (5.33) can be rewritten as

$$x^{m+1} = \hat{\varPhi}(x^m, b, A) = x^m - N[T_\ell A]T_\ell(Ax^m - b) \quad (5.34)$$

and defines a new iteration $\hat{\varPhi} \in \mathcal{L}$.

Remark 5.43. To perform the iteration (5.34), one needs

- (i) computing the defect $d := Ax - b$ (using the original quantities A, b),
- (ii) multiplying T_ℓ by a vector,
- (iii) a method for solving $W[T_\ell A]\delta = d$ or the explicit multiplication by $N[T_\ell A]$.

Note that we do not need the matrix $\hat{A} = T_\ell A$ when we proceed according to Remark 5.43. \hat{A} appears only indirectly in $W[T_\ell A]$. If, for instance, the Jacobi method is the underlying iteration Φ , one has only to evaluate the diagonal entries of $\hat{A} = T_\ell A$ since $W[T_\ell A] = \text{diag}\{T_\ell A\}$.

Defining

$$\hat{N}[A] := N[T_\ell A] T_\ell, \quad \hat{W}[A] := T_\ell^{-1} W[T_\ell A], \quad (5.35)$$

we generate a new iterative method $\hat{\Phi}$ by $A \mapsto \hat{N}[A]$ which is identical to (5.34):

$$x^{m+1} = x^m - \hat{N}[A](Ax^m - b).$$

This shows that the left transformation T_ℓ by (5.35) is able to generate a new iteration $\hat{\Phi}$ from Φ , which we denote by

$$\hat{\Phi} = \Phi \circ T_\ell \quad (5.36)$$

to express that we apply the iteration Φ after transforming the system by T_ℓ .

5.6.1.2 Examples

Proposition 5.44. Any $\Phi \in \mathcal{L}$ can be regarded as the Richardson iteration with $\Theta = 1$ applied to the transformed system

$$\hat{A}x = \hat{b} \quad \text{with } \hat{A} := NA, \hat{b} := Nb,$$

where $N = N_\Phi[A]$ is the matrix of the second normal form of $\Phi(\cdot, \cdot, A)$. With the notation introduced above, this statement can be reformulated as

$$\Phi = \Phi_1^{\text{Rich}} \circ N_\Phi.$$

One advantage of generating iterations $\hat{\Phi}$ by transformations is that no new convergence analysis is necessary, provided that the convergence of $\Phi(\cdot, \cdot, T_\ell A)$ is known.

Remark 5.45. Let $\hat{\Phi} = \Phi \circ T_\ell$. The convergence properties of $\hat{\Phi}$ (applied to the matrix A) are identical to those of Φ applied to $T_\ell A$. Care is only advisable for the interpretation of convergence statements with respect to a norm depending on A (e.g., the energy norm⁷).

⁷ A positive definite matrix A becomes $T_\ell A$ after the transformation. If $T_\ell A$ is not positive definite, $\|\cdot\|_{T_\ell A}$ is meaningless and does not define a norm.

For the purpose of illustration, we choose the left transformation by $T_\ell = A^H$. Then, (5.34) becomes

$$x^{m+1} = x^m - N[A^H A] A^H (Ax^m - b). \quad (5.37a)$$

Since $A^H A$ is positive definite for regular A , almost all methods mentioned above can be applied to $A^H A$. The Richardson iteration is characterised by $N_\Theta^{\text{Rich}}[A^H A] = \Theta I$. Choosing the Richardson iteration with the optimal damping factor

$$\Theta_{\text{opt}} = \frac{2}{\lambda_{\max} + \lambda_{\min}} \quad \text{with} \quad \begin{cases} \lambda_{\max} := \lambda_{\max}(A^H A) = \|A\|_2^2, \\ \lambda_{\min} := \lambda_{\min}(A^H A) = \|A^{-1}\|_2^{-2}, \end{cases}$$

we obtain the Landweber iteration (cf. Remark 5.17):

$$x^{m+1} = x^m - \Theta_{\text{opt}} A^H (Ax^m - b). \quad (5.37b)$$

For the new method (5.37b), we draw the following conclusion from the convergence properties of the Richardson method (cf. §3.5.1).

Remark 5.46. The Landweber iteration $\Phi_{\Theta_{\text{opt}}}^{\text{Landw}} = \Phi_{\Theta_{\text{opt}}}^{\text{Rich}} \circ A^H$ defined by (5.37b) converges for all regular matrices A with the rate

$$\rho(I - \Theta_{\text{opt}} A^H A) = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} = \frac{\text{cond}_2(A)^2 - 1}{\text{cond}_2(A)^2 + 1}.$$

5.6.1.3 Rules for the Left Transformation

The following statements are easy to check or already stated. The standard assumption is that T_ℓ be regular. Nevertheless, most of the statements remain valid for singular T_ℓ . Then the resulting matrix $N_{\Phi \circ T_\ell}$ is also singular so that the inverse $W_{\Phi \circ T_\ell} = N_{\Phi \circ T_\ell}^{-1}$ does not exist. Therefore all statements involving $W_{\Phi \circ T_\ell}$ have to be omitted for singular T_ℓ .

Proposition 5.47. If $\Phi \in \mathcal{L}$, also $\Phi \circ T_\ell \in \mathcal{L}$. The left transformation satisfies the following rules:

$$\mathfrak{D}(\Phi \circ T_\ell) = \{A \in \mathbb{K}^{I \times I} : T_\ell A \in \mathfrak{D}(\Phi)\}, \quad (5.38a)$$

$$N_{\Phi \circ T_\ell}[A] = N_\Phi[T_\ell A] T_\ell, \quad (5.38b)$$

$$W_{\Phi \circ T_\ell}[A] = T_\ell^{-1} W_\Phi[T_\ell A], \quad (5.38c)$$

$$M_{\Phi \circ T_\ell}[A] = I - N_{\Phi \circ T_\ell}[A] A = M_\Phi[T_\ell A], \quad (5.38d)$$

$$\rho(M_{\Phi \circ T_\ell}[A]) = \rho(M_\Phi[T_\ell A]), \quad (5.38e)$$

$$(\Phi \circ T_1) \circ T_2 = \Phi \circ (T_1 T_2), \quad (5.38f)$$

$$\Psi = \Phi \circ T_\ell \iff \Phi = \Psi \circ T_\ell^{-1}. \quad (5.38g)$$

The neutral element is the identity matrix: $T_\ell = I$.

5.6.2 Right Transformation

5.6.2.1 Definitions

The unknown vector $x \in \mathbb{K}^I$ in $Ax = b$ can be substituted by

$$x = T_r \hat{x}, \quad (5.39)$$

where T_r is a regular matrix. Inserting (5.39) into $Ax = b$, we obtain the right-sided transformed equation

$$A T_r \hat{x} = b.$$

First, we discuss the naive approach: compute the matrix \hat{A} in

$$\hat{A} \hat{x} = b \quad \text{with } \hat{A} := AT_r \quad (5.40)$$

in a preliminary phase and then apply a linear iteration directly to (5.40). Finally, we obtain $x = T_r \hat{x}$ from the (approximation of the) solution \hat{x} by (5.39).

Applying $\Phi \in \mathcal{L}$ to the system (5.40), we can rewrite the iteration $\Phi(\cdot, \cdot, \hat{A})$ as

$$\hat{x}^{m+1} = \hat{x}^m - N_\Phi[AT_r](AT_r\hat{x}^m - b).$$

Introducing

$$x^m := T_r \hat{x}^m,$$

we obtain the iteration

$$x^{m+1} = x^m - T_r N_\Phi[AT_r](Ax^m - b). \quad (5.41)$$

This is a newly generated iteration $\hat{\Phi}$ for solving the original equation $Ax = b$ with the following matrices of the second and third normal forms:

$$\hat{N}[A] := T_r N[AT_r], \quad \hat{W}[A] := W[AT_r] T_r^{-1}.$$

In analogy to (5.36), $\hat{\Phi}$ is denoted by

$$\hat{\Phi} = T_r \circ \Phi.$$

5.6.2.2 Examples

Remark 5.48. The convergence rate of $\hat{\Phi} = T_r \circ \Phi$ applied to the matrix A is identical to the convergence rate of Φ applied to AT_r . Convergence properties of Φ referring to a norm of $e^m = x^m - x$ carry over to the corresponding properties of $\hat{\Phi}$ with respect to the norm of $T_r^{-1}e^m = T_r^{-1}(x^m - x)$.

The analogue to (5.37a) is the right transformation $T_r = A^H$, leading to

$$x^{m+1} = x^m - A^H N[AA^H](Ax^m - b) \quad (5.42)$$

Choosing the Richardson method $\Phi = \Phi^{\text{Rich}}$ so that $N_\Theta^{\text{Rich}}[AA^H] = \Theta I$, the product $\hat{\Phi} = A^H \circ \Phi$ generates again the method (5.37b).⁸

More generally, (5.42) proves the following statement.

Remark 5.49. Let the matrix A be regular, choose $T_r = A^H$ and set $\hat{\Phi} = T_r \circ \Phi$. If $\Phi \in \mathcal{L}_{\text{sym}}$, then $M_{\hat{\Phi}} = M_\Phi^H$. If $\Phi \in \mathcal{L}_{\text{pos}}$, then $\hat{\Phi} \in \mathcal{L}_{>0}$ holds, i.e., $N_{\hat{\Phi}} A > 0$.

5.6.2.3 Rules for the Right Transformation

The analogue of Proposition 5.47 reads as follows.

Proposition 5.50. *If $\Phi \in \mathcal{L}$, also $T_r \circ \Phi \in \mathcal{L}$. The right transformation satisfies the following rules:*

$$\mathfrak{D}(T_r \circ \Phi) = \{A \in \mathbb{K}^{I \times I} : AT_r \in \mathfrak{D}(\Phi)\}, \quad (5.43a)$$

$$N_{T_r \circ \Phi}[A] = T_r N_\Phi[AT_r], \quad (5.43b)$$

$$W_{T_r \circ \Phi}[A] = W_\Phi[AT_r] T_r^{-1}, \quad (5.43c)$$

$$M_{T_r \circ \Phi}[A] = I - N_{T_r \circ \Phi}[A]A = T_r M_\Phi[AT_r] T_r^{-1}, \quad (5.43d)$$

$$\rho(M_{T_r \circ \Phi}[A]) = \rho(M_\Phi[AT_r]), \quad (5.43e)$$

$$T_2 \circ (T_1 \circ \Phi) = (T_2 T_1) \circ \Phi, \quad (5.43f)$$

$$\Psi = T_r \circ \Phi \iff \Phi = T_r^{-1} \circ \Psi. \quad (5.43g)$$

Note that (5.43d) is a bit different from (5.38d), but the similarity transformation by T_r does not change the statement (5.43e).

5.6.3 Kaczmarz Iteration

In 1937, Kaczmarz [230]⁹ described an iteration for which he could prove convergence for all regular matrices A .

5.6.3.1 Original Formulation

In the original formulation, the projections

⁸ The optimal value of Θ_{opt} in $N_\Theta^{\text{Rich}}[AA^H]$ depends on the extreme eigenvalues of AA^H (cf. Theorem 3.23). Because of $\sigma(AA^H) = \sigma(A^H A)$ according to Theorem A.10, the same Θ_{opt} holds for $N_\Theta^{\text{Rich}}[A^H A]$.

⁹ An English translation of the original German paper can be found in [231].

$$x \longmapsto P_i(x, b) := x - \mathbf{a}_i \langle Ax - b, \mathbf{e}_i \rangle / \langle \mathbf{a}_i, \mathbf{a}_i \rangle \quad (1 \leq i \leq n) \quad (5.44)$$

onto the hyperplanes $\{x \in \mathbb{K}^I : \langle Ax - b, \mathbf{e}_i \rangle = 0\}$ are used, where \mathbf{e}_i is the i -th unit vector and $\mathbf{a}_i = A^H \mathbf{e}_i$ the transposed i -th row of the matrix A . Using $\langle A\mathbf{a}_i, \mathbf{e}_i \rangle = \langle \mathbf{a}_i, A^H \mathbf{e}_i \rangle = \langle \mathbf{a}_i, \mathbf{a}_i \rangle$, one learns that $x' = P_i(x, b)$ lies in the plane $\langle Ax - b, \mathbf{e}_i \rangle = 0$. On the other hand, P_i describes a linear iteration, which is a projection in the sense of Definition 5.12. The complete iteration is the product of all projections in the succession $i = 1, \dots, n$:

$$\varPhi^{\text{Kacz}} := P_n \circ P_{n-1} \circ \dots \circ P_1.$$

We note that Kaczmarz' method admits interesting applications to overdetermined systems of equations (cf. Tanabe [361]). The Kaczmarz method is also applied to problems from tomography and is then termed the *algebraic reconstruction technique* and abbreviated as *ART* (cf. Natterer [287, §V.3–4]). However, note that in the latter application one is not at all interested in the solution of the (normal) equation since the underlying problems are ill-posed.¹⁰ Few iteration steps of the method are used as a kind of filter.

5.6.3.2 Interpretation as Gauss–Seidel Iteration

The right transformation $T_r = A^H$ generates the system

$$AA^H \hat{x} = b \quad (5.45)$$

for \hat{x} with $x = A^H \hat{x}$. Using the \hat{x} variables, we can rewrite the projection (5.44) as $\hat{x} = A^{-H} x \longmapsto \hat{P}_i(\hat{x}, b)$ with

$$\hat{P}_i(\hat{x}, b) := A^{-H} P_i(A^H \hat{x}, b) = \hat{x} - \mathbf{e}_i \langle AA^H \hat{x} - b, \mathbf{e}_i \rangle / \langle \mathbf{a}_i, \mathbf{a}_i \rangle,$$

since $A^{-H} \mathbf{a}_i = A^{-H} A^H \mathbf{e}_i = \mathbf{e}_i$. \hat{P}_i is the projection onto $\langle AA^H \hat{x} - b, \mathbf{e}_i \rangle = 0$, i.e., $\hat{x} \mapsto \hat{x}' := \hat{P}_i(\hat{x}, b)$ yields the solution of the scalar equation $(AA^H \hat{x})_i = b_i$ with respect to \hat{x}_i . Therefore, performing $\hat{x} \mapsto \hat{P}_i(\hat{x}, b)$ for $i = 1, \dots, n$ executes the Gauss–Seidel method for the system (5.45). The denominators $\langle \mathbf{a}_i, \mathbf{a}_i \rangle = \langle AA^H \mathbf{e}_i, \mathbf{e}_i \rangle$ represent the diagonal entries of AA^H . Hence, the following lemma is proved.

Lemma 5.51. *The Kaczmarz iteration for solving $Ax = b$ coincides with the Gauss–Seidel iteration for $AA^H \hat{x} = b$:*

$$\varPhi^{\text{Kacz}} = A^H \circ \varPhi^{\text{GS}} \in \mathcal{L}_{>0}.$$

¹⁰ This fact implies that here the theory of iterative methods (concerning converge and convergence speed) does not matter. Instead, x^m has to satisfy a certain ‘smoothness property’. Therefore one is looking for some x^m with still enough smoothness and not too large residual error $Ax^m - b$.

Since AA^H is positive definite for regular A , the theorem of Ostrowski (Theorem 3.41) yields convergence.

Theorem 5.52. *The Kaczmarz method converges for all regular A .*

To obtain a quantitative statement, one has to determine and estimate the constants γ and Γ in (3.46a,b) for the decomposition $AA^H = D - E - F$.

Exercise 5.53. Instead of the right transformation, one may also choose a left one by $T_\ell = A^H$. Prove that one step of the Gauss–Seidel iteration applied to $A^H Ax = A^H b$ has the form

$$\text{for } i := 1 \text{ to } n \text{ do } x := x - \mathbf{e}_i \langle Ax - b, \mathbf{a}_i \rangle / \langle \mathbf{a}_i, \mathbf{a}_i \rangle,$$

where, different from (5.44), \mathbf{a}_i represents the i -th column of A : $\mathbf{a}_i := A\mathbf{e}_i$.

5.6.3.3 Numerical Examples

The convergence rates for the Poisson model case are given in Table 5.2. The ratios $\rho_{m,m-1}$ behave as $1 - (\alpha h)^4$ with $2.5 \leq \alpha \leq 3$. The convergence order $\tau = 4$ makes the Kaczmarz method very unattractive.

m	$\ e^m\ _2$	$\rho_{m,m-1}$	m	$\ e^m\ _2$	$\rho_{m,m-1}$
10	0.465	0.984542	10	0.624	0.993655
30	0.371	0.990501	30	0.575	0.996921
50	0.309	0.991074	50	0.546	0.997806
70	0.258	0.990876	70	0.525	0.998272
80	0.235	0.990790	80	0.517	0.998434
90	0.214	0.990728	90	0.509	0.998564
100	0.195	0.990687	100	0.502	0.998671

Table 5.2 Results of the Kaczmarz iteration for the Poisson model case for $h = 1/8$ (left) and $h = 1/16$ (right).

5.6.4 Cimmoni Iteration

In 1938, Cimmoni [97] described the following iteration (see also Benzi [41]):

$$\Phi(x, b, A) := x - \vartheta A^H D^{-1} (Ax - b) \quad \text{with}$$

$$D := \text{diag}\{(AA^H)_{ii}/\mu_i : i \in I\}, \quad \vartheta := 2 / \sum_{i \in I} \mu_i,$$

where $\mu_i > 0$ are some weights. For simplicity, we set $\mu_i := 1$ and obtain the damping factor $\vartheta = 2/\#I$ in

$$\Phi(x, b, A) := x - \vartheta A^H D^{-1} (Ax - b) \quad \text{with} \quad D := \text{diag}\{AA^H\}.$$

Remark 5.54. Cimmoni's iteration is closely related to the damped Jacobi iteration applied to AA^H . In fact, $\Phi_\vartheta^{\text{Cimm}} = A^H \circ \Phi_\vartheta^{\text{Jac}} \in \mathcal{L}_{>0}$ holds. The iteration matrix $M_\vartheta^{\text{Cimm}}[A] = I - \vartheta A^H D^{-1} A$ is similar to

$$A^{-H} M_\vartheta^{\text{Cimm}}[A] A^H = I - \vartheta D^{-1} AA^H = M_\vartheta^{\text{Jac}}[AA^H].$$

Proposition 5.55. *Cimmoni's iteration converges for the original choice of $\vartheta = \frac{2}{\#I}$.*

Proof. Let \mathbf{a}_i be the i -th row of A (column of A^H). Then $D = \text{diag}\{\|\mathbf{a}_i\|_2^2 : i \in I\}$ can be written as $D = \Lambda^2$ with $\Lambda = \text{diag}\{\|\mathbf{a}_i\|_2\}$. Note that $A^H D^{-1} A = B^H B$ with $B = \Lambda^{-1} A$. The rows of B are $\mathbf{b}_i = \mathbf{a}_i / \|\mathbf{a}_i\|_2$ so that $\|\mathbf{b}_i\|_2 = 1$.

Let $x \in \mathbb{K}^I$ be normed: $\|x\|_2 = 1$, and set $y := Bx$. The components of y are the Euclidean scalar products $y_i = \langle \mathbf{b}_i, x \rangle$ so that $|y_i| \leq \|\mathbf{b}_i\|_2 \|x\|_2 = 1$. Hence $\|y\|_2^2 \leq \#I$ follows. Since x with $\|x\|_2 = 1$ is arbitrary, $\|B\|_2 \leq \sqrt{\#I}$ is proved.

Assume that there is some x with $\|Bx\|_2 = \sqrt{\#I}$ and $\|x\|_2 = 1$. The previous proof shows that $|y_i| = 1$ must hold for all $i \in I$. However, the Schwarz inequality only holds with an equal sign if all \mathbf{b}_i are multiples of x . This is a contradiction to the regularity of A and B . Hence, $\|B\|_2 < \sqrt{\#I}$ is proved. \square

$\lambda_{\max}(D^{-1}A^H A) = \lambda_{\max}(A^H D^{-1} A) = \|B^H B\|_2 \leq \|B\|_2^2 < \#I$ implies that $\vartheta = 2/\#I$ satisfies the convergence condition (6.9): $0 < \vartheta < 2/\lambda_{\max}$ for the damped Jacobi iteration.

5.6.5 Two-Sided Transformation

5.6.5.1 Definition and Properties

Applying transformations by T_ℓ from the left and T_r from the right, we obtain the two-sided transformed iteration $\hat{\Phi} = T_r \circ \Phi \circ T_\ell$ generated by

$$\begin{aligned}\hat{N}[A] &:= T_r N[T_\ell A T_r] T_\ell, \\ x^{m+1} &= x^m - T_r N[T_\ell A T_r] T_\ell (Ax^m - b).\end{aligned}\tag{5.46}$$

Exercise 5.56. Prove that $(T_r \circ \Phi) \circ T_\ell = T_r \circ (\Phi \circ T_\ell)$.

Concerning convergence properties of $T_r \circ \Phi \circ T_\ell$, the same statements apply as for $T_r \circ \Phi$ in Remark 5.48.

One-sided transformations often destroy the symmetry of the underlying iteration Φ . Choosing two-sided transformations satisfying $T_r = T_\ell^H$, we can save this property.

Proposition 5.57. (a) If $\Phi \in \mathcal{L}_{\text{sym}}$, then also $T^H \circ \Phi \circ T \in \mathcal{L}_{\text{sym}}$.

(b) Let $\Phi \in \mathcal{L}_{\text{pos}}$ and T be regular. Then $T^H \circ \Phi \circ T \in \mathcal{L}_{\text{pos}}$ also holds.

Proof. The matrix of the second normal form of $\hat{\Phi} := T^H \circ \Phi \circ T$ is

$$\hat{N}[A] = T^H N[T A T^H] T.$$

Concerning the symmetry for $\hat{\Phi}$ we have to show that $\hat{N}[A] = \hat{N}[A^H]^H$. Note that

$$\hat{N}[A^H]^H = (T^H N[TA^H T^H] T)^H = T^H N[TA^H T^H]^H T = T^H N[(TAT^H)^H]^H T.$$

Symmetry of Φ implies that $N[(TAT^H)^H]^H = N[TAT^H]$; hence

$$(T^H N[TA^H T^H] T)^H = T^H N[(TAT^H)] T = \hat{N}[A]$$

proves the symmetry of $\hat{\Phi}$.

Assume $A > 0$. Since the transformation T_ℓ is regular, $TAT^H > 0$ also holds (cf. (C.3a)). Positive definiteness of Φ implies $N[(TAT^H)] > 0$. Applying (C.3a) again, we obtain $\hat{N}[A] > 0$, i.e., $\hat{\Phi}$ is positive definite. \square

For completeness, we list the properties of the two-sided transformation corresponding to Propositions 5.47 and 5.50 including the new results of Proposition 5.57.

Proposition 5.58. *If $\Phi \in \mathcal{L}$, also $\hat{\Phi} := T_r \circ \Phi \circ T_\ell \in \mathcal{L}$. The two-sided transformation satisfies the following rules:*

$$\mathfrak{D}(T_r \circ \Phi \circ T_\ell) = \{A \in \mathbb{K}^{I \times I} : T_\ell A T_r \in \mathfrak{D}(\Phi)\}, \quad (5.47a)$$

$$N_{\hat{\Phi}}[A] = T_r N_\Phi[T_\ell A T_r] T_\ell, \quad (5.47b)$$

$$W_{\hat{\Phi}}[A] = T_\ell^{-1} W_\Phi[T_\ell A T_r] T_r^{-1}, \quad (5.47c)$$

$$M_{\hat{\Phi}}[A] = I - N_{\hat{\Phi}}[A] A = T_r M_\Phi[T_\ell A T_r] T_r^{-1}, \quad (5.47d)$$

$$\rho(M_{\hat{\Phi}}[A]) = \rho(M_\Phi[T_\ell A T_r]), \quad (5.47e)$$

$$T'_2 \circ (T'_1 \circ \Phi \circ T''_1) \circ T''_2 = (T'_2 T'_1) \circ \Phi \circ (T''_1 T''_2), \quad (5.47f)$$

$$\Psi = T_r \circ \Phi \circ T_\ell \iff \Phi = T_r^{-1} \circ \Psi \circ T_\ell^{-1}. \quad (5.47g)$$

The combination with the other algebraic operations are listed below:

$$\vartheta \cdot (T_r \circ \Phi \circ T_\ell) = T_r \circ (\vartheta \cdot \Phi) \circ T_\ell, \quad (5.48a)$$

$$(T_r \circ \Phi \circ T_\ell)^* = T_\ell^H \circ \Phi^* \circ T_r^H, \quad (5.48b)$$

$$T_r \circ (\Phi + \Psi) \circ T_\ell = (T_r \circ \Phi \circ T_\ell) + (T_r \circ \Psi \circ T_\ell), \quad (5.48c)$$

$$T_r \circ (\Phi \circ \Psi) \circ T_\ell = (T_r \circ \Phi \circ T_\ell) \circ (T_r \circ \Psi \circ T_\ell). \quad (5.48d)$$

5.6.5.2 Invariance of Iterations

Definition 5.59. (a) $\Phi \in \mathcal{L}$ is called *invariant* with respect to a left transformation T_ℓ if

$$\Phi \circ T_\ell = \Phi.$$

Analogously, Φ is invariant with respect to a right transformation T_r if $T_r \circ \Phi = \Phi$.

(b) $\Phi \in \mathcal{L}$ is called *diagonally left-invariant* if $\Phi \circ T_\ell = \Phi$ holds for all diagonal matrices T_ℓ . The diagonally right-invariant iterations are defined analogously.

The diagonal (left- or right-) invariance is of practical importance. Several properties of the matrix A depend on a suitable scaling of A . If the vectors x and b of the system contain physical quantities of different nature, their values depend on the choice of physical units. Replacing millimetre by kilometre, we change the scaling of a part of x or b by a factor of 10^6 . Such a scaling changes the condition of the matrix. As the simplest example consider the matrix $A = I$ with $\text{cond}(I) = 1$. Scaling of $A = I$ by a diagonal matrix $D = \text{diag}\{d_i : i \in I\}$ yields D with $\text{cond}(D) = \max_i\{d_i\} / \min\{d_i\}$ which can become arbitrarily large. In several cases, an unfavourable scaling may have a very negative effect.

A suitable scaling is already a problem for the Gauss elimination since for a wrong scaling the column pivot choice may fail. A possible remedy might be an equilibration, i.e., the matrix is rescaled such that the entries are similar in size (cf. van der Sluis [369] and Skeel [342]).

Such difficulties do not occur if the linear iteration is diagonally invariant. In the case of a diagonal left-invariance, the results x^m do not change when $Ax = b$ is replaced by $DAx = Db$ for some diagonal matrix D . If the iteration is diagonally right-invariant, a scaling $A \mapsto AD$ does not influence the iterates.

Next, we need the *Hadamard product* $A \odot B$ of two matrices which is defined by the entries $(A \odot B)_{ij} = A_{ij}B_{ij}$ ($i, j \in I$).

Lemma 5.60. (a) Φ is invariant with respect to a left transformation T_ℓ if and only if

$$W_\Phi[T_\ell A] = T_\ell W_\Phi[A].$$

(b) Φ is invariant with respect to a right transformation T_r if and only if

$$W_\Phi[AT_r] = W_\Phi[A]T_r.$$

(c) Assume that $W_\Phi[A] = X \odot A$ for some $X \in \mathbb{K}^{I \times I}$. Then Φ is diagonally left- and right-invariant.

(d) Let both Φ and Ψ be invariant with respect to a left transformation T_ℓ . Then $\vartheta \cdot \Phi$, $\Phi + \Psi$ and $\Phi \circ \Psi$ also have this property. The analogous statement holds for the right transformation.

(e) If Φ is invariant with respect to a left transformation T , then Φ^* is invariant with respect to a right transformation T^H . Vice versa, if Φ is invariant with respect to a right transformation T , then Φ^* is invariant with respect to a left transformation T^H .

(f) If Φ is diagonally left- and right-invariant, then Φ^* is also.

Proof. The parts (a) and (b) follow from (5.47c). Concerning (c), note that $X \odot (D'AD'') = D'(X \odot A)D''$ holds for diagonal matrices D', D'' . Part (d) is trivial. The definition of W_{Φ^*} proves part (e). Part (f) follows from part (e). \square

Typically, the matrix X in Lemma 5.60c has entries being either 1 or 0. If, e.g., $X_{ij} = 1$ holds only for the diagonal entries $i = j$, then $X \odot A = \text{diag}\{A\}$ defines $W^{\text{Jac}}[A]$. In the case of Gauss–Seidel, $X_{ij} = 1$ holds only for $j \geq i$.

Conclusion 5.61. *The Jacobi, Gauss–Seidel, SOR, the symmetric Gauss–Seidel, and the SSOR iterations are diagonally left- and right-invariant.*

Proof. Concerning the Jacobi, Gauss–Seidel, SOR iterations, use Lemma 5.60c. The adjoint versions (backward Gauss–Seidel and SOR) share the same property because of Lemma 5.60f. The symmetric counterparts inherit the invariance because of Lemma 5.60d. \square

Exercise 5.62. Generalise the diagonal invariance to the case of block-diagonal matrices and show that Conclusion 5.61 can be generalised to the respective block versions of the iterations.

5.6.6 Similarity Transformation

In many cases, A and $N = N[A]$ are positive definite, but not the product NA . As a consequence, NA and the iteration matrix $M = I - NA$ are not even symmetric. At least for *theoretical* considerations it would be advantageous to apply a *similarity transformation* $M \mapsto S_T(M) := TMT^{-1}$ with either $T = A^{1/2}$ or $T = W^{1/2} = N^{-1/2}$. If $A, N > 0$, the latter transformations yield $S_T(M) > 0$.

$\check{M} := S_T(M)$ is the iteration matrix of the linear iteration

$$\begin{aligned}\check{x}^{m+1} &= \check{M} \check{x}^m + \check{N}b = \check{x}^m - \check{N}(\check{A}\check{x}^m - b) \\ &= \check{x}^m - \check{W}^{-1}(\check{A}\check{x}^m - b) \quad \text{with} \\ \check{M} &= TMT^{-1}, \quad \check{N} = TN, \quad \check{W} = WT^{-1}, \quad \check{A} = AT^{-1}.\end{aligned}\tag{5.49}$$

$T = A^{1/2}$ yields $\check{N}\check{A} = A^{1/2}NA^{1/2} > 0$, while $T = N^{-1/2}$ produces $\check{N}\check{A} = N^{1/2}AN^{1/2} > 0$. Indirectly the iterates \check{x}^m define the sequence $\{x^m\}$ with

$$x^m = T^{-1}\check{x}^m.$$

In contrast to the previous transformation, the iteration (5.49) does not produce the solution x of $Ax = b$ but—in the case of convergence—the solution \check{x} of $\check{A}\check{x} = b$.

As emphasised above, the iteration (5.49) is used for theoretical purpose for an intermediate formulation (as, e.g., in §5.6.2.1). If one really wants to perform (5.49) with a transformation as $T = W^{1/2}$ in order to obtain a positive definite matrix $\check{N}\check{A} = W^{1/2}NAW^{-1/2} = W^{-1/2}AW^{-1/2}$, one should generalise the definition of a square root.

Exercise 5.63. Any factorisation of $W > 0$ into $W = V^H V$ induces the transformation by $T = V$. Prove that $VNAV^{-1}$ is positive definite. A particular example of $W = V^H V$ is Cholesky decomposition (cf. Quarteroni–Sacco–Saleri [314, §3.4.2]).

Chapter 6

Analysis of Positive Definite Iterations

Abstract This chapter gathers convergence statements about iterations satisfying suitable requirements connected with positive definiteness. Section 6.1 enumerates six cases which are analysed in Section 6.2. In several cases, convergence holds for a suitably damped version of the iteration. Of particular interest are symmetric and positive definite iterations constructed in the previous chapter. In Section 6.3 we analyse traditional symmetric iterative methods: the symmetric Gauss–Seidel iteration and the symmetric SOR method, abbreviated by SSOR. The convergence properties of SSOR are investigated in §§6.3.1–6.3.2, while modifications are described in §§6.3.3–6.3.4. Finally, in §6.3.5, numerical examples illustrate the convergence behaviour.

6.1 Different Cases of Positivity

We distinguish six cases of positivity. Consider any $\Phi \in \mathcal{L}$ and denote the corresponding matrices by

$$M = M[A], \quad N = N[A], \quad W = W[A].$$

- **Case 1:** positive spectrum of NA .

The weakest condition considered in this chapter is a positive spectrum of NA :

$$\sigma(NA) \subset (0, \infty). \tag{6.1a}$$

- **Case 2:** directly positive definite iterations $\Phi \in \mathcal{L}_{>0}$.

Positive definiteness appears in two versions. For directly positive definite iterations (cf. Definition 5.14) we have

$$NA > 0. \tag{6.1b}$$

Note that in this case no conditions on A are required except for regularity.

- **Case 3:** positive definite iterations $\Phi \in \mathcal{L}_{\text{pos}}$.

The standard situation is the case of an positive definite iteration (cf. Definition 5.8). Application to a positive definite matrix A yields

$$A > 0, \quad N > 0, \quad W > 0. \quad (6.1c)$$

Simple conclusions concerning the iteration matrix M are gathered in the next remark.

Remark 6.1. (a) Each condition in (6.1a–c) implies that $\sigma(M) \subset (-\infty, 1)$.

(b) In the case of (6.1b), M is Hermitian and satisfies $M < I$.

The positive definite matrices in (6.1b,c) induce the corresponding vector and matrix norms $\|\cdot\|_X$ for $X \in \{NA, A, N, W\}$ as defined in (C.5a,d).

Remark 6.2. Assume that convergence holds. (a) Then each of the conditions (6.1a–c) implies $\sigma(M) \subset (-1, 1)$.

(b) In the case of (6.1b), the convergence is monotone with respect to the Euclidean norm $\|\cdot\|$ and the norms $\|\cdot\|_{NA}$ and $\|\cdot\|_{(NA)^{-1}}$. The identities $\sigma(M) = \|M\| = \|M\|_{NA} = \|M\|_{(NA)^{-1}}$ hold.

(c) In the case of (6.1c), the convergence is monotone with respect to the norms $\|\cdot\|_A$ and $\|\cdot\|_W$, and $\sigma(M) = \|M\|_A = \|M\|_W$ holds.

Exercise 6.3. Assume (6.1c). The energy scalar product $\langle \cdot, \cdot \rangle_A$ is defined in (C.5b). Prove that M is symmetric with respect to $\langle \cdot, \cdot \rangle_A$, i.e., $\langle Mx, y \rangle_A = \langle x, My \rangle_A$, and that this statement is equivalent to $M = A^{-1}M^H A$.

Let $A > 0$. The symmetry with respect to $\langle \cdot, \cdot \rangle_A$ can be transferred to the usual symmetry by the following similarity transformation: $\hat{M} = \hat{M}^H$ holds for

$$\hat{M} := A^{1/2} M A^{-1/2} = I - A^{1/2} N A^{1/2} = I - A^{1/2} W^{-1} A^{1/2}. \quad (6.2a)$$

Similarly, $W > 0$ induces the similarity transformation

$$\check{M} := W^{1/2} M W^{-1/2} = I - W^{-1/2} A W^{-1/2} = I - N^{1/2} A N^{1/2}. \quad (6.2b)$$

The statement of Remark 6.2c can be expressed by

$$\rho(M) = \rho(\hat{M}) = \|\hat{M}\|_2 = \|M\|_A, \quad (6.2c)$$

$$\rho(M) = \rho(\check{M}) = \|\check{M}\|_2 = \|M\|_W. \quad (6.2d)$$

The proof follows from (A.6c) and (B.21b).

- **Case 4:** positive definite $W + W^H$.

The positive definiteness of W can be generalised to

$$W + W^H > A > 0.$$

The weaker condition

$$W + W^H > 0$$

will also be discussed, i.e., the Hermitian part of W is positive definite. An equivalent condition is

$$N + N^H > 0.$$

- **Case 5:** symmetrised iteration $\Phi^{\text{sym}} \in \mathcal{L}_{\text{sym}}$.

We recall the construction of a symmetric iteration $\Phi^{\text{sym}} = \Phi^* \circ \Phi$ described in §5.4.2. Theorem 5.29 states that $A = A^H$ leads to the matrices

$$\begin{aligned} M^{\text{sym}} &= (I - N^H A)(I - N A) = I - N^{\text{sym}} A, \\ N^{\text{sym}} &= N + N^H - N^H A N, \\ W^{\text{sym}} &= W(W + W^H - A)^{-1} W^H \end{aligned} \tag{6.3}$$

with N and W belonging to Φ , while the Hermitian matrices M^{sym} , N^{sym} , and W^{sym} are associated with Φ^{sym} .

- **Case 6:** perturbed positive definite A .

A non-Hermitian matrix A may be split into $A = A_0 + iA_1$ with positive definite $A_0 := \frac{1}{2}(A + A^H)$ (cf. (3.27)). If A_1 is small in a suitable sense, A can be regarded as a perturbation of the positive definite matrix A_0 .

6.2 Convergence Analysis

6.2.1 Case 1: Positive Spectrum

We assume (6.1a): $\sigma(NA) \subset (0, \infty)$. Sufficient conditions for (6.1a) are given in Lemma 5.18.

In §3.5.1, convergence of the Richardson iteration is investigated under the condition $\sigma(A) \subset (0, \infty)$. Using Proposition 5.44, we can transfer the results in §3.5.1 to NA . The quantities \varTheta and A in Lemma 3.21 and in Theorems 3.22, 3.23 have to be replaced with ϑ and NA . The matrices corresponding to the damped iteration Φ_ϑ are denoted by $M_\vartheta = I - \vartheta NA$, $N_\vartheta = \vartheta N$, and $W_\vartheta = N_\vartheta^{-1}$.

Lemma 6.4. *Assume that $\sigma(NA) \subset \mathbb{R}$ and denote the extreme eigenvalues of NA by λ_{\min} and λ_{\max} . Then the spectrum of the iteration matrix M_ϑ is real for any $\vartheta \in \mathbb{R}$, i.e., $\sigma(M_\vartheta) \subset \mathbb{R}$. The spectral radius is characterised by*

$$\rho(M_\vartheta) = \max\{|1 - \vartheta\lambda_{\min}|, |1 - \vartheta\lambda_{\max}|\} \quad \text{for all } \vartheta \in \mathbb{R}. \tag{6.4}$$

Exercise 6.5. Characterise $\rho(M_\vartheta)$ under the above assumptions for complex ϑ .

Theorem 6.6. Assume that condition (6.1a) holds and let $\lambda_{\max}(NA)$ be the maximal eigenvalue of NA . Then, for real ϑ , the damped iteration Φ_ϑ converges if and only if

$$0 < \vartheta < 2/\lambda_{\max}(NA). \quad (6.5)$$

The convergence rate is described by (6.4).

Theorem 6.7 (optimal ϑ). Under the assumptions of Theorem 6.6, the optimal convergence rate of Φ_ϑ is attained for

$$\vartheta_{\text{opt}} = \frac{2}{\lambda_{\max} + \lambda_{\min}} \quad \text{with} \quad \rho(M_{\vartheta_{\text{opt}}}) = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} = \frac{\kappa(NA) - 1}{\kappa(NA) + 1}. \quad (6.6a)$$

$\kappa(NA) = \lambda_{\max}/\lambda_{\min}$ is the spectral condition number of NA (cf. (B.13)). For large $\kappa(NA) \gg 1$, the asymptotic behaviour is

$$\frac{\kappa(NA) - 1}{\kappa(NA) + 1} = 1 - \frac{2}{\kappa(NA)} + \mathcal{O}(\kappa(NA)^{-2}). \quad (6.6b)$$

The expression $1 - 2/\kappa$ has to be compared with the rate $1 - 1/\kappa$ for iterations with $\sigma(M) \subset [0, 1)$.

Remark 6.8. Assume (6.1a) and $\sigma(M) \subset [0, \infty)$. The optimal scaling factor ϑ satisfying $\sigma(M_\vartheta) \subset [0, 1)$ is $\vartheta_+ = 1/\lambda_{\max}$. The corresponding rate is $\rho(M_{\vartheta_+}) = 1 - \frac{1}{\kappa(NA)}$.

For a complex spectrum of NA , compare with Exercise 3.26 and Theorem 3.27.

6.2.2 Case 2: Positive Definite NA

Theorem 6.9. Assume $NA > 0$ and $\vartheta \in \mathbb{R}$. Then iteration (5.8) converges if and only if

$$0 < \vartheta < 2/\|NA\|_2.$$

The convergence is monotone with respect to the Euclidean norm $\|\cdot\|_2$ and the energy norm $\|\cdot\|_{NA}$. Furthermore, the convergence rate and the contraction number coincide:

$$\rho(M_\vartheta) = \|M_\vartheta\|_2 = \|M_\vartheta\|_{NA}.$$

The optimal convergence rate (6.6a) can be expressed as a function of the condition number $\kappa(NA) = \text{cond}_2(NA)$:

$$\rho(M_{\vartheta_{\text{opt}}}) = \frac{\kappa(NA) - 1}{\kappa(NA) + 1} \quad \text{for } \vartheta_{\text{opt}} = \frac{2}{\lambda_{\max}(NA) + \lambda_{\min}(NA)}. \quad (6.7)$$

Proof. Use the results in §6.2.1, $\lambda_{\max}(NA) = \|NA\|_2$, and Remark 6.2b. \square

6.2.3 Case 3: Positive Definite Iteration

Now we assume (6.1c). This case is already treated by Theorem 3.34. Since the proof is still missing, we repeat the statements in short. Note that (6.8a) describes a sufficient and necessary condition for convergence.

Theorem 6.10. *Let (6.1c) be valid. Then, for $0 \leq \lambda \leq \Lambda$, the following equivalence relations hold:*

$$2W > A > 0 \iff \rho(M) < 1, \quad (6.8a)$$

$$0 < \lambda W \leq A \leq \Lambda W \iff \sigma(M) \subset [1 - \Lambda, 1 - \lambda], \quad (6.8b)$$

$$0 \leq \lambda W < A < \Lambda W \iff \sigma(M) \subset (1 - \Lambda, 1 - \lambda), \quad (6.8c)$$

$$W \geq A > 0 \iff \sigma(M) \subset [0, 1]. \quad (6.8d)$$

Proof. Using the matrix \hat{M} in (6.2a), $\sigma(M) = \sigma(\hat{M})$ allows us to reformulate $\sigma(\hat{M}) \subset [1 - \Lambda, 1 - \lambda]$ as

$$(1 - \Lambda) I \leq \hat{M} = I - A^{1/2} N A^{1/2} \leq (1 - \lambda) I$$

(cf. (C.3e)). Applying (C.3b') with $C := A^{-1/2}$, we get the equivalent inequalities

$$(1 - \Lambda) A^{-1} \leq A^{-1} - N \leq (1 - \lambda) A^{-1}.$$

The left inequality yields $-\Lambda A^{-1} \leq -N \Leftrightarrow \Lambda A^{-1} \geq N$. Applying (C.3g), we arrive at $\frac{1}{\Lambda} A \leq N^{-1} = W$, i.e., $A \leq \Lambda W$. The proof of $\lambda W \leq A$ is analogous. This proves (6.8b). Replacing ' \leq ' with ' $<$ ', we obtain (6.8c). The implications (6.8a,d) follow for special values of λ and Λ . \square

Denote the iteration defined by the matrices (6.1c) by Φ . Below we discuss the damped iteration Φ_ϑ . Theorems 6.6 and 6.7 and (6.2c,d) yield the following result.

Theorem 6.11. *Assume (6.1c). The damped iteration Φ_ϑ defined by (5.8) converges if and only if ϑ satisfies*

$$\begin{aligned} 0 < \vartheta < 2/\lambda_{\max} \quad &\text{with} \\ \lambda_{\max} := \|N^{1/2} A N^{1/2}\|_2 &= \|A^{1/2} N A^{1/2}\|_2 = \rho(NA). \end{aligned} \quad (6.9)$$

An equivalent formulation of condition (6.9) using $W = N^{-1}$ is

$$0 < \vartheta A < 2W.$$

The convergence rate (even for general $\vartheta \in \mathbb{C}$) is

$$\rho(M_\vartheta) = \|M_\vartheta\|_A = \|M_\vartheta\|_W = \max\{|1 - \vartheta\lambda_{\min}|, |1 - \vartheta\lambda_{\max}|\},$$

where λ_{\min} is the minimal eigenvalue of NA . The optimal value of ϑ minimising $\rho(M_\vartheta)$ is ϑ_{opt} in (6.7).

Corollary 6.12. (a) If $A > 0$ and $N < 0$ ($\Leftrightarrow W < 0$), then Φ_ϑ converges if and only if $0 > \vartheta > 2/\rho(NA)$.

(b) If $A > 0$, $N = N^H$, and neither $N > 0$ nor $N < 0$, Φ_ϑ diverges for all $\vartheta \in \mathbb{C}$.

6.2.4 Case 4: Positive Definite $W + W^H$ or $N + N^H$

First, we assume

$$W + W^H > A > 0. \quad (6.10)$$

The first part of the next theorem coincides with Theorem 3.35.

Theorem 6.13. *Under condition (6.10), the iteration converges monotonically with respect to the energy norm:*

$$\rho(M) \leq \|M\|_A < 1 \quad \text{for } M = I - W^{-1}A.$$

$W + W^H > A$ is also necessary for $\|M\|_A < 1$ (but even without condition (6.10), $\rho(M) < 1$ is possible).

Proof. Assume that $W + W^H - A$ has a nonpositive eigenvalue. Then, by (3.37), $\hat{M}^H \hat{M} = I - A^{1/2} W^H (W + W^H - A) W A^{1/2}$ has an eigenvalue ≥ 1 implying $\|M\|_A \geq 1$. \square

Next, we assume

$$W + W^H > 0 \quad \text{and} \quad A > 0.$$

To regain inequality (6.10), we have to apply a suitable damping, since Φ_ϑ is associated with $W_\vartheta = \frac{1}{\vartheta} W$. For instance, the choice

$$\vartheta < \frac{\lambda_{\min}(W + W^H)}{\lambda_{\max}(A)} \quad (6.11)$$

ensures that $W_\vartheta + W_\vartheta^H > A > 0$.

Exercise 6.14. The sharper estimate $\vartheta < \lambda_{\min}(A^{-1/2} (W + W^H) A^{-1/2})$ also implies $W_\vartheta + W_\vartheta^H > A > 0$.

Remark 6.15. Theorem 6.13 proves that Φ_ϑ with ϑ in (6.11) is convergent. The convergence is monotone with respect to the energy norm: $\|M_\vartheta\|_A = \|\hat{M}_\vartheta\|_2 = \rho(\hat{M}_\vartheta^H \hat{M}_\vartheta)^{1/2}$ (\hat{M}_ϑ as in (6.2a)).

Optimising the damping factor ϑ leads us to the quadratic inequality

$$\vartheta (W + W^H) \geq \vartheta^2 A + \alpha W^H A W, \quad \alpha = \alpha(\vartheta) > 0. \quad (6.12)$$

For each sufficiently small $\vartheta > 0$, there is a maximal $\alpha(\vartheta)$ satisfying (6.12). ϑ_{opt} is the maximiser of $\alpha(\vartheta)$.

Theorem 6.16. *Let Φ_ϑ satisfy (6.12). Then Φ_ϑ converges with the contraction number*

$$\|M_\vartheta\|_A = \sqrt{1 - \alpha}.$$

Proof. Repeat the estimate of $\hat{M}_\vartheta^H \hat{M}_\vartheta$ in (3.37) and use (6.12). \square

The assumption $N + N^H > 0$ does not yield new results because of the next lemma, but in concrete cases the matrix $N + N^H$ may be easier to analyse than $W + W^H$.

Lemma 6.17. $N + N^H > 0$ and $W + W^H > 0$ are equivalent.

Proof. $N + N^H > 0 \Leftrightarrow W^H(N + N^H)W = W + W^H > 0$ by (C.3a). \square

Remark 6.18. Assume $N + N^H > 0$. With a suitable scaling, N_ϑ satisfies

$$N_\vartheta + N_\vartheta^H > N_\vartheta^H A N_\vartheta$$

which is equivalent to $W_\vartheta + W_\vartheta^H > A > 0$ and allows applying Theorem 6.13. The estimate

$$N_\vartheta + N_\vartheta^H - N_\vartheta^H A N_\vartheta \geq \alpha A$$

is equivalent to (6.12).

6.2.5 Case 5: Symmetrised Iteration Φ^{sym}

Below we use the notation defined in (6.3). In particular, M^{sym} and M are the respective iteration matrices of $\Phi^{\text{sym}} = \Phi^* \circ \Phi$ and Φ .

Remark 6.19. Assume $A > 0$. Then

$$\sigma(M^{\text{sym}}) = \|\hat{M}^{\text{sym}}\|_2 = \|M^{\text{sym}}\|_A \subset [0, \infty)$$

holds, where $\hat{M}^{\text{sym}} = A^{1/2} M^{\text{sym}} A^{-1/2} > 0$. The connection to the iteration Φ is given by

$$\sigma(M^{\text{sym}}) = \|M\|_A^2 = \|\hat{M}\|_2^2 \quad (\hat{M} := A^{1/2} M A^{-1/2}). \quad (6.13)$$

If Φ^{sym} converges, the convergence is monotone with respect to the energy norm $\|\cdot\|_A$, and $\sigma(M^{\text{sym}}) \subset [0, 1)$ holds.

Proof. Use $\hat{M}^{\text{sym}} = \hat{M}^H \hat{M} \geq 0$ with $\hat{M} = A^{1/2} M A^{-1/2}$ and the similarity of M^{sym} and \hat{M}^{sym} . \square

Equation (6.13) yields the following important conclusion. In general, the condition $\|M\|_A < 1$ (monotone convergence with respect to the energy norm) is only sufficient for convergence. Because of the next statement this is even a necessary condition for Φ^{sym} . Therefore estimates of $\|M\|_A$ become important.

Conclusion 6.20. $\Phi^{\text{sym}} = \Phi^* \circ \Phi$ converges if and only if Φ is monotonically converging with respect to the energy norm, i.e., $\|M_\Phi\|_A < 1$.

The construction of $\Phi^{\text{sym}} = \Phi^* \circ \Phi$ in §5.4.2 ensures that $A > 0$ implies $N^{\text{sym}} = (N^{\text{sym}})^H$. N^{sym} is Hermitian, but not necessarily positive definite. By Corollary 6.12b, convergence of the damped version of Φ^{sym} requires either $N^{\text{sym}} > 0$ or $N^{\text{sym}} < 0$. The second case is completely nonstandard. Since $N^{\text{sym}} = N + N^H - N^H A N$ (cf. (6.3)), the condition $N^{\text{sym}} > 0$ is equivalent to the identical conditions $N + N^H > N^H A N$ and $W + W^H > A > 0$ in Remark 6.18. As stated in Remark 6.18, these inequalities can be guaranteed by a suitable scaling if $N + N^H > 0$ or equivalently $W + W^H > 0$.

Next, we investigate the properties of $(\Phi_\vartheta)^{\text{sym}} = \Phi_\vartheta^* \circ \Phi_\vartheta$. For a proof, use Remark 6.19.

Proposition 6.21. *Assume $A > 0$. Let M, N, W and $M_\vartheta, N_\vartheta, W_\vartheta$ be the matrices associated with Φ and the damped iteration Φ_ϑ , while $M^{\vartheta, \text{sym}}, N^{\vartheta, \text{sym}}, W^{\vartheta, \text{sym}}$ are those of $(\Phi_\vartheta)^{\text{sym}} = \Phi_\vartheta^* \circ \Phi_\vartheta$.*

(a) *Positive definite case $N + N^H > 0$: For a suitable scaling factor $\vartheta > 0$, $W_\vartheta + W_\vartheta^H > A$ holds and $(\Phi_\vartheta)^{\text{sym}}$ converges. Since $N^{\vartheta, \text{sym}} = N_\vartheta + N_\vartheta^H - N_\vartheta^H A N_\vartheta$, the statements of Remark 6.18 apply. In the convergent case, the transformed iteration matrix $\hat{M}^{\vartheta, \text{sym}} := A^{1/2} M^{\vartheta, \text{sym}} A^{-1/2}$ satisfies*

$$0 \leq \hat{M}^{\vartheta, \text{sym}} < I,$$

and $(\Phi_\vartheta)^{\text{sym}}$ is a positive definite iteration.

- (b) *Negative definite case $N + N^H < 0$: A negative ϑ leads us back to case (a).*
- (c) *Otherwise, $(\Phi_\vartheta)^{\text{sym}}$ diverges for any choice of ϑ .*

Let ϑ be a suitable scaling of Φ so that $W_\vartheta + W_\vartheta^H > A$ holds. Rename $\Phi_\vartheta, M_\vartheta, N_\vartheta, W_\vartheta, (\Phi_\vartheta)^{\text{sym}}$ by $\Phi, M_\Phi, N_\Phi, W_\Phi, \Phi^{\text{sym}}$. The statements of Remark 6.19, together with the convergence criterion $W_\Phi + W_\Phi^H > A > 0$, yield the next result.

Theorem 6.22. *Assume that*

$$W_\Phi + W_\Phi^H > A > 0.$$

Then the symmetrised iteration $\Phi^{\text{sym}} := \Phi^ \circ \Phi$ converges monotonically:*

$$\rho(M_{\Phi^{\text{sym}}}) = \|M_\Phi\|_A^2 < 1. \quad (6.14)$$

Moreover the spectrum is nonnegative:

$$\sigma(M_{\Phi^{\text{sym}}}) \subset [0, \rho(M_{\Phi^{\text{sym}}})] \subset [0, 1].$$

Proof. For the last equality combine (6.13) in the form $\sigma(M_{\Phi^{\text{sym}}}) \subset [0, \rho(M_{\Phi^{\text{sym}}})]$ with (6.14). \square

Since $\sigma(M_{\Phi^{\text{sym}}}) \subset [0, 1)$ holds, Remark 6.8 shows that the convergence rate can be improved by damping (extrapolation).

Concerning the contraction number with respect to the energy norm, Φ and Φ^{sym} behave the same: Φ^{sym} consists of two iteration steps and yields the same bound $\|M_\Phi\|_A^2$ as two steps of Φ . However, concerning the convergence rate, the symmetric iteration performs worse. While $\rho(M_\Phi)$ may be strictly smaller than $\|M_\Phi\|_A$ (cf. Remark 6.15), $\rho(M_{\Phi^{\text{sym}}})$ is equal to $\|M_\Phi\|_A^2$; i.e., the inequality

$$\rho(M_{\Phi^{\text{sym}}}) \geq \rho(M_\Phi)^2$$

holds and may possibly be a strict inequality.

When assessing Φ^{sym} and Φ only with regard to convergence speed, Φ should be preferred. The advantage of Φ^{sym} will be seen in connection with Krylov methods. Another advantage is the possibility to perform Φ^{sym} with less cost than two steps of Φ (cf. Remark 6.27).

6.2.6 Case 6: Perturbed Positive Definite Case

The next generalisation splits A into $A_0 + iA_1$ according to (3.27). The condition $A > 0$ is weakened by $A_0 > 0$.

Theorem 6.23. Assume that $A = A_0 + iA_1$ according to (3.27) satisfies $A_0 > 0$. Let $W = N[A]^{-1} > 0$ hold for the matrix of the third normal form of $\Phi(\cdot, \cdot, A)$. The optimal constants $0 < \lambda \leq \Lambda$ and $\tau \geq 0$ in

$$\lambda W \leq A_0 \leq \Lambda W, \quad -\tau W \leq A_1 \leq \tau W \quad (6.15)$$

are $\lambda = \lambda_{\min}(NA_0)$, $\Lambda = \lambda_{\max}(NA_0)$, and $\tau := \rho(NA_1)$. Then the damped iteration (5.8) converges for

$$0 < \vartheta < \frac{2\lambda}{\lambda\Lambda + \tau^2}$$

monotonically with respect to the norm $\|\cdot\|_W$:

$$\rho(M_\vartheta) \leq \|M_\vartheta\|_W \leq \frac{1}{2}\vartheta(\Lambda - \lambda) + \sqrt{\left[1 - \frac{1}{2}\Theta(\Lambda + \lambda)\right]^2 + \Theta^2\tau^2} < 1.$$

The optimal ϑ can be determined as in (3.31c).

Proof. M_ϑ is similar to $M := N^{-1/2}M_\vartheta N^{1/2} = I - \vartheta N^{1/2}AN^{1/2}$. M can be regarded as the iteration matrix of the Richardson method for $\Theta := \vartheta$ and $A' := N^{1/2}AN^{1/2}$ instead of A . The splitting $A = A_0 + iA_1$ induces the splitting $A' = A'_0 + iA'_1$ with the Hermitian matrices

$$A'_0 = N^{1/2}A_0 N^{1/2}, \quad A'_1 = N^{1/2}A_1 N^{1/2}.$$

The inequalities (3.30a,b) applied to A' are equivalent to (6.15). The estimate (3.31b) following from Theorem 3.30 refers to the iteration matrix M and reads as $\|M\|_2 = \|W^{1/2}M_\vartheta W^{-1/2}\|_2 = \|M_\vartheta\|_W$. \square

The counterpart of Theorem 3.31 reads as follows.

Theorem 6.24. *Under the assumption (3.30a,b), the estimate*

$$\rho(M_\vartheta) \leq r_\vartheta := \sqrt{\vartheta^2 \tau^2 + \max \{ |1 - \vartheta \lambda|, |1 - \vartheta \Lambda| \}}$$

holds for the damped iteration (5.8) with λ and Λ as in Theorem 6.23. The convergence is ensured in the form $r_\vartheta < 1$ if

$$0 < \vartheta < \bar{\vartheta} \quad \text{with } \bar{\vartheta} := \begin{cases} 2\Lambda / (\Lambda^2 + \tau^2) & \text{if } \tau^2 < \lambda\Lambda, \\ 2\lambda / (\Lambda^2 + \tau^2) & \text{if } \tau^2 \geq \lambda\Lambda. \end{cases}$$

r_ϑ is minimal for $\vartheta' := \min\{\frac{\lambda}{\lambda^2 + \tau^2}, \frac{2}{\lambda + \Lambda}\}$. Moreover, the norm estimate (6.16) holds:

$$\|(M_\vartheta)^m\|_W \leq 2r_\vartheta^m \quad (m \geq 0). \quad (6.16)$$

Exercise 6.25. Reformulate Corollary 3.32 for the damped iteration (5.8).

In the case of a matrix $NA = C_0 + iC_1$ decomposed into a Hermitian part $C_0 := (NA + A^H N^H)/2$ and a skew-Hermitian part $C_1 := (NA - A^H N^H)/(2i)$, we can apply the counterparts of Theorems 3.28, 3.30, 3.31 and Corollaries 3.32, 3.33 to get similar results as above.

6.3 Symmetric Gauss–Seidel Iteration and SSOR

The symmetric Gauss–Seidel method $\Phi^{\text{symGS}} = \Phi_{\text{backw}}^{\text{GS}} \circ \Phi^{\text{GS}} \in \mathcal{L}_{\text{sym}}$ and the symmetric SOR method (SSOR) $\Phi_\omega^{\text{SSOR}} = \Phi_\omega^{\text{backwSOR}} \circ \Phi_\omega^{\text{SOR}} \in \mathcal{L}_{\text{sym}}$ are defined in §5.4.3. In 1955, the SSOR method is first described by Sheldon [339].

Since $\Phi_1^{\text{SOR}} = \Phi^{\text{GS}}$ (cf. Proposition 3.13c), the symmetric Gauss–Seidel iteration also satisfies $\Phi^{\text{symGS}} = \Phi_1^{\text{SSOR}}$. Therefore the symmetric Gauss–Seidel method does not require a separate analysis.

6.3.1 The Case $A > 0$

Theorem 6.26. *Let A be positive definite. The symmetric SOR method $\Phi_\omega^{\text{SSOR}}$ converges for $0 < \omega < 2$ with*

$$\rho(M_\omega^{\text{SSOR}}) = \|M_\omega^{\text{SOR}}\|_A^2 < 1, \quad \text{where} \quad M_\omega^{\text{SSOR}} = M_\omega^{\text{backwSOR}} M_\omega^{\text{SOR}}$$

(cf. Remark 5.2 and (3.15b)). The spectrum $\sigma(M_\omega^{\text{SSOR}})$ is contained in $[0, 1]$. $\Phi_\omega^{\text{SSOR}}$ diverges for all real $\omega \notin (0, 2)$. The same statements hold for the block-SSOR version.

Proof. Combine the result of Theorem 3.41 (Ostrowski) with Theorem 6.22. Concerning $\omega \notin (0, 2)$ use $\rho(M_\omega^{\text{SSOR}}) = \|M_\omega^{\text{SOR}}\|_A^2 \geq \rho(M_\omega^{\text{SOR}})^2$ and (3.41). \square

The amount of work required by the symmetric SOR iteration seems to be twice as large as that for the original SOR method, since one SSOR step consists of two SOR steps (cf. (5.14)). However, this disadvantage can be overcome.

Remark 6.27 (Niethammer [292, 293]). The SSOR iteration requires essentially the same amount of work as the SOR method if one tolerates additional storage needed for an auxiliary vector. The cost factor (cf. §2.3) amounts to

$$C_{\Phi}^{\text{SSOR}} = C_{\Phi}^{\text{SOR}} + 5/C_A = 2 + 6/C_A$$

for an optimal implementation instead of $2C_{\Phi}^{\text{SOR}} = 4 + 2/C_A$ for the naive implementation (5.14).

Proof. The first SSOR half-step $x^m \mapsto x^{m+1/2}$ can be rewritten as

$$x^{m+1/2} = x^m + \omega (Lx^{m+1/2} - x^m + Ux^m + D^{-1}b) \quad (6.17a)$$

(cf. (3.15f)). The second backward SOR step

$$x^{m+1} = x^{m+1/2} + \omega (Ux^{m+1} - x^{m+1/2} + Lx^{m+1/2} + D^{-1}b) \quad (6.17b)$$

contains the term $Lx^{m+1/2}$ which is already evaluated in (6.17a). Analogously, the term Ux^{m+1} computed in (6.17b) can be used in the following half-step:

$$x^{m+3/2} = x^{m+1} + \omega (Lx^{m+3/2} - x^{m+1} + Ux^{m+1} + D^{-1}b).$$

On the average, one SSOR step requires one evaluation of Lx and Ux . □

The statements of Theorem 3.44 can be translated into the following statement about the SSOR method.

Theorem 6.28. Let $A = D - E - E^H > 0$ and $0 < \omega < 2$. Furthermore, assume that there are constants $\gamma > 0$ and Γ with (6.18a,b) (cf. (3.46a,b)):

$$0 < \gamma D \leq A, \quad (6.18a)$$

$$\left(\frac{1}{2}D - E \right) D^{-1} \left(\frac{1}{2}D - E^H \right) \leq \frac{1}{4}\Gamma A. \quad (6.18b)$$

Then the following estimate holds:

$$\rho(M_{\omega}^{\text{SSOR}}) = \|M_{\omega}^{\text{SSOR}}\|_A \leq 1 - \frac{2\Omega}{\frac{\Omega^2}{\gamma} + \Omega + \frac{\Gamma}{4}} \quad \text{with } \Omega := \frac{2 - \omega}{2\omega}. \quad (6.18c)$$

For $\omega' = 2/(1 + \sqrt{\gamma\Gamma})$, the bound in (6.18c) becomes a minimum:

$$\rho(M_{\omega}^{\text{SSOR}}) \leq \frac{\sqrt{\Gamma} - \sqrt{\gamma}}{\sqrt{\Gamma} + \sqrt{\gamma}} = \frac{1 - \sqrt{\gamma/\Gamma}}{1 + \sqrt{\gamma/\Gamma}}.$$

Proof. Combine (6.14) with Theorem 3.44. □

The following statement is analogous to Conclusion 3.46.

Conclusion 6.29 (order improvement). *In the case of $\rho(D^{-1}ED^{-1}E^H) \leq 1/4$ (or $\leq 1/4 + \mathcal{O}(1 - \rho(M^{Jac}))$), the choice $\omega = \omega'$ enables an order improvement. If τ is the order of the Jacobi (and of the symmetric Gauss–Seidel) method, then $\tau/2$ is the order of the SSOR method with $\omega = \omega'$.*

The condition $\rho(D^{-1}ED^{-1}E^H) \leq 1/4$ is essential. This inequality does not hold for the model problem with chequer-board ordering. Then, as we shall see in §6.3.4, no order improvement is possible.

For completeness, we repeat the properties of the symmetric Gauss–Seidel iteration.

Proposition 6.30. (a) *The iteration matrix of the symmetric Gauss–Seidel iteration and the matrices of the second and third normal forms are*

$$\begin{aligned} M^{\text{symGS}} &= (D - F)^{-1}E(D - E)^{-1}F, \\ N^{\text{symGS}} &= (D - F)^{-1}D(D - E)^{-1}, \\ W^{\text{symGS}} &= (D - E)D^{-1}(D - F) = A + ED^{-1}F. \end{aligned}$$

(b) *The symmetric Gauss–Seidel iteration is a symmetric iteration in the sense of Definition 5.3, provided that $D \in \mathbb{R}^{I \times I}$.*

(c) *If $A > 0$, the matrix W^{symGS} of the third normal form is also positive definite, so that the symmetric Gauss–Seidel iteration is a positive definite iteration.*

(d) *The symmetric Gauss–Seidel iteration converges and the spectrum of the iteration matrix is nonnegative:*

$$\sigma(M^{\text{symGS}}) \subset [0, 1].$$

6.3.2 SSOR in the 2-Cyclic Case

In the 2-cyclic case, we can rewrite the backward SOR iteration as $\Phi_\omega^{\text{backwSOR}} = \Phi_\omega^{(1)} \circ \Phi_\omega^{(2)}$ with the partial steps defined in (6.19a,b). Therefore, the symmetric SOR iteration takes the form

$$\Phi_\omega^{\text{SSOR}} = \Phi_\omega^{(1)} \circ \Phi_\omega^{(2)} \circ \Phi_\omega^{(2)} \circ \Phi_\omega^{(1)} \in \mathcal{L}_{\text{sym}}.$$

Exercise 6.31. Prove: (a) The SSOR iteration matrix $M_\omega^{\text{SSOR}} = M_\omega^{(1)}M_\omega^{(2)}M_\omega^{(2)}M_\omega^{(1)}$ leads to the rate

$$\rho(M_\omega^{(1)}M_\omega^{(2)}M_\omega^{(2)}M_\omega^{(1)}) = \rho(M_\omega^{(2)}M_\omega^{(2)}M_\omega^{(1)}M_\omega^{(1)}).$$

- (b) $M_\omega^{(1)}M_\omega^{(1)} = M_{\omega'}^{(1)}$ and $M_\omega^{(2)}M_\omega^{(2)} = M_{\omega'}^{(2)}$ hold with $\omega' := \omega(2 - \omega)$.
(c) $0 < \omega < 2$ implies $0 < \omega' \leq 1$. $\omega' = 1$ is only achieved for $\omega = 1$.

Exercise 6.31 entails the following negative conclusion.

Conclusion 6.32. *In the 2-cyclic case, $\rho(M_\omega^{\text{SSOR}}) = \rho(M_{\omega'}^{\text{SOR}})$ holds with $\omega' := \omega(2 - \omega) \leq 1$ for all $0 < \omega < 2$. According to Theorem 4.27, underrelaxation ($\omega' < 1$) is always slower than the Gauss–Seidel iteration ($\omega' = 1$). Hence, $\omega = 1$ is the optimal parameter and SSOR simplifies to the symmetric Gauss–Seidel iteration (cf. Alefeld [2]).*

The reason for the missing order improvement is that, differently from the situation discussed in Remark 6.29, the condition $\rho(D^{-1}ED^{-1}E^H) \leq \frac{1}{4}$ is not satisfied. In the 2-cyclic case, we have $\rho(D^{-1}ED^{-1}E^H) = \rho(D_1^{-1}A_1D_2^{-1}A_2) = \rho(M^{\text{GS}}) \approx 1$ (cf. Theorem 4.20).

Exercise 6.33. Let (A, D) be 2-cyclic. Prove that

$$M^{\text{symGS}} = \begin{bmatrix} 0 & -D_1^{-1}A_1D_2^{-1}A_2D_1^{-1}A_1 \\ 0 & D_2^{-1}A_2D_1^{-1}A_1 \end{bmatrix}$$

is the iteration matrix of the symmetric Gauss–Seidel method and that

$$\rho(M^{\text{symGS}}) = \rho(M^{\text{GS}}).$$

6.3.3 Modified SOR

In the 2-cyclic case, we can regard the SOR method as a product iteration $\Phi_\omega^{\text{SOR}} = \Phi_\omega^{(2)} \circ \Phi_\omega^{(1)}$ (cf. §5.4), where $\Phi_\omega^{(1)}$ involves only the first block of the vector and $\Phi_\omega^{(2)}$ only the second one:

$$\Phi_\omega^{(1)}(x, b) = \begin{pmatrix} x^1 - \omega [x^1 - D_1^{-1}(A_1x^2 - b^1)] \\ x^2 \end{pmatrix} \quad (6.19a)$$

$$\Phi_\omega^{(2)}(x, b) = \begin{pmatrix} x^1 \\ x^2 - \omega [x^2 - D_2^{-1}(A_2x^1 - b^2)] \end{pmatrix} \quad (6.19b)$$

where $x = \begin{bmatrix} x^1 \\ x^2 \end{bmatrix}$, $b = \begin{bmatrix} b^1 \\ b^2 \end{bmatrix}$, and A is split as in (4.3). The corresponding iteration matrices are

$$M_\omega^{(1)} = \begin{bmatrix} (1 - \omega)I & \omega D_1^{-1}A_1 \\ 0 & I \end{bmatrix}, \quad M_\omega^{(2)} = \begin{bmatrix} I & 0 \\ \omega D_2^{-1}A_2 & (1 - \omega)I \end{bmatrix}.$$

Thus we have $M_\omega^{\text{SOR}} = M_\omega^{(2)}M_\omega^{(1)}$. The modified SOR iteration (MSOR) makes use of different relaxation parameters ω and ω' in both of the half-steps:

$$\Phi_{\omega, \omega'}^{\text{modSOR}} = \Phi_{\omega'}^{(2)} \circ \Phi_\omega^{(1)}.$$

Again the comment in Remark 3.6 about multiple parameters applies. Concerning convergence analysis and optimal parameters, we refer to Young [412, §8] and Hadjidimos [211, §3].

6.3.4 Unsymmetric SOR Method

The only reason for mentioning the unsymmetric SOR method is that it is constructed in analogy to the modified Gauss–Seidel method in §6.3.3. The SSOR method $\Phi_{\omega}^{\text{SSOR}} = \Phi_{\omega}^{\text{backwSOR}} \circ \Phi_{\omega}^{\text{SOR}}$ (cf. (5.15)) can be modified by choosing different parameters ω, ω' in both factors. Accordingly, the unsymmetric SOR iteration reads

$$\Phi_{\omega, \omega'}^{\text{unsymSOR}} := \Phi_{\omega'}^{\text{backwSOR}} \circ \Phi_{\omega}^{\text{SOR}}.$$

Again, this method is not notably better than the SSOR method. For more details and further references, see Hadjidimos [211, §4.1].

6.3.5 Numerical Results for the SSOR Iteration

For methods with an iteration matrix satisfying $0 < A^{1/2}MA^{-1/2} < \rho(M)I$, Remark 2.22d is applicable: the quotients $\|e^m\|_A / \|e^{m-1}\|_A$ converge monotonically to $\rho(M)$. Since $M = M_{\omega}^{\text{SSOR}}$ satisfies this assumption, we observe this monotone behaviour for the SSOR iteration and the symmetric Gauss–Seidel method ($\omega = 1$). Table 6.1 (left) contains the results of the SSOR method with lexicographical ordering. For the Poisson model problem with step size $h = 1/32$, we obtain the convergence rate 0.98092. According to Table 3.2, $\omega = \omega' = 1.8213$ is the optimal value for the bound (3.55c), which becomes $\|M_{\omega}^{\text{SSOR}}\|_A \leq 0.9065$. Table 6.1 shows the convergence rates for different ω . Obviously, $\rho(M_{\omega}^{\text{SSOR}})$ attains its minimum not at $\omega = \omega'$ but for $\omega_{\text{opt}} \in [1.845, 1.846]$. The values of Table 6.1 demonstrate that, differently from the SOR method (cf. Fig. 4.1), the convergence rate has a flat minimum. Small errors in the choice of $\omega = \omega_{\text{opt}}$ deteriorate the convergence rate only insignificantly. In this respect, the choice $\omega = \omega'$ is sufficiently good.

symmetric Gauss–Seidel iteration				SSOR with $\omega = 1.8213$		ω	$\rho(M_{\omega}^{\text{SSOR}})$
m	$\ e^m\ _{\infty}$	$\ e^m\ _A$	$\frac{\ e^m\ _{\infty}}{\ e^{m-1}\ _{\infty}}$	$\ e^m\ _A$	$\frac{\ e^m\ _A}{\ e^{m-1}\ _A}$		
1	1.48	202	0.79011	0.579572	$2.3_{10}+02$	0.67588	1 0.98092
2	1.35	159	0.91627	0.790646	$1.6_{10}+02$	0.71534	1.8 0.88376
3	1.27	137	0.94025	0.858495	$1.2_{10}+02$	0.72622	1.81 0.88163
4	1.20	122	0.94528	0.891046	$9.0_{10}+01$	0.73679	1.8213 0.87962
5	1.14	111	0.94734	0.910237	$6.7_{10}+01$	0.74876	1.83 0.87845
94	0.158	11.2	0.98074	0.980884	$3.4_{10}-04$	0.87961	1.84 0.87765
95	0.155	11.0	0.98075	0.980891	$2.8_{10}-04$	0.87961	1.8450 0.877529
96	0.152	10.8	0.98075	0.980897	$2.5_{10}-04$	0.87961	1.8455 0.877528
97	0.149	10.6	0.98076	0.980903	$2.2_{10}-04$	0.87961	1.8460 0.877528
98	0.146	10.4	0.98076	0.980909	$1.9_{10}-04$	0.87961	1.847 0.877528
99	0.144	10.2	0.98077	0.980914	$1.7_{10}-04$	0.87961	1.85 0.87762
100	0.141	10.0	0.98077	0.980919	$1.5_{10}-04$	0.87961	1.86 0.87855

Table 6.1 Left: Symmetric Gauss–Seidel iteration and SSOR for $h = 1/32$. Right: Convergence rates of the SSOR method for $h = 1/32$ and different ω .

Chapter 7

Generation of Iterations

Abstract The algebraic operations described in Chapter 5 are tools for generating linear iterations. In this chapter we discuss how these tools can be used to build new iterative methods. The product of iterations is recalled in Section 7.1 and refers to later applications in Part III. Many traditional iterations are constructed by the *additive splitting* technique of Section 7.2. The *regular splitting* and *weakly regular splitting* defined in §7.2.2 yield sufficient convergence criteria. Another kind of splitting is the *P-regular splitting* defined in §7.2.4. A special kind of additive splitting is the *incomplete triangular decomposition (ILU)* discussed in Section 7.3. The transformations introduced in §5.6 will reappear in Section 7.4 under the name *preconditioning*.

7.1 Product Iterations

We recall that new iterations can be constructed by the product of simpler ones:

$$\Pi := \Phi \circ \Psi \quad \text{for } \Phi, \Psi \in \mathcal{L}.$$

Of particular interest are *symmetric iterations*. If Φ is not symmetric, it can be symmetrised: $\Phi^{\text{sym}} := \Phi^* \circ \Phi$ (also $\Phi \circ \Phi^*$ would be possible). The Krylov methods of Part II are best to combine with *positive definite iterations*, for which $A > 0$ implies $N[A] > 0$.

Symmetric products of three factors will also appear (see, e.g., Lemma 11.44). Corollary 5.30 states that $\Phi^* \circ \Psi \circ \Phi$ is symmetric if Ψ is so. The corresponding statement about positive definiteness follows. Note that Criterion 5.10 yields a criterion for $\Phi^* \circ \Phi$ to be positive definite.

Lemma 7.1. *If $\Phi^* \circ \Phi \in \mathcal{L}_{\text{pos}}$ and $\Psi \in \mathcal{L}_{\text{semi}}$, then the product satisfies*

$$\Phi^* \circ \Psi \circ \Phi \in \mathcal{L}_{\text{pos}}.$$

Proof. Let $A > 0$. One verifies that

$$N_{\Phi^* \circ \Psi \circ \Phi} = M_{\Phi^*} N_{\Psi} M_{\Phi^*}^H + N_{\Phi^* \circ \Phi}.$$

Positive semidefiniteness of Ψ yields $N_{\Psi} \geq 0$ and $M_{\Phi^*} N_{\Psi} M_{\Phi^*}^H \geq 0$, while $N_{\Phi^* \circ \Phi} > 0$ follows since $\Phi^* \circ \Phi$ is positive definite. \square

In §12 we shall produce iterations from A -orthogonal projections.

Definition 7.2. $\Phi \in \mathcal{L}$ is called an A -orthogonal projection if $\mathfrak{D}(\Phi) \ni A > 0$ implies that the matrix $A^{1/2} N[A] A^{1/2}$ is an orthogonal projection.

An orthogonal projection has a spectrum contained in $\{0, 1\}$. For our purpose, the following generalisation is sufficient:

$$\Phi \in \mathcal{L}_{\text{sym}} \text{ with } \sigma(N[A] A) \subset [0, 2]. \quad (7.1)$$

Definition 7.3. The iteration $\Phi(\cdot, \cdot, A) \in \mathcal{L}$ is called nonexpansive (with respect to an associated norm $\|\cdot\|$) if

$$\|M_{\Phi}[A]\| \leq 1.$$

Exercise 7.4. $A > 0$ and (7.1) imply that Φ is nonexpansive with respect to $\|\cdot\|_A$.

Lemma 7.5. Assume that $A > 0$. Let $\Phi_i \in \mathcal{L}$ satisfy (7.1) for $1 \leq i \leq k$. Then the product iteration

$$\Pi(\cdot, \cdot, A) := \Phi_k(\cdot, \cdot, A) \circ \dots \circ \Phi_2(\cdot, \cdot, A) \circ \Phi_1(\cdot, \cdot, A)$$

converges if and only if (5.10) holds (cf. Proposition 5.23).

Proof. (i) Let $x \in \bigcap_{i=1}^k \ker(N_{\Phi_i})$. For an indirect proof, assume $x \neq 0$ and set $y := A^{-1}x \neq 0$. Since $x \in \ker(N_{\Phi_1})$, $y = M_{\Phi_1}y$ holds for the iteration matrix $M_{\Phi_1} = I - N_{\Phi_1}A$. By $y = M_{\Phi_2}y$, etc., we obtain $y = (M_{\Pi})y$ for the iteration matrix $M_{\Pi} = \prod_{i=1}^k M_{\Phi_i}$ of $\Pi(\cdot, \cdot, A)$. The eigenvalue 1 of M_{Π} proves divergence of Π . Hence, convergence implies (5.10).

(ii) Assume that (5.10) holds and define

$$\hat{M}_i := A^{1/2} M_{\Phi_1} A^{-1/2} = I - A^{1/2} N_{\Phi_1} A^{1/2} \quad \text{and} \quad \hat{M}_{\Pi} := \prod_{i=1}^k \hat{M}_i.$$

The product iteration Π converges monotonically with respect to the energy norm if $\|\hat{M}_{\Pi}\|_2 < 1$. By (7.1), $\sigma(\hat{M}_i) \subset (-1, 1]$ and $\|\hat{M}_i x\|_2 \leq \|x\|_2$ hold for all $x \in \mathbb{K}^I$. In addition, $\|\hat{M}_i x\|_2 = \|x\|_2$ is equivalent to $A^{-1/2}x \in \ker(N_{\Phi_i})$. As a consequence $\|\hat{M}_{\Pi} x\|_2 \leq \|x\|_2$ holds for all $x \in \mathbb{K}^I$ and $\|\hat{M}_{\Pi} x\|_2 = \|x\|_2$ implies $A^{-1/2}x \in \bigcap_{i=1}^k \ker(N_{\Phi_i}[A])$. The assumption (5.10) yields $x = 0$. Hence $\|\hat{M}_{\Pi}\|_A = \|\hat{M}_{\Pi}\|_2 < 1$ follows. \square

7.2 Additive Splitting Technique

7.2.1 Definition and Examples

Most of the classical iterations are constructed by an additive¹ splitting as explained below. Given the system of equations

$$Ax = b \quad (A \in \mathbb{K}^{I \times I}, b \in \mathbb{K}^I), \quad (7.2)$$

we split A into the difference

$$A = W - R \quad (W \text{ regular}). \quad (7.3)$$

The system (7.2) is equivalent to

$$Wx = Rx + b.$$

This suggests the iterative method

$$Wx^{m+1} = Rx^m + b \quad (7.4)$$

which is well defined since W is required to be regular.

Lemma 7.6. (a) Assume (7.3). Then the iterative method (7.4) is consistent. The matrices of the first normal form (2.8) are

$$M = W^{-1}R, \quad N = W^{-1}.$$

The notation ‘ W ’ for the matrix in (7.3) is chosen because the third normal form (2.12),

$$W(x^m - x^{m+1}) = Ax^m - b,$$

is valid with the same matrix W .

(b) Vice versa, any iteration $\Phi \in \mathcal{L}$ with regular N can be obtained from an additive splitting (7.3).

Proof. (a) A comparison of the representation

$$x^{m+1} = W^{-1}Rx^m + W^{-1}b$$

derived from (7.4) with (2.8) shows that $M = W^{-1}R$ and $W = N^{-1}$.

(b) Choose $W := N_\Phi[A]^{-1}$ and $R := W - A$ in (7.3). □

Because of Lemma 7.6b, the additive splitting technique does not produce a special class of iterations but *all* linear iterations. This is a similar situation as the combination of the Richardson iteration Φ_1^{Rich} with a right transformation $T_\ell = N$

¹ The term ‘additive’ distinguishes this technique from the multiplicative factorisation in §7.3.

(cf. Proposition 5.44). In the case of the additive splitting, $W = N^{-1}$ is the primary quantity, whereas in the latter case, N determines the transformation.

Remark 7.7. The fact that the additive splitting can generate any linear iteration leads to the question: what are the data on which the choice of W can be based? The following cases can be distinguished:

- (i) The choice is only based on the data of the matrix A . This means that there is an explicitly available mapping $A \mapsto W[A]$ or $A \mapsto N[A]$. In this case, the iteration is *algebraic* (cf. Definition 2.2b).
- (ii) The matrix A may be the result of a discretised partial differential equation. Correspondingly, additional data of the partial differential equation not contained in the matrix data (e.g., geometric data, coarser discretisations, etc.) can be used for constructing W .
- (iii) An intermediate situation between (i) and (ii) is the following one. The element matrices $B = \{B^{(\nu)} : \nu \in J\}$ introduced in §E.3 contain more data than A . Therefore a mapping $B \mapsto W[B]$ may be well defined, but cannot be obtained from A (cf. Remark E.8b).

In §7.4.5 we shall give an example for case (ii). There the proposed matrix W cannot be derived from the matrix A .

A typical example of cases (ii) or (iii) are domain decomposition iterations involving submatrices discretising Neumann boundary problems in subdomains (cf. §12.3). These subproblems lead to matrices A_1 and A_2 such that $A = A_1 + A_2$. Obviously, A is a result of A_1 and A_2 , but these matrices cannot be determined from A .

All splittings discussed in this section and in §7.3 correspond to the case (i) of Remark 7.7.

Example 7.8. (a) A natural choice of W is some part of the matrix A . The splitting $A = D - (A - D)$ with the diagonal $W = D$ of A yields the Jacobi iteration.

(b) Starting from the splitting $A = D - E - F$ in (1.16), we choose $W = D - E$ and $R = F$. The resulting iteration (7.4) is the Gauss–Seidel iteration. Alternatively, the choice $W = D - F$ and $R = E$ yields the backward Gauss–Seidel method (cf. Proposition 5.1).

(c) Using the blockwise version of $A = D_{\text{block}} - E_{\text{block}} - F_{\text{block}}$ in (3.19a–d), the respective splitting yields the block-Jacobi and block-Gauss–Seidel iterations.

Note that in the previous examples the matrices D , $D - E$, $D_{\text{block}} - E_{\text{block}}$ contain increasing parts of the matrix A . In Theorem 7.13 and §7.2.3 we shall see that this fact may improve the convergence. On the other hand, W must still be (easily) invertible, since we have to solve the system (2.12').

The additive splitting can be combined with the summation introduced in §5.3 and yields the *multi-splitting method* (cf. O’Leary–White [296]).

7.2.2 Regular Splittings

In this section we shall make use of M-matrices (cf. §C.3). Accordingly, we use the notation

$$A < B, \quad A \leq B, \quad x < y, \quad x \leq y, \quad \dots$$

for matrices and vectors in the sense of *componentwise inequalities*. In particular, $A > 0$ denotes a positive matrix, not a positive definite one.

The following definition of a ‘regular splitting’ is due to Varga [375]. It allows not only qualitative convergence statements but also a comparison of different iterative methods.

Definition 7.9 (regular splitting). The real matrix $W \in \mathbb{R}^{I \times I}$ describes a *regular splitting* of $A \in \mathbb{R}^{I \times I}$ if

$$W \text{ regular}, \quad W^{-1} \geq 0, \quad W \geq A \text{ (i.e., } R := W - A \geq 0\text{)}. \quad (7.5)$$

Condition (7.5) may be compared with (3.35g) in the positive definite case. The iteration matrix of the iteration (7.4) is

$$M = W^{-1}R \quad \text{with} \quad R := W - A$$

(cf. Lemma 7.6a). Condition (7.5) implies that

$$M \geq 0 \quad \text{for regular splittings} \quad (7.6)$$

because of $R \geq 0$. Using (7.6), we can weaken Definition 7.9 (cf. Ortega [298]).

Definition 7.10 (weakly regular splitting). The splitting (7.3) is *weakly regular* if

$$W \text{ regular}, \quad W^{-1} \geq 0, \quad M = W^{-1}R \geq 0. \quad (7.7)$$

Theorem 7.11 (convergence). Let A be inverse positive: $A^{-1} > 0$ (a sufficient condition is that A be an M-matrix). Assume that W describes a weakly regular splitting of A . Then the induced iteration (7.4) converges:

$$\rho(M) = \rho(W^{-1}R) = \frac{\rho(A^{-1}R)}{1 + \rho(A^{-1}R)} < 1. \quad (7.8)$$

Proof. (i) Obviously, it is sufficient to show $\rho(W^{-1}R) = \rho(C)/(1 + \rho(C))$ for $C := A^{-1}R$. The weak regularity (7.7) implies that

$$\begin{aligned} 0 \leq M &= W^{-1}R = [A^{-1}W]^{-1}A^{-1}R \\ &= [A^{-1}(A + R)]^{-1}A^{-1}R = [I + C]^{-1}C. \end{aligned}$$

By Theorem C.34 and $M \geq 0$, there is an eigenvector $x \geq 0$ belonging to the eigenvalue $\lambda = \rho(M) \in \sigma(M)$. Rewriting $\lambda x = Mx = (I + C)^{-1}Cx$, we obtain

$$\lambda x + \lambda Cx = Cx \quad (7.9a)$$

The value $\lambda = 1$ is excluded, since (7.9a) would yield $x = 0$. Hence,

$$Cx = \frac{\lambda}{1-\lambda}x \quad (7.9b)$$

follows. In part (iii) we shall show that $C \geq 0$. Equation (7.9b), together with $x \geq 0$ and $Cx \geq 0$, ensures the inequality $\frac{\lambda}{1-\lambda} \geq 0$, i.e., $0 \leq \lambda = \rho(M) < 1$.

(ii) (7.9b) proves that λ is an eigenvalue of M if and only if $\mu = \frac{\lambda}{1-\lambda}$ is an eigenvalue of C . The inequality $0 \leq \lambda < 1$ shows that $\mu \geq 0$. Since $\mu = \frac{\lambda}{1-\lambda}$ increases monotonically in λ , $|\mu| = \mu$ is maximal for $\lambda = \rho(M) \in \sigma(M)$. By Theorem C.34, $\mu = \rho(C) \in \sigma(C)$ is the maximal eigenvalue of C ; therefore we have $\rho(C) = \rho(M)/[1 - \rho(M)]$. Solving this equation for $\rho(M)$, we arrive at assertion (7.8): $\rho(M) = \rho(C)/[1 + \rho(C)]$.

(iii) From

$$0 \leq \left[\sum_{\nu=0}^{m-1} M^\nu \right] W^{-1}, \quad W^{-1} = (I - M)A^{-1}, \quad \text{and}$$

$$\sum_{\nu=0}^{m-1} M^\nu (I - M) = I - M^m,$$

we conclude that

$$0 \leq (I - M^m)A^{-1} \leq A^{-1} \quad \text{and} \quad 0 \leq M^m A^{-1} \leq A^{-1}.$$

Therefore, M^m is bounded. This fact proves that $\kappa = \rho(M) < 1$. Since $\lambda = 1$ is already excluded, $\rho(M) < 1$ holds and implies

$$C = A^{-1}R = [W(I - M)]^{-1}R = (I - M)^{-1}W^{-1}R = \left[\sum_{\nu=0}^{\infty} M^\nu \right] M \geq 0. \quad \square$$

It might be expected that the iteration converges faster the closer W is to A , i.e., the smaller the remainder $R = W - A$ is. This property is stated more precisely in the following *comparison theorem*.

Theorem 7.12. *Let A be inverse positive: $A^{-1} \geq 0$. Let W_1 and W_2 define two regular splittings. If W_1 and W_2 are comparable in the sense of*

$$A \leq W_1 \leq W_2, \quad (7.10a)$$

then the convergence rates satisfy the corresponding inequalities

$$0 \leq \rho(M_1) \leq \rho(M_2) < 1, \quad \text{where } M_i := W_i^{-1}R_i, \quad R_i := W_i - A. \quad (7.10b)$$

Proof. The matrices $B := A^{-1}R_1$ and $C := A^{-1}R_2$ satisfy $0 \leq B \leq C$ and therefore $0 \leq \rho(B) \leq \rho(C)$ (cf. (C.15)). From representation (7.8) we obtain

$$0 \leq \rho(M_1) = \rho(B)/[1 + \rho(B)] \leq \rho(C)/[1 + \rho(C)] = \rho(M_2) < 1. \quad \square$$

The comparisons in (7.10a,b) can be strengthened into strict inequalities.

Theorem 7.13. *From $A^{-1} > 0$ and*

$$A \trianglelefteq W_1 \trianglelefteq W_2, \quad W_1, W_2: \text{ regular splittings}, \quad (7.11a)$$

the strict inequalities

$$0 < \rho(M_1) < \rho(M_2) < 1 \quad \text{with } M_i := W_i^{-1}R_i, \quad R_i := W_i - A \quad (7.11b)$$

follows.

Proof. Define B and C as in the previous proof. Since $B = A^{-1}R_1$ may be reducible, Theorem C.25 is not directly applicable. $R_1 \geq 0$ holds since the splitting is regular. Define

$$I_+ := \{\beta \in I : R_{1,\alpha\beta} > 0 \text{ for some } \alpha \in I\} \quad \text{and} \quad I_0 := I \setminus I_+.$$

Any column $s = (R_{1,\alpha\beta})_{\alpha \in I}$ of R_1 corresponding to an index $\beta \in I_+$ satisfies $s \trianglerighteq 0$ and therefore $A^{-1}s > 0$ by Exercise C.20b. Hence, B has the form

$$B = \begin{bmatrix} B_1 & 0 \\ B_2 & 0 \end{bmatrix} \quad \text{with positive blocks } B_1 > 0 \quad \text{and} \quad B_2 > 0$$

corresponding to the block structure $\{I_+, I_0\}$. In particular,

$$\rho(B) = \rho(B_1) > 0 \quad (7.11c)$$

holds (cf. (C.11a)). Because of $R_2 - R_1 = W_2 - W_1 \trianglerighteq 0$, there is a pair (α, β) with $(R_2 - R_1)_{\alpha\beta} > 0$. Hence, the column of $C - B = A^{-1}(R_2 - R_1)$ for the index β is positive. Assume $\beta \in I_+$. In this case, $C_1 \trianglerighteq B_1$ and $C_2 \trianglerighteq B_2$ hold for the blocks in $C = \begin{bmatrix} C_1 & C_3 \\ C_2 & C_4 \end{bmatrix}$. Lemma C.30 and (C.11c) yield the inequality

$$\rho(C) \geq \rho(C_1) > \rho(B_1).$$

In the remaining case of $\beta \in I_0$, we conclude that

$$C_3 \trianglerighteq B_3 = 0, \quad C_4 \trianglerighteq B_4 = 0,$$

and

$$\rho(C) > \rho(C_1) \geq \rho(B_1)$$

(cf. Lemma C.30). In any case, using (7.11c), we arrive at the strict inequality $\rho(C) > \rho(B) > 0$, which via (7.8) leads us to the assertion. \square

7.2.3 Applications

Theorem 7.14. Let A be an M-matrix. Then the point- and blockwise Jacobi iterations converge. Moreover, the blockwise iteration is faster:

$$\rho(M^{\text{blockJac}}) \leq \rho(M^{\text{Jac}}) < 1. \quad (7.12a)$$

Let D be the pointwise diagonal D^{ptw} or the block diagonal D^{block} of A . Then

$$D \text{ describes a regular splitting.} \quad (7.12b)$$

Assuming explicitly (7.12b), we may replace the assumption ‘ A is an M-matrix’ by the inverse positivity: $A^{-1} \geq 0$. The strict inequality

$$0 < \rho(M^{\text{blockJac}}) < \rho(M^{\text{Jac}}) < 1$$

holds instead of (7.12a) if $A^{-1} > 0$ and $D^{\text{ptw}} \neq D^{\text{block}} \neq A$.

Proof. For an M-matrix A , the diagonals $D = D^{\text{ptw}}$ and $D = D^{\text{block}}$ satisfy the inequality $D > A$ and the sign condition (C.18b). By Theorem C.53, D is again an M-matrix, so that $D^{-1} \geq 0$ and (7.12b) follow. Because of $D^{\text{ptw}} \geq D^{\text{block}}$, Theorem 7.12 proves inequality (7.12a). Concerning the strict inequality, compare with Theorem 7.13. \square

Theorem 7.15. Split $A = D - E - F$ according to (3.11a–d) or (3.19a–d). The statements of Theorem 7.14 carry over to analogous ones for the pointwise and blockwise Gauss–Seidel iteration, where the statements (7.12a,b) become

$$\rho(M^{\text{blockGS}}) \leq \rho(M^{\text{GS}}) < 1, \quad D - E \text{ describes a regular splitting.}$$

We omit the proof, since it is completely analogous to the previous one. The comparison between the Jacobi and Gauss–Seidel iteration is more interesting. The quantitative relation $\rho(M^{\text{GS}}) = \rho(M^{\text{Jac}})^2$, which according to Conclusion 4.30 holds for consistent orderings, can no longer be shown for the general case. However, a corresponding qualitative statement derived from $D - E \leq D$ is valid.

Theorem 7.16. For an M-matrix A , the following inequalities hold:

$$\rho(M^{\text{GS}}) \leq \rho(M^{\text{Jac}}) < 1, \quad \rho(M^{\text{blockGS}}) \leq \rho(M^{\text{blockJac}}) < 1.$$

This statement can be generalised to other than M-matrices.

Theorem 7.17 (Stein–Rosenberg [352]). Exactly one of the following four alternatives holds for the pointwise Jacobi and Gauss–Seidel iterations if A fulfils the sign condition (C.18b), $a_{\alpha\beta} \leq 0$ for $\alpha \neq \beta$:

$$0 = \rho(M^{\text{GS}}) = \rho(M^{\text{Jac}}),$$

$$0 < \rho(M^{\text{GS}}) < \rho(M^{\text{Jac}}) < 1,$$

$$\rho(M^{\text{GS}}) = \rho(M^{\text{Jac}}) = 1,$$

$$\rho(M^{\text{GS}}) > \rho(M^{\text{Jac}}) > 1.$$

In particular, both methods converge or diverge simultaneously. The statement of the theorem remains valid if M^{Jac} and M^{GS} are replaced by $L + U$ and $(I - L)^{-1}U$ with $L \geq 0$ being an arbitrary, strictly lower triangular matrix and $U \geq 0$ a strictly upper one.

The proof can be found in Varga [375, §3.3] or in the original paper [352]. For generalisations, see Buoni–Varga [88, 89].

In the case of overrelaxation (i.e., for $\omega > 1$), the SOR iteration does not lead to a regular splitting. To ensure regularity of the splitting, we have to restrict the parameter ω to $0 < \omega < 1$ (underrelaxation).

Exercise 7.18. Prove that the SOR iteration arises from a splitting (7.3) with $W = \omega^{-1}D - E$. Let A be an M-matrix and D its diagonal. For $0 < \omega \leq 1$, the matrix W describes a regular splitting. What conclusion can be drawn from $\omega^{-1}D - E \geq D - E$?

In the case of a regular splitting, the property (7.7) (i.e., $M \geq 0$) allows an enclosure of the solution $x = A^{-1}b$, provided that we find suitable starting iterates.

Theorem 7.19. Let $M \geq 0$ be the iteration matrix of a convergent iteration. Starting with initial iterates x^0 and y^0 satisfying

$$x^0 \leq x^1, \quad x^0 \leq y^0, \quad y^1 \leq y^0,$$

we obtain iterates x^m and y^m with the enclosure property

$$x^0 \leq x^1 \leq \dots \leq x^m \leq \dots \leq x = A^{-1}b \leq \dots \leq y^m \leq \dots \leq y^1 \leq y^0.$$

Proof. It follows from the estimates $x^{m+1} - x^m = M^m(x^1 - x^0) \geq 0$, and $y^m - y^{m+1} = M^m(y^0 - y^1) \geq 0$, $y^m - x^m = M^m(y^0 - x^0) \geq 0$ (cf. (2.16b)). \square

We recall the generalisation of the M-matrices by the H-matrices in Definition C.60 and the definition of diagonal dominance in §C.3.3.

Theorem 7.20. Each of the following conditions (7.13a,b) is sufficient for the convergence of the pointwise Jacobi and Gauss–Seidel iterations:

$$A \text{ is an H-matrix}, \tag{7.13a}$$

$$A \text{ is strictly, irreducibly, or essentially diagonally dominant}. \tag{7.13b}$$

Exercise 7.21. Prove that (7.13b) implies (7.13a) and $\|M^{\text{Jac}}\|_{\infty} \leq 1$, $\|M^{\text{GS}}\|_{\infty} \leq 1$.

Proof. (i) The case (7.13b) is reduced to (7.13a) because of Exercise 7.21.

(ii) Define $B := |D| - |A - D|$ as in Definition C.60 and denote the iteration matrix of the Jacobi iteration for B by $M_B^{\text{Jac}} := |D|^{-1}|A - D|$. Theorem 7.16 yields $\rho(M_B^{\text{Jac}}) < 1$. By $|M_B^{\text{Jac}}| = M_B^{\text{Jac}}$, the convergence $\rho(M^{\text{Jac}}) < 1$ follows from the next lemma, which remains to be proved.

Lemma 7.22. $\rho(A) \leq \rho(|A|)$ for all $A \in \mathbb{C}^{I \times I}$.

(iii) Split $A = D - E - F$ according to (3.11a–d) and define $L := D^{-1}E$, $U := D^{-1}F$. Since $B = |D| - |E| - |F| = |D|(I - |L| - |U|)$, the iteration matrices belonging to A and B are:

$$M^{\text{GS}} = (I - L)^{-1}U = \sum_{\nu=0}^{\infty} L^\nu U, \quad M_B^{\text{GS}} = (I - |L|)^{-1}|U| = \sum_{\nu=0}^{\infty} |L|^\nu |U|$$

(cf. Lemma A.13). Hence, $|M^{\text{GS}}| = |\sum_{\nu=0}^{\infty} L^\nu U| \leq \sum_{\nu=0}^{\infty} |L|^\nu |U| = M_B^{\text{GS}}$. From Lemma 7.22 and Theorem 7.15, we conclude that $\rho(M^{\text{GS}}) \leq \rho(M_B^{\text{GS}}) < 1$. \square

Proof of Lemma 7.22. By $\|A^\nu\|_\infty = \| |A^\nu| \|_\infty \leq \| |A|^\nu \|_\infty$, Theorem B.27 yields

$$\rho(A) = \lim_{\nu \rightarrow \infty} \|A^\nu\|_\infty^{1/\nu} \leq \lim_{\nu \rightarrow \infty} \| |A|^\nu \|_\infty^{1/\nu} = \rho(|A|). \quad \square$$

The diagonal dominance in (7.13b) is often used as a convergence criterion since the proof becomes very simple. Strict diagonal dominance is historically the first convergence criterion for the Jacobi iteration (see the paper of R. von Mises and H. Pollaczek-Geiringer [381, Satz 2] from 1929).

Proposition 7.23. If the strict diagonal dominance (C.16) can be quantified by a number $q > 1$ such that

$$|a_{\alpha\alpha}| \geq q \sum_{\beta \in I \setminus \{\alpha\}} |a_{\alpha\beta}| \quad \text{for all } \alpha \in I, \quad (7.14a)$$

then the Jacobi and Gauss–Seidel iterations converge monotonically with respect to the maximum norm with the contraction numbers

$$\|M^{\text{Jac}}\|_\infty, \|M^{\text{GS}}\|_\infty \leq 1/q < 1. \quad (7.14b)$$

Proof. Using (7.14a), the estimate of $M^{\text{Jac}} = D^{-1}(A - D)$ by $\|M^{\text{Jac}}\|_\infty \leq 1/q$ follows immediately from (B.8).

In the Gauss–Seidel case, we use the description of the iteration by (1.15). The components of the error $e^m = x^m - x$ satisfy

$$e_i^{m+1} = - \left(\sum_{j=1}^{i-1} a_{ij} e_j^{m+1} + \sum_{j=i+1}^n a_{ij} e_j^m \right) / a_{ii}.$$

Induction on i yields $\|e^{m+1}\|_\infty \leq \|e^m\|_\infty / q$. Since $e^{m+1} = M^{\text{GS}} e^m$, the inequality (7.14b) follows. \square

Concerning the convergence of the SSOR iteration for H-matrices, we refer to Alefeld–Varga [3] and Neumaier–Varga [289].

7.2.4 P-Regular Splitting

The P-regular splitting defined below is of different nature. In particular, it is based on the order relation of positive definite matrices (cf. §C.1). The term ‘P-regular’ is introduced by Ortega [298], but the following convergence statement goes back to Weissinger [392] in 1953 (see also Weissinger [391]).

Lemma 7.24. *Let X be any general matrix, while Z is positive definite; i.e., $Z > 0$. Then $Z - X^H ZX > 0$ implies that*

$$\rho(X) < 1 \text{ and } \|Z^{1/2} X Z^{-1/2}\|_2 < 1.$$

Proof. Set $Y := Z^{1/2} X Z^{-1/2}$. Multiplying $Z - X^H ZX > 0$ by $Z^{-1/2}$ from both sides yields $I - Y^H Y > 0$ or $Y^H Y < I$ (cf. (C.3a')). Hence $\|Y\|_2^2 = \rho(Y^H Y) < \rho(I) = 1$ proves the last statement. Since X and Y are similar matrices, $\rho(X) = \rho(Y) \leq \|Y\|_2$ proves $\rho(X) < 1$. \square

Definition 7.25. The splitting $A = W - R$ is called *P-regular* if W is regular and the Hermitian part $\frac{1}{2}(C + C^H)$ of $C := W + R$ is positive definite.

The last condition can be written as $0 < \frac{1}{2}(C + C^H) = \frac{1}{2}(W + W^H + R + R^H) = W + W^H - \frac{1}{2}(A + A^H)$, i.e.,

$$W + W^H > \frac{1}{2}(A + A^H) =: \hat{A}. \quad (7.15)$$

Theorem 7.26 (Weissinger [392]). *Assume $A + A^H > 0$ and consider a P-regular splitting $A = W - R$. The corresponding iteration (7.4) converges monotonically with respect to the norm $\|\cdot\|_{\hat{A}}$ with \hat{A} defined in (7.15):*

$$\rho(M) \leq \|M\|_{\hat{A}} < 1 \quad \text{for } M = I - W^{-1}A. \quad (7.16)$$

Proof. The splitting $A = W - R$ yields the iteration matrix $M = W^{-1}R$. Note that

$$\begin{aligned} A - M^H A M &= A - (I - W^{-1}A)^H A (I - W^{-1}A) \\ &= (W^{-1}A)^H A + AW^{-1}A - (W^{-1}A)^H A (W^{-1}A) \\ &= (W^{-1}A)^H (W + W^H - A)(W^{-1}A) =: B. \end{aligned}$$

Forming the expression $\frac{1}{2}(B + B^H)$ and using \hat{A} in (7.15), we arrive at

$$\hat{A} - M^H \hat{A} M = (W^{-1}A)^H (W + W^H - \hat{A})(W^{-1}A) > 0$$

because of (7.15). Lemma 7.24 with $Z := \hat{A}$ yields (7.16). \square

7.3 Incomplete Triangular Decompositions

One learns from Theorem 7.13 that the convergence speed of Jacobi and Gauss–Seidel iterations could be improved if even larger parts of the matrix A were contained in W . The practical obstacle is that we must be able to solve the system $W\delta = d$ efficiently. In particular, this requirement seems to exclude splittings with W containing larger portions of A than the lower and upper triangular parts. However, if we are able to decompose W into triangular factors²

$$W = LU \quad (L \text{ lower triangular, } U \text{ upper triangular matrix}),$$

the solution of $LU\delta = d$ can easily be performed using the forward and backward substitution (cf. Quarteroni–Sacco–Saleri [314, §3.2]).

Therefore, we are looking for a suitable matrix $W = LU$. In general, $W = A$ is not a good candidate since its LU decomposition leads to a fill-in, i.e., to larger nonzero parts of the matrix. In the case of sparse factors L, U with $A \neq W = LU$, this factorisation is called an *incomplete LU decomposition* of A and abbreviated as ILU.

Besides the use of ILU as a linear iteration (possibly accelerated by techniques of Part II), ILU is also of interest as smoothing iteration of the multigrid method (cf. §11.9.2).

7.3.1 Introduction and ILU Iteration

In the following, the index set I is ordered. Here the standard choice in the model case is the lexicographical ordering. By Conclusion 1.11, the LU decomposition $A = LU$ has proved to be inappropriate for sparse matrices, since the factors L and U contain many more nonzero entries than the original matrix A . Computing the LU decomposition is completely identical to Gauss elimination: U is the upper triangular matrix remaining after eliminating the entries below the diagonal, whereas L contains the elimination factors $L_{ji} = a_{ji}^{(i)} / a_{ii}^{(i)}$ ($j \geq i$) (cf. Quarteroni–Sacco–Saleri [314, §3.3]). Instead of computing L and U by Gauss elimination, we may determine the $n^2 + n$ unknown entries L_{ji}, U_{ij} ($j \geq i$) directly from the n normalisation conditions

$$L_{ii} = 1 \quad (1 \leq i \leq n) \tag{7.17a}$$

and the n^2 equations involved in $A = LU$:

$$\sum_{j=1}^n L_{ij} U_{jk} = A_{ik} \quad (1 \leq i, k \leq n). \tag{7.17b}$$

² In this section, L and U are general (nonstrict) triangular matrices and do not coincide with the matrices L, U defined in (3.15d).

The incomplete LU decomposition is based on the idea of not eliminating all matrix entries of A to avoid the fill-in of the matrix during the elimination process. Since, after an *incomplete elimination*, entries remain in the lower triangular part, an exact solution of the system is not possible. Instead, the previous equality $A = LU$ holds up to remainder R :

$$A = LU - R. \quad (7.18)$$

For the exact description of the ILU process, we choose a subset $E \subset I \times I$ of the product of the ordered index set $I = \{1, 2, \dots, n\}$. The elimination is restricted to the pairs $(i, j) \in E$. Concerning E , we always require

$$(i, i) \in E \quad \text{for all } i \in I. \quad (7.19a)$$

In general, one should choose E large enough, so that the graph $G(A)$ of A is contained in E (cf. Definition C.12):

$$G(A) \subset E. \quad (7.19b)$$

E is called the (*elimination*) *pattern* of the ILU decomposition. Examples of E will be given in §7.3.2. Through the definition of the triangular matrices, we have

$$L_{ij} = U_{ji} = 0 \quad \text{for } 1 \leq i < j \leq n. \quad (7.20a)$$

To construct *sparse* matrices L and U , nonzero entries are allowed only at positions of the pattern E ; otherwise, we require

$$L_{ij} = U_{ij} = 0 \quad \text{for } (i, j) \notin E. \quad (7.20b)$$

Exercise 7.27. Prove that there are $\#E$ matrix entries of L and U which are not directly determined by (7.17a), (7.20a), and (7.20b).

In analogy to (7.17b), we pose $\#E$ equations for the same number of unknowns:

$$\sum_{j=1}^n L_{ij} U_{jk} = A_{ik} \quad \text{for all } (i, k) \in E. \quad (7.20c)$$

The remainder $R = LU - A$ is obtained from (7.20d,e):

$$R_{ik} = 0 \quad \text{for all } (i, k) \in E, \quad (7.20d)$$

$$R_{ik} = \sum_{j=1}^n L_{ij} U_{jk} - A_{ik} \quad \text{for all } (i, k) \notin E. \quad (7.20e)$$

Under assumption (7.19b), the term A_{ik} in 7.20e may be omitted because of $A_{ik} = 0$.

The ILU factors satisfying (7.17a) and (7.20a–c) can, e.g., be constructed by the following algorithm:

```

 $L := 0; U := 0;$ 
for  $i := 1$  to  $n$  do
  begin  $L_{ii} := 1;$ 
    for  $k := 1$  to  $i - 1$  do if  $(i, k) \in E$  then  $L_{ik} := \frac{A_{ik} - \sum' L_{ij} U_{jk}}{U_{kk}}$ ;   (7.21a)
    for  $k := 1$  to  $i$  do if  $(k, i) \in E$  then  $U_{ki} := A_{ki} - \sum'' L_{kj} U_{ji}$    (7.21b)
  end;
```

The sums \sum' and \sum'' are taken over all j with $j \neq k$. Since all indices referring to vanishing terms can be omitted, we may write:

$$\Sigma' = \sum_{j \in I \text{ with } j < k, (i,j) \in E, (j,k) \in E}, \quad \Sigma'' = \sum_{j \in I \text{ with } j < k, (k,j) \in E, (j,i) \in E}.$$

The definition of L_{ik} in (7.21a) is obtained from (7.20c). To prove (7.21b), interchange i and k in (7.20c). One verifies that only those components of L and U are involved in the right-hand sides of (7.21a,b) that are already computed. Remark 7.28 will enable a simplification of the algorithm.

Remark 7.28. The definitions $D := \text{diag}\{U\}$, $U' := U - D$, $L' := (L - I)D$ lead to a strictly lower triangular matrix L' and a strictly upper triangular matrix U' . Equation (7.18) rewritten with the new quantities becomes

$$A = (D + L')D^{-1}(D + U') - R. \quad (7.22)$$

The quantities D , L' , and U' are the result of the following algorithm:

```

 $D := 0; L' := 0; U' := 0;$ 
for  $i := 1$  to  $n$  do
  begin
    for  $k := 1$  to  $i - 1$  do if  $(i, k) \in E$  then  $L'_{ik} := A_{ik} - \sum' L'_{ij} D_{jj}^{-1} U'_{jk}$ ;   (7.23a)
    for  $k := 1$  to  $i - 1$  do if  $(k, i) \in E$  then  $U'_{ki} := A_{ki} - \sum'' L'_{kj} D_{jj}^{-1} U'_{ji}$ ;   (7.23b)
     $D_{ii} := A_{ii} - \sum'' L'_{ij} D_{jj}^{-1} U'_{ji}$    (7.23c)
  end;
```

Hence, ILU iteration based on L', D, U' is algebraic.

Remark 7.29. (a) If A is Hermitian, (7.23a–c) immediately implies the symmetries $L' = U'^H$ and $D = D^H$.

(b) The incomplete Cholesky decomposition $A = L'' L''^H - R$ for positive definite matrices A follows from (7.22) with $L'' := (D + L')D^{-1/2}$.

Tacitly, we assume that the quantities U_{kk} (pivot entries) in (7.21a) and D_{jj} in (7.23a) do not vanish and that, in the case of Remark 7.29b, even $D_{jj} > 0$ holds. Concerning these assumptions, we refer to the analysis in §7.3.5.

Exercise 7.30. Complete LU decompositions are characterised by $R = 0$ in (7.18). Prove: (a) $R = 0$ holds for cases (i) $E = I \times I$ or (ii) $E = \{(i, j) : |i - j| \leq w\}$ for band matrices of band width $w \geq 0$.

(b) $D = \text{diag}\{A\}$ and $L' = U' = 0$ hold for the diagonal elimination pattern $E = \{(i, i) : i \in I\}$, which is the minimal pattern satisfying (7.19a).

The additive splitting $A = W - R$ of A given by (7.18) or (7.22) defines the corresponding ILU iteration:

$$W(x^m - x^{m+1}) = Ax^m - b \quad \text{with} \quad (7.24a)$$

$$W = LU \quad \text{or} \quad W = (D + L')D^{-1}(D + U'), \text{ respectively.} \quad (7.24b)$$

The matrices of the first and second normal forms are

$$M = NR \quad \text{with} \quad N = U^{-1}L^{-1} \quad \text{or} \quad N = (D + U')^{-1}D(D + L')^{-1}.$$

Remark 7.31. In addition to the factors L, U (or D, L', U' , respectively), we can either store A and use (7.24a) or store R and apply the representation (7.25):

$$Wx^{m+1} = b + Rx^m. \quad (7.25)$$

Concerning the computational work, we recall §2.3.1: the decomposition (7.23a-c) defines the initialisation cost denoted by $\text{Init}(\Phi^{\text{ILU}}, A)$, while $\text{Work}(\Phi^{\text{ILU}}, A)$ is the cost required by (7.25).

7.3.2 Incomplete Decomposition with Respect to a Star Pattern

For the description of the pattern E , we should not use the ordered indices $1, \dots, n$. In the case of the model problem, the pairs (i, j) for $1 \leq i, j \leq N - 1$ are taken as indices of I . The edges of the graph $G(A)$ are described by the pairs $((i, j), (i \pm 1, j))$ (horizontal neighbours) and $((i, j), (i, j \pm 1))$ (vertical neighbours). For the case of a regular grid, the *star notation* was already used in §1.3.2 as short-hand notation of the matrices. In the following, we use the so-called star patterns. The entries ‘*’ in the examples

$$\begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \end{bmatrix}, \quad \begin{bmatrix} * & * \\ * & * & * \\ * & * \end{bmatrix}, \quad \begin{bmatrix} \cdot & \cdot & \cdot & * \\ \cdot & \cdot & \cdot \\ * & \cdot & \dots \end{bmatrix}$$

refer to elements in the set E . If, for instance, ‘*’ is the right neighbour of the midpoint, this means that for all $\alpha \in I$ having a right neighbour $\beta \in I$, the pair (α, β) belongs to E . Unmarked positions or the sign ‘·’ signify that the corresponding pairs (α, β) do not belong to E .

Remark 7.32. The 1×1 star $[*]$ characterises the minimal set $E = \{(i, i) : i \in I\}$ of Exercise 7.30b. The corresponding ILU iteration coincides with the Jacobi iteration.

7.3.3 Application to General Five-Point Formulae

Algorithm (7.23a,b) should be regarded more as a definition than a method for practically computing the matrices D , L' , U' . For the example of a general five-point formula A , we demonstrate how to derive a cheaper computation. For the sake of convenience, we assume that the coefficients are constant:

$$A = \begin{bmatrix} -e \\ -a & d - b \\ -c \end{bmatrix} \quad (\text{cf. (1.13a)}). \quad (7.26)$$

To ensure that A be an M-matrix, we require that

$$a, b, c, e \geq 0, \quad d \geq a + b + c + e.$$

The smallest pattern satisfying (7.19b) is

$$E = G(A), \quad \text{i.e., } E = \begin{bmatrix} * \\ * & * & * \\ * \end{bmatrix} \quad (\text{five-point pattern}). \quad (7.27a)$$

Using lexicographical ordering, the strictly triangular matrix L' has the pattern

$$\begin{bmatrix} & \cdot & \\ * & \cdot & \cdot \\ & * \end{bmatrix},$$

since the $*$ -marked positions are the only matrix entries corresponding to the pattern E and located below the diagonal. Correspondingly, U' has the pattern

$$\begin{bmatrix} & * & \\ \cdot & \cdot & * \\ \cdot & \cdot \end{bmatrix}.$$

In (7.23a,b), we replace the indices $i, j, k \in \{1, \dots, n\}$ by $\alpha, \beta, \gamma \in I = \Omega_h$ and, subsequently, we identify $\alpha = (x, y) = (k_\alpha h, l_\alpha h) \in \Omega_h$ with the pair (k_α, ℓ_α) , where now $1 < k_\alpha, \ell_\alpha < N - 1$ holds (cf. (1.3)). First, one has to discuss the sum Σ' in (7.23a). $L'_{\alpha\gamma} \neq 0$ can only be true for $\gamma = (k_\gamma, \ell_\gamma) = (k_\alpha - 1, \ell_\alpha)$ or $\gamma = (k_\alpha, \ell_\alpha - 1)$, whereas $U'_{\gamma\beta} \neq 0$ leads to $\beta = (k_\gamma + 1, \ell_\gamma)$ or $\beta = (k_\gamma, \ell_\gamma + 1)$. Hence,

$$L'_{\alpha\gamma} D_{\gamma\gamma}^{-1} U'_{\gamma\beta} \neq 0$$

requires $\beta = \alpha$ or $\beta = (k_\alpha + 1, \ell_\alpha - 1)$. Both possibilities contradict the inequality $\alpha \neq \beta$ —in (7.23a) written as $k \leq i - 1$ —and $(\alpha, \beta) \in E$. Therefore, Σ' is an empty sum and (7.23a) reduces to $L'_{\alpha\beta} = A_{\alpha\beta}$ for $\alpha > \beta$ and $(\alpha, \beta) \in E$. Hence, L' is the constant two-point star

$$L' = \begin{bmatrix} 0 & & \\ -a & 0 & 0 \\ & -c & \end{bmatrix}. \quad (7.27b)$$

Similarly, we obtain

$$U' = \begin{bmatrix} -e & & \\ 0 & 0 & -b \\ 0 & \end{bmatrix}. \quad (7.27c)$$

Only for $\alpha = \beta$, is the sum Σ'' in (7.23c) not empty and does contain the two indices $\gamma = (i_\alpha - 1, j_\alpha)$ and $\gamma = (i_\alpha, j_\alpha - 1)$. We abbreviate the diagonal entry $D_{\alpha\alpha}$ by $d_\alpha = d_{i_\alpha, j_\alpha}$. Because of $A_{\alpha\alpha} = d$ and the already known values in (7.27b,c), definition (7.23c) can be rewritten as

$$d_{i,j} = d - \frac{ab}{d_{i-1,j}} - \frac{ce}{d_{i,j-1}} \quad (1 \leq i, j \leq N-1), \quad (7.27d)$$

where the terms with $j-1=0$ or $i-1=0$ have to be ignored. In particular, we obtain $d_{11} = d$ for the first grid point. For the five-point formula (7.26), the double loop in (7.23a–c) is reduced to a simple loop over all $(i, j) \in I = \Omega_h$.

It is also possible to determine the remainder matrix R . Equations (7.20d,e) become

$$\begin{aligned} R_{\alpha\beta} &= 0 && \text{for } (\alpha, \beta) \in E, \\ R_{\alpha\beta} &= (L'D^{-1}U')_{\alpha\beta} && \text{for } (\alpha, \beta) \notin E. \end{aligned} \quad (7.27e)$$

One verifies that R has two (variable) coefficients per row:

$$R = \begin{bmatrix} r_{ij} & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & s_{ij} \end{bmatrix} \quad \text{with } r_{ij} = \frac{ae}{d_{i-1,j}}, \quad s_{ij} = \frac{cb}{d_{i,j-1}}, \quad (7.27f)$$

where $r_{ij} = 0$ holds for $i = 1$ and $s_{ij} = 0$ for $j = 1$.

Remark 7.33. The ILU decomposition of a five-point formula with constant or variable coefficients requires $6n$ operations for computing the d_{ij} values in (7.27d). The solution of

$$W\delta = (D + L')D^{-1}(D + U')\delta = d$$

takes $10n$ operations; hence because of the additional $10n$ operations for computing $d = Ax^m - b$, one ILU iteration step (7.24a) requires, in total, $21n$ operations. Note that the d_{ij} values in (7.27d) have to be determined only once. An alternative is determining R by additional $4n$ operations. Afterwards, the iteration (7.25) requires only $14n$ operations. Together with $C_A = 5$, the following cost factors result:

$$C^{\text{ILU}} = 4.2 \quad \text{or} \quad C^{\text{ILU}} = 2.8 \quad \text{respectively for } E \text{ in (7.27a).}$$

7.3.4 Modified ILU Decompositions

So far we ignored matrix entries a_{ij} for $(i, j) \notin E$ completely. One may pose the question of whether or not this is a good strategy. The following approach will indirectly use all a_{ij} .

We recall the Gauss–Seidel iteration, where the matrix $W = D - E$ is changed into $W = \frac{1}{\omega}D - E$ for the SOR method. Hence, overrelaxation, which in general leads to improved convergence, corresponds to diminishing the diagonal in W . Following Wittum [403], we introduce a modification which also leads to a diminishing or enlargement of the diagonal depending on the choice of ω .

Let $\mathbf{1}$ be the vector $(1)_{\alpha \in I}$ consisting of the entries $\mathbf{1}_\alpha = 1$. Gustafsson [172] proposes replacing the equation $R_{ii} = 0$ (i.e., (7.20d) for $i = k$) by

$$A\mathbf{1} = W\mathbf{1}, \quad \text{i.e.,} \quad R\mathbf{1} = 0. \quad (7.28)$$

One may view $\mathbf{1}$ as a *test vector*. By condition (7.28), W is gauged in such a way that A and W coincide with respect to their application to $\mathbf{1}$. We generalise the condition $R\mathbf{1} = 0$ by

$$R_{ii} = \omega \sum_{j \neq i} R_{ij} \quad (\omega \in \mathbb{R}) \quad (7.29)$$

and denote the corresponding decomposition as ILU_ω decomposition (its existence is not yet claimed). The corresponding ILU_ω iteration is denoted by $\varPhi_\omega^{\text{ILU}}$.

Remark 7.34. (a) For $\omega = 0$, Eq. (7.29) coincides with (7.20d) for $i = k$: $R_{ii} = 0$. Hence, the unmodified ILU decomposition is the ILU_0 decomposition.

(b) For $\omega = -1$, the conditions (7.28) and (7.29) are identical, i.e., the ILU_{-1} decomposition describes the modification by Gustafsson [172].

In the case of the five-point formula (7.26) and the five-point pattern (7.27a), L' and U' are still obtainable from (7.27b,c), whereas recursion (7.27d) for the entries d_{ij} of D becomes

$$d_{ij} := d + \frac{(\omega e - b)a}{d_{i-1,j}} + \frac{(\omega b - e)c}{d_{i,j-1}} \quad (7.30)$$

(terms with $i - 1 = 0$ and $j - 1 = 0$ are again to be ignored).

7.3.5 Existence and Stability of the ILU Decomposition

In this section, the inequalities $A \leq B$ have to be understood in the sense of elementwise inequalities $A_{\alpha\beta} \leq B_{\alpha\beta}$ ($\alpha, \beta \in I$) as in §C.3.

It is well known that the (complete) LU decomposition exists if and only if all principal submatrices $(a_{ij})_{1 \leq i,j \leq k}$ are regular for $1 \leq k \leq n$. However, even if the decomposition $A = LU$ exists, it can be useless since the solution process of

the equations $Ly = b$ and $Ux = y$ may be unstable. Choose, e.g., $A = LU$ with $U = L^\top$ and $L = \text{tridiag}\{\alpha, 1, 0\}$ for $\alpha < 1$, and investigate the error propagation (cf. Elman [121]). The criterion involving the principal submatrices is satisfied for positive definite matrices. However, there are positive definite matrices for which the ILU decomposition fails because of $U_{kk} = 0$ in (7.21a). The first part of the following criterion is stated by Meijerink–van der Vorst [280], while the second part is due to Manteuffel [273].

Theorem 7.35. *Let $E \subset I \times I$ satisfy (7.19a). (a) M-matrices A permits an ILU decomposition $A = W - R$ with W in (7.24b), which, in addition, represents a splitting (7.4) in the sense of Definition 7.9.*

(b) *If an H-matrix A has a positive diagonal D , the ILU decomposition*

$$A = (D + L')D^{-1}(D + U')$$

exists. $\hat{A} := D - |A - D|$ (cf. Definition C.60) has also an ILU decomposition $(\hat{D} + \hat{L}')\hat{D}^{-1}(\hat{D} + \hat{U}')$. Then the following inequalities hold:

$$0 \leq \hat{D} \leq D, \quad \hat{L}'\hat{D}^{-1} \leq -|L'D^{-1}| \leq 0, \quad \hat{D}^{-1}\hat{U}' \leq -|D^{-1}U'| \leq 0.$$

Meijerink–van der Vorst [280] prove part (a) by interpreting the ILU decomposition as a sequence of Gauss elimination steps which conserve the M-matrix property (cf. Lemma C.59). We give another proof directly referring to the defining equations (7.20c) and requiring weaker assumptions.

X_E denotes the restriction of a matrix X to the index subset E :

$$(X_E)_{\alpha\beta} := \begin{cases} X_{\alpha\beta} & \text{if } (\alpha, \beta) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

The matrices denoted in the following by the letters D , L , and U with different indices should always be of diagonal structure or strictly lower or upper triangular structure, respectively. Note that the triple (D, L, U) is uniquely defined by the sum $X = D + L + U$. To express the single components of this triple, we write $X = \text{diag}\{X\} + L(X) + U(X)$.

In the following, it is not necessarily assumed that $A_{\alpha\beta} \leq 0$ holds for $\alpha \neq \beta$, as it is necessary for M-matrices. We define

$$(A_-)_{\alpha\beta} := \begin{cases} A_{\alpha\beta} & \text{if } \alpha = \beta \text{ or } A_{\alpha\beta} \leq 0, \\ 0 & \text{otherwise.} \end{cases}$$

The matrix A is assumed to fulfil the following conditions:

$$A_{\alpha\beta} \leq (L(A_-)_E \cdot \text{diag}\{A\}^{-1} \cdot U(A_-)_E)_{\alpha\beta} \quad \text{for all } \begin{cases} \alpha \neq \beta, \\ (\alpha, \beta) \in E, \end{cases} \quad (7.31a)$$

A has a complete LU decomposition

$$A = (\underline{D} + \underline{L})\underline{D}^{-1}(\underline{D} + \underline{U}) = \underline{D} + \underline{L} + \underline{U} + \underline{L}\underline{D}^{-1}\underline{U} \quad (7.31b)$$

with $\underline{D} \geq 0$, $\underline{L} \leq 0$, $\underline{U} \leq 0$.

Remark 7.36. All M-matrices A satisfy the assumptions (7.31a,b). (7.31b) implies the inverse positivity of A , i.e., $A^{-1} \geq 0$. Condition (7.31a) is always satisfied if A fulfills the sign condition $A_{\alpha\beta} \leq 0$ ($\alpha \neq \beta$) for all $(\alpha, \beta) \in E$.

Proof. Since the Gauss elimination yields the complete LU decomposition, the inequalities in (7.31b) follow from Lemma C.59. Vice versa, the inequalities in (7.31b) imply $(D + L)^{-1} \geq 0$, $D^{-1} \geq 0$, $(D + U)^{-1} \geq 0$, from which $A^{-1} \geq 0$ can be concluded. If A is an M-matrix and therefore $A_{\alpha\beta} \leq 0$ for $\alpha \neq \beta$, $(A_-)_{\alpha\beta} \leq 0 \leq (L(A_-)_E \operatorname{diag}\{A\}^{-1} U(A_-)_E)_{\alpha\beta}$ follows. \square

Theorem 7.37. Assume that $E \subset I \times I$ satisfies (7.19a) and that the matrix A fulfills (7.31a,b). Then A permits an ILU decomposition $A = W - R$ with W in (7.24b). $A = W - R$ is a regular splitting if $A_{\alpha\beta} \leq 0$ for $(\alpha, \beta) \notin E$ (the minimal condition (7.19b) is sufficient). The enclosure (7.32) holds with D , L , U from (7.31b):

$$(\underline{D} + \underline{L} + \underline{U})_E \leq D + L' + U' \leq (A_-)_E. \quad (7.32)$$

Proof. The conditions (7.20d) can be written as $R_E = 0$. Inserting the remainder $R = D + L' + U' + L'D^{-1}U' - A$, we obtain $(D + L' + U' + L'D^{-1}U' - A)_E = 0$, i.e.,

$$(D + L' + U')_E = (A - L'D^{-1}U')_E. \quad (7.33)$$

Using the mapping

$$X \mapsto \Phi(X) := (A - L(X) \cdot \operatorname{diag}\{X\}^{-1} \cdot U(X))_E, \quad (7.34a)$$

we may write the defining equation (7.33) as a *fixed-point equation*

$$D + L' + U' = \Phi(D + L' + U'). \quad (7.33')$$

Assume the monotonicity properties

$$\left. \begin{array}{l} C_1 \leq C_2, \operatorname{diag}\{C_1\} \geq 0, \\ L(C_2) + U(C_2) \leq 0 \end{array} \right\} \implies \Phi(C_1) \leq \Phi(C_2). \quad (7.34b)$$

Equation (7.31b) states that $A = \underline{D} + \underline{L} + \underline{U} + \underline{L} \underline{D}^{-1} \underline{U}$. We set

$$A_0 := \underline{D} + \underline{L} + \underline{U} \quad \text{and} \quad A^0 := (A_-)_E. \quad (7.34c)$$

$\underline{L} \underline{D}^{-1} \underline{U} \geq 0$ yields

$$A_0 = (A_0)_- \leq ((A_0)_-)_E \leq ((A_0 + \underline{L} \underline{D}^{-1} \underline{U})_-)_E = (A_-)_E = A^0,$$

i.e.,

$$A_0 \leq A^0. \quad (7.34d)$$

Next, we show that

$$A_0 \leq \Phi(A_0) \quad \text{and} \quad \Phi(A^0) \leq A^0. \quad (7.34e)$$

$\Phi(A_0) = (A - \underline{L} \underline{D}^{-1} \underline{U})_E = (\underline{D} + \underline{L} + \underline{U})_E = (A_0)_E \geq A_0$ holds because of $\underline{L}, \underline{U} \leq 0$. The second inequality in (7.34e) is identical to (7.31a). Φ defines the following fixed-point iterations:

$$A_{m+1} := \Phi(A_m), \quad A^{m+1} := \Phi(A^m). \quad (7.34f)$$

The monotonicity (7.34b) and the inequalities (7.34d,e) lead to

$$A_0 \leq A_1 \leq \dots \leq A_m \leq \dots \leq A^m \leq \dots \leq A^1 \leq A^0 \quad (7.34g)$$

(cf. Theorem 7.19). Hence, both sequences must converge to a unique limit $C = D + L' + U'$ satisfying the fixed-point equation (7.33'). (7.32) follows from (7.34c) and $A_0 \leq D + L' + U' \leq A^0$. $W^{-1} = (D + U')^{-1}D(D + L')^{-1} \geq 0$ is a consequence of the inequalities $D \geq 0$ and $L', U' \leq 0$. Remainder R vanishes on E : $R_E = 0$; otherwise, $R_{\alpha\beta} = (L'D^{-1}U' - A)_{\alpha\beta}$ holds. The inequality $A_{\alpha\beta} \leq 0$ for indices $(\alpha, \beta) \notin E$ implies $R_{\alpha\beta} \geq (L'D^{-1}U')_{\alpha\beta} \geq 0$. Hence, the splitting $A = W - R$ is regular. \square

The *stability*³ of the ILU decomposition is expressed in (7.32) by the estimate of the diagonal D from below by \underline{D} .

To generalise Theorem 7.37 to the ILU_ω decomposition with $\omega \neq 0$, we may write the equations $R_{ij} = 0$ for $i \neq j$, $(i, j) \in E$, and (7.29) as

$$R_E - \omega \operatorname{diag}\{R_E \mathbf{1}\} = 0 \quad \text{with } R = D + L + U + LD^{-1}U - A,$$

Here, $E' := (I \times I) \setminus E$ is the complement. For a vector $v = (v_1, \dots, v_n)^\top$, $\operatorname{diag}\{v\}$ denotes the diagonal matrix $\operatorname{diag}\{v_1, \dots, v_n\}$. Carrying over the proof technique, we are led to the fixed-point equation $C = \Phi_\omega(C)$ with

$$\Phi_\omega(C) := \Phi(C) - \omega \operatorname{diag}\{(A - L(C) \cdot \operatorname{diag}\{C\}^{-1} \cdot U(C))_{E'} \mathbf{1}\} \quad (7.35)$$

and Φ defined in (7.34a). In general, however, Φ_ω does not have the desired properties. The monotonicity corresponding to (7.34b) may be violated for $\omega > 0$, whereas for $\omega < 0$, it may happen that no A_0 exists with $\Phi_\omega(A_0) \geq A_0$ (and hence, no solution exists).

For a precise discussion, we study the five-point formula (7.26) with the five-point pattern (7.27a). Since L', U' are already uniquely determined (cf. (7.27b,c)), the fixed-point equation simplifies to a scalar equation for D :

$$\begin{aligned} D = \Phi_\omega(D) &:= \operatorname{diag}\{A - L'D^{-1}U'\} - \omega \operatorname{diag}\{(A - L'D^{-1}U')_{E'} \mathbf{1}\} \\ &= \operatorname{diag}\{d + (\omega a - c)e/D_{i-1,j} + (\omega c - a)b/D_{i,j-1}\} \end{aligned} \quad (7.36a)$$

³ Concerning the problem that the solution of the systems $(D + L)x = b$ or $(D + U)x = b$ may lead to instabilities, we refer to Elman [121], where ILU decompositions for nonsymmetric matrices are discussed.

(cf. (7.30)). For analysing this equation, we investigate the one-dimensional fixed-point equation

$$d = \varphi_\omega(d) := d + [(\omega e - b)a + (vb - e)c] / d. \quad (7.36b)$$

A discussion of the function φ_ω , which is left to the reader, shows the following.

(i) The fixed-point equation (7.36b) is solvable if and only if

$$4\gamma < d^2 \quad \text{for } \gamma := ce + ab - \omega(ae + cb). \quad (7.36c)$$

(ii) If (7.36c) is satisfied, the solutions of (7.36b) are

$$\delta_{\pm} = \frac{1}{2} \left(d \pm \sqrt{d^2 - 4\gamma} \right). \quad (7.36d)$$

(iii) δ_+ is the stable fixed point because (7.36e) leads to (7.36f):

$$\varphi_\omega(\delta) < \delta \quad \text{for } \delta > \delta_+, \quad \varphi_\omega(\delta) > \delta \quad \text{for } \delta_- < \delta < \delta_+, \quad (7.36e)$$

$$\lim \delta_m = \delta_+ \quad \text{for } \delta_0 > \delta_-, \quad \delta_{m+1} := \varphi_\omega(\delta_m). \quad (7.36f)$$

(iv) On the other hand, starting values $\delta_0 < \delta_-$ generate sequences $\{\delta_m\}$ which contain at least one element $\delta_m \leq 0$.

Exercise 7.38. Let A in (7.26) be diagonally dominant and symmetric:

$$a = b \geq 0, \quad c = e \geq 0, \quad s := a + c > 0, \quad d = 2\sigma + \varepsilon \quad \text{with } \varepsilon \geq 0. \quad (7.37a)$$

Prove that for $\omega = -1$, the value δ_+ is obtained from (7.36d) with $\gamma = \sigma^2$. For small ε , this value has the expansion

$$\delta_+ = \sigma + \sqrt{\varepsilon\sigma} + \mathcal{O}(\varepsilon). \quad (7.37b)$$

Assuming (7.37a), we obtain for $\omega = 0$ that

$$\delta_+ = a + c + \sqrt{2ac} + \mathcal{O}(\sqrt{\varepsilon}).$$

Theorem 7.39. Let $\omega \in [-1, \omega^*]$, where $\omega^* := \min\{\frac{c}{a}, \frac{a}{c}\}$. Assume that the matrix A in (7.26) satisfies (7.37a). Then the ILU_ω decomposition exists, and the entries d_{ij} of the diagonal D are enclosed by

$$\delta_+ = \frac{d + \sqrt{d^2 - 4(c^2 + a^2 - 2\omega ac)}}{2} < d_{ij} \leq d \quad \text{for } (i, j) \in I. \quad (7.38)$$

The fixed-point iteration (7.36a) with the starting iterate $D^0 := \text{diag}\{d\mathbf{1}\}$ converges from above to D .

Proof. (7.36c) is satisfied for $\omega > -1$, while Φ_ω is monotone for $\omega \leq \omega^*$. One verifies that $D_0 \leq \Phi_\omega(D_0)$ and $\Phi_\omega(D^0) \leq D^0$ hold for $D_0 := \text{diag}\{\delta_+\mathbf{1}\}$ and $D^0 := \text{diag}\{d\mathbf{1}\}$. Hence, we can draw the same conclusions as in the proof of Theorem 7.37. \square

7.3.6 Properties of the ILU Decomposition

An immediate consequence of Theorem 7.11 is the following convergence statement.

Theorem 7.40. *If A is an M-matrix or, if according to Theorem 7.37, $A = W - R$ describes a regular splitting, the ILU iteration (7.24a,b) converges with the convergence rate $\rho(A^{-1}R)/(1 + \rho(A^{-1}R))$.*

In the standard case, one may assume $\|R\| = \mathcal{O}(\|A\|)$, so that $\rho(A^{-1}R) \leq \|A^{-1}\| \|R\| \leq C\|A^{-1}\| \|A\| = C \text{cond}(A) \gg 1$ leads to the convergence rate $(1 + 1/\rho(A^{-1}R))^{-1} \approx 1 - \mathcal{O}(1/\text{cond}(A))$. Hence, the ILU decomposition has the same order as the Jacobi or Gauss–Seidel iteration. A better result can be derived for the modified ILU₋₁ decomposition (cf. (7.28) or (7.29) with $\omega = -1$). We prepare its analysis with the following lemma (cf. Wittum [402]).

Lemma 7.41. *Assume (7.39a), where A , D_A , and D are positive definite:*

$$A = D_A - L - L^H, \quad W = (D + L')D^{-1}(D + L'^H). \quad (7.39a)$$

The spectrum $\sigma(W^{-1}A)$ is contained in $[0, \Gamma]$ if

$$(2 - \frac{1}{\Gamma})D - D_A + L + L^H + L' + L'^H \quad \text{is positive semidefinite.} \quad (7.39b)$$

Proof. We write $D + L'$ as $\frac{1}{\Gamma}D + C$ with $C := (1 - \frac{1}{\Gamma})D + L'$. From

$$\begin{aligned} \Gamma W - A &= (\frac{1}{\Gamma}D + C)(\frac{1}{\Gamma}D)^{-1}(\frac{1}{\Gamma}D + C)^H - A \geq \frac{1}{\Gamma}D + C + C^H - A \\ &= (2 - \frac{1}{\Gamma})D - D_A + L + L^H + L' + L'^H \geq 0 \end{aligned}$$

with ‘ \geq ’ in the sense of semidefiniteness, it follows that $\sigma(W^{-1}A) \in [0, \Gamma]$. \square

Theorem 7.42. *Let $-1 \leq \omega \leq \omega^*$ (cf. Theorem 7.39). The five-point formula (7.26) and the five-point pattern (7.27a) are assumed to satisfy (7.37a). Then the inequality*

$$\gamma W \leq A \leq \Gamma W \quad \text{with } \begin{cases} \gamma = 1/[1 + (1 + \omega)\frac{2ac}{\delta_+ \lambda_{\min}}], \\ \Gamma = \delta_+/[2\delta_+ - d], \end{cases} \quad (7.40)$$

holds with ‘ \leq ’ in the sense of semidefiniteness, where δ_+ is defined in (7.38) and $\lambda_{\min} = \varepsilon + 4(a + c) \sin^2 \frac{\pi h}{2}$ is the smallest eigenvalue of A . In particular, (7.41) holds:

$$\gamma = 1, \quad \Gamma = \frac{1}{2}\sqrt{\frac{\sigma}{\varepsilon}} - \frac{1}{4} + \mathcal{O}\left(\sqrt{\frac{\varepsilon}{\sigma}}\right) \quad \text{for } \omega = -1. \quad (7.41)$$

Proof. (i) (7.39b) becomes $(2 - \frac{1}{\Gamma})D - D_A \geq 0$, since by (7.27b,c), $L = -L'$ holds in Lemma 7.41. Thanks to $D_A = dI$ and $\delta_+ I \leq D$ (cf. (7.38)), Γ with $(2 - \frac{1}{\Gamma})\delta_+ = d$ is sufficient for (7.39b). Solving for Γ , we obtain $\Gamma = \delta_+/(2\delta_+ - d)$.

(ii) The entries r_{ij} , s_{ij} of R (cf. (7.27f)) are bounded from above by $2ac/\delta_+$. According to (7.29), the diagonal entries of R are equal to $\omega(r_{ij} + s_{ij})$. The eigenvalues of R lie in the Gershgorin circles around $\omega(r_{ij} + s_{ij})$ with the radius $r_{ij} + s_{ij}$ (cf. Hackbusch [193, Criterion 4.3.4], Varga [376]) and, hence, they are bounded by $(1+\omega)(r_{ij} + s_{ij}) \leq 2(1+\omega)ac/\delta_+$, implying $R \leq [2(1+\omega)ac/\delta_+]I$. From $\lambda_{\min}I \leq A$, we deduce $R \leq \rho A$ with $\rho := 2(1+\omega)ac/(\delta_+ + \lambda_{\min})$. $A = W - R \geq W - \rho A$ yields $W \leq (1+\rho)A$. Hence, $\gamma = 1/(1+\rho)$ leads to the representation of γ .

(iii) For $\omega = -1$, insert the representation (7.37b) into (7.40). \square

Conclusion 7.43. (a) Replacing the Poisson equation $-\Delta u = f$ with the Helmholtz equation $-\Delta u + \varepsilon u = f$ with $\varepsilon > 0$, we obtain the coefficients $a = b = c = e = -h^{-2}$, $d = 4h^{-2} + \varepsilon$ in (7.37a). Equation (7.41) yields the bound and condition number $\Gamma = \Gamma/\gamma = h^{-1}/\sqrt{2\varepsilon} + \mathcal{O}(1)$ indicating the order improvement.

(b) Let $\omega = -1$. The (modified) ILU₋₁ iteration damped by $\vartheta_{\text{opt}} = 2/(\gamma + \Gamma) = 2\sqrt{2\varepsilon}h + \mathcal{O}(h^2)$ has the convergence speed

$$\rho(M_{\vartheta_{\text{opt}}}^{\text{ILU}}) \leq (\Gamma - 1)/(\Gamma + 1) \approx 1 - 2/\Gamma \approx 1 - 2\sqrt{2\varepsilon}h.$$

Hence, similar to the SSOR method with an optimal relaxation parameter ω_{SSOR} , it is of first order as long as $\varepsilon > 0$.

Proof. Use Theorem 6.7. \square

The applicability of the ILU₋₁ decomposition is not at all restricted to strict diagonal dominance in Theorem 7.42 and Remark 7.43b, as shown by the following remark.

Remark 7.44 (enlargement of the diagonal). Let $A = A_\varepsilon$ be a matrix satisfying (7.37a) with $\varepsilon > -4(a+c)\sin^2\frac{\pi h}{2}$ (i.e., $\lambda_{\min}(A) > 0$) instead of $\varepsilon > 0$. Then the ILU₋₁ decomposition $A_\eta = W_\eta - R_\eta$ has to be applied to the matrix $A_\eta := A + (\eta - \varepsilon)I$ with $\eta > 0$ in order to re-establish diagonal dominance $d > 2\sigma$. W_η can be viewed as the ILU decomposition of $A = A_\varepsilon$ with remainder $R = W_\eta - A = R_\eta - (\eta - \varepsilon)I$. Conclusion 7.43 yields the spectral condition number $\kappa(W_\eta^{-1}A_\eta)$. Let $\lambda = \lambda_{\min}(A)$ and $\Lambda = \lambda_{\max}(A)$ be the extreme eigenvalues of A . Because of

$$\kappa(A_\eta^{-1}A) = \kappa(A_\eta^{-1}A_\varepsilon) = \frac{\Lambda(\lambda + \eta - \varepsilon)}{\lambda(\Lambda + \eta - \varepsilon)} \approx 1 + \frac{\eta - \varepsilon}{\lambda_{\min}(A)},$$

Lemma 7.55 shows that

$$\kappa(W_\eta^{-1}A_\varepsilon) \lesssim h^{-1} \left(1 + \frac{\eta - \varepsilon}{\lambda_{\min}(A)} \right) / \sqrt{2\eta}. \quad (7.42)$$

Exercise 7.45. Prove that the right-hand side of (7.42) becomes minimal for $\eta = 4(a+c)\sin^2\frac{\pi h}{2}$.

Exercise 7.46. Prove that the ILU decomposition coincides with the exact LU decomposition if A has the tridiagonal pattern $\begin{bmatrix} * & * & * \\ \cdot & * & * \\ * & \cdot & * \end{bmatrix}$ or $\begin{bmatrix} * & * & * \\ * & * & * \\ \cdot & * & * \end{bmatrix}$. Then the ILU iteration solves $Ax = b$ directly.

7.3.7 ILU Decompositions Corresponding to Other Patterns

Strengthening (7.19b) by $E \supseteq G(A)$ is the minimal requirement to construct new methods. When choosing a pattern E larger than $G(A)$, we should add those positions where $R = 0$ is violated: According to (7.27f), these are the positions $\begin{bmatrix} * & * & * \\ \cdot & * & * \\ * & * & * \end{bmatrix}$. Adding $\begin{bmatrix} * & * & * \\ \cdot & * & * \\ * & * & * \end{bmatrix}$ to the five-point pattern, we obtain

$$E = \begin{bmatrix} * & * \\ * & * & * \\ * & * \end{bmatrix} \quad (\text{'seven-point pattern'}). \quad (7.43)$$

Now the lower triangular matrix L' and upper triangular matrix U' have the form

$$L' = - \begin{bmatrix} 0 & 0 \\ a_{ij} & 0 & 0 \\ c & f_{ij} \end{bmatrix}, \quad U' = - \begin{bmatrix} g_{ij} & e \\ 0 & 0 & b_{ij} \\ 0 & 0 \end{bmatrix},$$

whose coefficients result from the recursions

$$\begin{aligned} d_{ij} &= d - ec/d_{i,j-1} + a_{ij}(\omega g_{i-1,j} - b_{i-1,j})/d_{i-1,j} \\ &\quad + f_{ij}(\omega b_{i+1,j-1} - g_{i+1,j-1})/d_{i+1,j-1}, \\ a_{ij} &= a + g_{i,j-1}c/d_{i,j-1}, \quad b_{ij} = b + ef_{ij}/d_{i+1,j-1}, \\ f_{ij} &= b_{i,j-1}c/d_{i,j-1}, \quad g_{ij} = a_{ij}e/d_{i-1,j} \end{aligned} \quad (7.44)$$

for $1 \leq i, j \leq N - 1$, where all terms with indices $i - 1 = 0$, $j - 1 = 0$, or $i + 1 = N$ have to be ignored. This seven-point ILU decomposition has properties similar to those of the five-point version in Theorem 7.42 (cf. Gustafsson [172], Axelsson–Barker [13, §7]).

Exercise 7.47. Prove: (a) For $-1 \leq \omega \leq 0$, the fixed-point iteration (7.35) converges for the starting iterate $C = A$ to values satisfying the inequalities $a_{ij} \leq \alpha := a/\Delta$, $b_{ij} \leq \beta := b/\Delta$, $f_{ij} \leq \beta c/\gamma$, $g_{ij} \leq \alpha e/\delta$, $d_{ij} \geq \delta$ with $\Delta := 1 - ec/\delta^2$, where d is the maximal solution of the fixed-point equation

$$\delta = \varphi(\delta) := d - \left[ec + \frac{ab}{\Delta^2} \left(1 + \frac{ec}{\delta^2} \right) - \frac{\omega}{\delta} (\alpha^2 e + \beta^2 c) \right] / \delta.$$

(b) For the next considerations, assume the symmetry $a = b$, $c = e$ as well as the diagonal dominance $d = 2(a + c) + \varepsilon$ with $\varepsilon \geq 0$. Furthermore, choose $\omega = -1$

(i.e., the modified ILU). Prove that the equation $\delta = \varphi(\delta)$ can be brought into the form $2a + \varepsilon = a(\xi + \xi^{-1})$ with $\xi := a\delta/(\delta - c)^2$. Hence, the solution is

$$\delta = c + a/(2\xi) + \sqrt{ac/\xi + a^2/(4\xi^2)}$$

with $\xi = 1 + \varepsilon/(2a) + \sqrt{\varepsilon/a + \varepsilon^2/(4a^2)}$.

(c) For $\varepsilon \geq 0$, a solution $\delta = \delta_0 + C\sqrt{\varepsilon} + \mathcal{O}(\varepsilon)$ exists.

(d) δ solves the equation $(\delta - \gamma - e - \beta)^2 = \varepsilon\delta$.

(e) The weak diagonal dominance, which is sufficient for (7.39b), leads to the condition $2\varphi + 2|a - \alpha| \leq (2 - \frac{1}{F})\delta - d$. Show that $\Gamma = \delta/(2\sqrt{\varepsilon\delta} - \varepsilon)$.

(f) As in (7.41), the estimate $\gamma W \leq A \leq \Gamma W$ holds with $\gamma = 1$.

Concerning ILU decompositions with a general k -point pattern, note that the amount of computational work increases more than linearly with the number k of pattern entries.

7.3.8 Approximative ILU Decompositions

The ILU decompositions, as defined in (7.27d) or (7.30), are strictly sequential algorithms. The same statement holds for solving the systems $(D + L)x = b$ and $(D + U)x = b$ arising during the solution of $W\delta = d$. This is a disadvantage for a parallel treatment. The parallel treatment of the systems is discussed by van der Vorst [371] (cf. also Ortega [298, §3.4]). Here we discuss the computation of the ILU decomposition. Note that the fixed-point iteration (7.34f) in the proof of Theorem 7.37 is suited to numerical computations. The upper starting iterate $A^0 = (A_-)_E$ (in general, $A^0 = A$) is available (in contrast to A_0), so that the iterates $A^{m+1} = \Phi(A^m)$ are computable.

Remark 7.48. The evaluation of the function Φ in (7.34a) can be performed in parallel for all coefficients $\Phi(X)_{\alpha\beta}$, $(\alpha, \beta) \in E$.

The equations (7.33'): $X = \Phi(X)$ or, more precisely, the recursions (7.30) and (7.44) represent simple systems of equations for the unknowns d_{ij} (and possibly a_{ij} , b_{ij} , f_{ij} , g_{ij}), which can be solved by backward substitutions. Independently of the starting iterate, the values for (i, j) with $\max\{i, j\} \leq m$ are exact after m iteration steps. If A and therefore also the starting iterate A^0 (cf. (7.34c)) have constant coefficients, the m -th iterate A^m has identical constant coefficients for all positions⁴ (i, j) with $\min\{i, j\} \geq m$. Since the coefficients of A^m coincide for $\min\{i, j\} \geq m$, one need not calculate all of them. This consideration leads us to the truncated ILU version introduced by Wittum [400] for constant coefficients:

⁴ At positions with $\min\{i, j\} < m$ other values are possible, since in (7.30) or (7.44) some terms may be absent because of $i - 1 = 0$ or $j - 1 = 0$.

Compute d_{ij} (and possibly $a_{ij}, b_{ij}, f_{ij}, g_{ij}$) from (7.30) or, respectively, (7.44) for all i, j with $\max\{i, j\} = k$ for $k = 1, 2, \dots, m$ and continue these values constantly by means of $d_{ij} := d_{\min\{i,m\}, \min\{j,m\}}$ for $\max\{i, j\} > m$

(analogously for $a_{ij}, b_{ij}, f_{ij}, g_{ij}$). The amount of computational work is $\mathcal{O}(m^2)$ independent of dimension n of the matrix. The same statement holds for the storage requirement. The truncated ILU decomposition is a good substitute for the standard ILU decomposition and has favourable stability properties (cf. Wittum–Liebau [406]).

7.3.9 Blockwise ILU Decomposition

Choosing the row or column variables as blocks, A has a block structure with tridiagonal matrix blocks in diagonal position as shown in (3.17). In the decomposition ansatz (7.22):

$$A = (D + L')D^{-1}(D + U') - R = D + L' + U' + L'D^{-1}U' - R,$$

we may also require that D be a block-diagonal matrix with blocks of tridiagonal structure and that L' and U' be strictly (lower/upper) block-triangular matrices. The algorithm is similar to (7.23a,b) (cf. (11.95a–c)). With the increased amount of computational work, one gains, in general, more robust convergence properties. Block-ILU decompositions were introduced in the early 1980s (cf. §7.3.11).

7.3.10 Numerical Examples

Table 7.1 shows the errors $\|x^m - x\|_2$ after $m = 20$ iterations and the convergence factors for different ILU variants. *ILU_5* refers to the five-point ILU defined by (7.27a), while *ILU_7* refers to (7.43). The step size of the Poisson model problem is $h = 1/32$. For $\omega = 0$ and $\omega = 1$, the ILU iteration is applied to the original matrix, whereas for the modified method with $\omega = -1$ an enlargement of the diagonal by $A_\eta := A + 5I$ is chosen according to Remark 7.44. ϑ is the damping factor in (5.8).

version	ω	ϑ	$\ x^{20} - x\ _2$	$\frac{\ x^{20} - x\ _2}{\ x^{19} - x\ _2}$
ILU_5	0	1.66	1.617_{10-1}	0.9455
ILU_5	-1	0.25	1.628_{10-3}	0.7666
ILU_5	1	1.9	2.349_{10-1}	0.9617
ILU_7	0	1	8.904_{10-2}	0.9185
ILU_7	0	1.66	2.690_{10-2}	0.8646
ILU_7	-1	0.4	4.722_{10-5}	0.6254

Table 7.1 Results of the ILU iteration for the Poisson model case.

Exercise 7.49. Count the arithmetic operations (separately for the decompositions and the solution phase) and compare *ILU_5* and *ILU_7* with regard to the effective amount of work.

7.3.11 Remarks

ILU decompositions are first mentioned in 1960 by Varga [374, §6] and Buleev [84]. The first precise analysis is due to Meijerink–van der Vorst [280]. Here, we also mention Jennings–Malik [228]. ILU methods have proved to be very robust. This means that good convergence properties are not restricted to the Poisson model problem, but hold for a large class of problems. Since the existence of an ILU decomposition is not always ensured, there are many stabilising variants. Concerning literature about the ILU method, we refer to Axelsson–Barker [13], Axelsson [12, §7], and Beauwens [37].

Because of the improved condition number Γ/γ in (7.41), the modified version ($\omega = -1$) of Gustafsson [172] is the preferred basis for applications of the conjugate gradient technique (cf. §10) to ILU iterations. Because of the consistency condition $R\mathbf{1} = \mathbf{0}$, this version is also called an ILU iteration of first order. A special decomposition for the Poisson model problem of second order is described by Stone [356]; however, because of other disadvantages, first-order variants are preferred.

The first publication of a blockwise ILU method in 1981 is due to Kettler [235], who refers to a ‘publication in preparation’ by Meijerink which appeared in [279] two years later. Additional early papers are those by Axelsson–Brinkkemper–Il’in [14] (1984 with a preprint in 1983) and Concus–Golub–Meurant [99] (1985 with a preprint in 1982).

In the literature, the distinction between SSOR and ILU methods is not very sharp. The SSOR method for $A = D + L' + U'$ corresponds to an ILU decomposition $W = (D + L')D^{-1}(D + U')$ with remainder $R = W - A = L'D^{-1}U'$. This R does not satisfy condition (7.20d); however, this condition is already weakened by (7.29) and addition of a diagonal part (cf. Remark 7.44). Vice versa, generalised SSOR methods have been introduced in which $D = \text{diag}\{A\}$ is replaced by another diagonal (cf. Axelsson–Barker [13]). The ILU iteration based on a five-point pattern also falls into this category.

In the literature, one finds a lot of abbreviations for different ILU variants. ‘IC’ refers to the ‘incomplete Cholesky’ variant of the ILU decomposition. Additional numbers like ‘(5)’ or ‘(7)’ denote the respective five- or seven-point pattern. In other papers, ‘(0)’ indicates the pattern $E = G(A)$, whereas ‘(1)’ means the pattern which is enlarged by one level, etc. The supplement ‘Tr’ characterises the truncated version (7.45). The letter ‘M’ stands for the modified method with $\omega = -1$, whereas ‘B’ may indicate a block variant. If the block corresponds to a grid line (row or column), sometimes the symbol ‘L’ is used.

In particular concerning the $\text{ILU}(p)$ variant, we refer to Saad [328, §§10.3]. The thresholding technique ILUT can also be found in [328, §§10.4]. See also Björck [48, §§4.4.3f]. Another kind of factorisation is proposed by Benzi–Tüma [42].

While ILU methods are less attractive as linear iterations, their combination with multigrid methods is successful (see §11.6.2 and Hackbusch–Wittum [208]).

7.4 Preconditioning

The term ‘preconditioning’ is rather ambiguous. In §7.4.1 we describe the preconditioning in the narrower sense. When it is used in the wider sense it is losing its original meaning and, in the extreme case, may mean any transformation in the sense of §5.6 (cf. §7.4.3).

7.4.1 Idea of Preconditioning

We recall the *spectral condition number* $\kappa(A) := \rho(A)\rho(A^{-1})$ of a regular⁵ matrix defined in (B.13). In the case of $A > 0$, the spectral condition number $\kappa(A)$ simplifies to the ratio $\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$ of the extreme eigenvalues. Alternatively, for a given matrix norm we can define the condition $\text{cond}(A) := \|A\| \|A^{-1}\|$. If A is normal, the Euclidean condition $\text{cond}_2(A)$ with respect to the spectral norm coincides with $\kappa(A)$. This holds in particular under the assumption $A > 0$. Furthermore, we consider the simplest linear iteration: the Richardson iteration defined in §3.2.1. The convergence analysis in §3.5.1 shows that, for the optimal parameter Θ_{opt} , the convergence rate and contraction number coincide with

$$\rho(M_{\Theta_{\text{opt}}}^{\text{Rich}}) = \frac{\kappa(A) - 1}{\kappa(A) + 1} = \frac{1 - \frac{1}{\kappa(A)}}{1 + \frac{1}{\kappa(A)}}$$

(cf. (3.26c)). The essential observation is that $\rho(M_{\Theta_{\text{opt}}}^{\text{Rich}})$ depends only on the spectral number $\kappa(A)$ (cf. (B.13)).

If $\kappa(A)$ is very close to 1, we have very fast convergence. If $\kappa(A)$ is of moderate size, a moderate convergence speed results. If, however, $\kappa(A)$ is large, the asymptotic approximation $\rho(M_{\Theta_{\text{opt}}}^{\text{Rich}}) = 1 - 2/\kappa(A) + \mathcal{O}(\kappa(A)^{-2})$ shows that the convergence is rather slow.

Hence, one can try to choose a left transformation with $T_\ell = N = W^{-1}$ so that

$$\hat{A} := T_\ell A = W^{-1}A \quad \text{has a positive spectrum,} \quad (7.46a)$$

$$\kappa(W^{-1}A) \quad \text{is as small as possible.} \quad (7.46b)$$

Note that under condition (7.46a) Theorem 6.7 implies that the optimally damped iteration

$$\varPhi_W(x, b) := x - \vartheta_{\text{opt}} W^{-1}(Ax - b) \quad (7.46c)$$

has the convergence rate

$$\rho(M_{\vartheta_{\text{opt}}}^{\text{Rich}}) = \frac{\kappa(W^{-1}A) - 1}{\kappa(W^{-1}A) + 1}.$$

Often, the matrix W is called the *preconditioning matrix*, *preconditioning*, or *preconditioner*. Sometimes these names also refer to the matrix $N = W^{-1}$.

⁵ We may set $\kappa(A) = \infty$ for singular A . For certain purposes it makes sense to extend the spectral condition to singular matrices $A \neq 0$ by $\kappa_0(A) := \max_{\lambda \in \sigma(A)} |\lambda| / \min_{\lambda \in \sigma(A) \setminus \{0\}} |\lambda|$.

of the second normal form. The mapping

$$\varPhi_{\Theta}^{\text{Rich}} \mapsto \varPhi_W = \varPhi_{\Theta}^{\text{Rich}} \circ W^{-1}$$

is also called ‘preconditioning’. Note that this term expresses the intention to improve the condition, but it is not a concrete description of the mapping $A \mapsto W[A]$. Besides the size of the condition (and therefore the convergence speed) one must have in mind the related cost (cf. §2.3.2).

The condition numbers will also appear in Part II in connection with the semi-iterative method applied to the basic iteration \varPhi_W . Instead of a real spectrum contained in $[\lambda_{\min}(A), \lambda_{\max}(A)]$, we may replace the interval by an ellipse (cf. §8.3.6).

The construction of the iteration (7.46c) is not restricted to the left transformation $\varPhi_W = \varPhi^{\text{Rich}} \circ W^{-1}$. The right transformation $T_r = W^{-1}$ applied to the Richardson method leads to the same iteration (5.41): $\varPhi_W(x, b) = x - W^{-1}(Ax - b)$. The two-sided transformation $\varPhi_W = W^{-1/2} \circ \varPhi^{\text{Rich}} \circ W^{-1/2}$ by (5.46) also leads to the same ‘preconditioned’ iteration.

7.4.2 Examples

As examples of preconditioning the positive definite matrix $A = D - E - F$ (cf. (1.16)) we recall the matrices W of the already described symmetric iterations:

$$\begin{aligned} W &= D = \text{diag}\{A\} && (\text{Jacobi}), \\ W &= (D - E)D^{-1}(D - F) && (\text{SSOR}). \end{aligned}$$

Here, the methods can be understood pointwise or blockwise.

Since the choice of ‘ $W = \text{diagonal matrix}$ ’ is especially simple and also computable in parallel, one might ask whether the Jacobi method with $D := \text{diag}\{A\}$ represents the optimal diagonal preconditioning. The answer is given by Theorems 7.50 and 7.51: $D := \text{diag}\{A\}$ is optimal in the 2-cyclic case, whereas D is close to the optimum in the general case (see also Higham [221, Theorem 7.5]).

Theorem 7.50 (Forsythe–Strauss [139]). *Assume that A is positive definite with $D := \text{diag}\{A\}$ and $A - D$ is weakly 2-cyclic. Then $D := \text{diag}\{A\}$ is the best diagonal preconditioner; i.e., $\kappa(D^{-1}A) \leq \kappa(\Delta^{-1}A)$ for all diagonal matrices Δ .*

Theorem 7.51 (van der Sluis [368]). *Let the matrix A be positive definite with $D := \text{diag}\{A\}$ and assume that each row of A contains at most C_A nonzero entries (cf. (2.28)). Then $\kappa(D^{-1}A) \leq C_A \kappa(\Delta^{-1}A)$ holds for all diagonal matrices Δ .*

Bank–Scott [32] describe a related result about the condition of finite element matrices in the presence of local refinements.

Let Γ be the constant in (8.39c). The SSOR preconditioning improves the condition number from $\kappa(A)$ to $\frac{1}{2}(1 + \sqrt{\Gamma\kappa(A)})$. The transition from $\kappa(A)$ to $\mathcal{O}(\sqrt{\kappa(A)})$ corresponds to the improvement of the order (cf. Conclusion 6.29).

7.4.3 Preconditioning in the Wider Sense

Let $A = Q \operatorname{diag}\{\lambda_i : 1 \leq i \leq n\} Q^H$ (Q is unitary) be any normal matrix with $0 < \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$. Obviously $\kappa(A) = \operatorname{cond}_2(A) = \lambda_n/\lambda_1$ is the condition. Now we replace λ_1 with $-\lambda_1$. $\hat{A} = Q \operatorname{diag}\{-\lambda_1, \lambda_2, \dots, \lambda_n\} Q^H$ is an indefinite matrix also satisfying $\kappa(\hat{A}) = \operatorname{cond}_2(\hat{A}) = \kappa(A)$. Although the condition is unchanged, the Richardson iteration has a problem because of Exercise 3.25. Obviously, it is not the condition which must be improved, but the indefinite matrix must be turned into a positive definite one. Again a transformation by $W^{-1} = \hat{A}$ helps: the resulting squared Richardson iteration \hat{A}^2 is positive definite. However, the condition $\kappa(A)$ is replaced with the larger condition $\kappa(\hat{A}^2) = \kappa(A)^2$. Calling $W^{-1} = \hat{A}$ a preconditioner, the original meaning of improving the condition is perverted. Nevertheless, we can try to precondition \hat{A}^2 in the narrower sense. One learns from this example that beside the condition other structural properties are important which may be improved by a transformation for which the name 'preconditioning' is not quite adequate.

Another systematic approach to indefinite Hermitian matrices A (cf. Remark 8.31) is the left transformation by a polynomial in A . Such 'preconditioners' are described, e.g., by Ashby–Manteuffel–Saylor [7]. Here the polynomial $T_\ell = p(A)$ should be close to the minimiser of $\min\{\rho(p(A)) : \operatorname{degree}(p) = d\}$ for a fixed degree $d \geq 1$.

In the case of non-Hermitian matrices A , even the convergence of the Richardson iteration cannot be described by $\kappa(A)$ or $\operatorname{cond}_2(A)$. Hence the term 'preconditioning' loses its meaning. On the other hand, a large condition number is not necessarily a disadvantage (see the multigrid iteration in §11.4). For the extreme example of a diagonal matrix A , the system is exactly solvable independently of the condition.

7.4.4 Rules for Condition Numbers and Spectral Equivalence

The Euclidean condition $\operatorname{cond}_2(\cdot)$ and the spectral condition number $\kappa(\cdot)$ satisfy the following equations and inequalities (cf. (B.12), (B.13)).

Exercise 7.52. Let the matrices A, B, C be regular. Prove the following:

$$\kappa(A) = \kappa(A^{-1}), \quad \operatorname{cond}_2(A) = \operatorname{cond}_2(A^{-1}), \quad (7.47a)$$

$$\kappa(A) = \kappa(\lambda A), \quad \operatorname{cond}_2(A) = \operatorname{cond}_2(\lambda A) \text{ for } \lambda \in \mathbb{C} \setminus \{0\}, \quad (7.47b)$$

$$\kappa(A) = \operatorname{cond}_2(A) \quad \text{for normal matrices } A, \quad (7.47c)$$

$$\operatorname{cond}_2(AB) \leq \operatorname{cond}_2(A) \operatorname{cond}_2(B), \quad (7.47d)$$

$$\operatorname{cond}_2(C^{-1}A) \leq \operatorname{cond}_2(C^{-1}B) \operatorname{cond}_2(B^{-1}A), \quad (7.47e)$$

$$\kappa(B^{-1}A) = \operatorname{cond}_2(B^{-1/2}AB^{-1/2}) \quad \text{for } A, B > 0, \quad (7.47f)$$

$$\kappa(AB) = \kappa(BA). \quad (7.47g)$$

Following considerations are restricted to positive definite matrices. The next lemma shows that the spectral number can be formulated by matrix inequalities.

Lemma 7.53. *Let A and B be positive definite. Then $\kappa(B^{-1}A)$ can be represented as*

$$\kappa(B^{-1}A) = \bar{\alpha}/\underline{\alpha} \quad (7.48a)$$

where $\bar{\alpha}$ and $\underline{\alpha}$ are the best bounds in the inequality

$$\underline{\alpha}B \leq A \leq \bar{\alpha}B \quad \text{with } \underline{\alpha} > 0. \quad (7.48b)$$

Vice versa, (7.48b) implies

$$\kappa(B^{-1}A) \leq \bar{\alpha}/\underline{\alpha}. \quad (7.48c)$$

Proof. The best bounds in (7.48b) are the extreme eigenvalues of $B^{-1}A$. Hence, (7.48a) follows from (B.14). \square

Exercise 7.54. Prove that (7.48b) is equivalent to either of the following inequalities:

$$\frac{1}{\underline{\alpha}}A \leq B \leq \frac{1}{\bar{\alpha}}A \quad \text{with } \underline{\alpha} > 0, \quad (7.48d)$$

$$\underline{\alpha}A^{-1} \leq B^{-1} \leq \bar{\alpha}A^{-1} \quad \text{with } \underline{\alpha} > 0, \quad (7.48e)$$

$$\underline{\alpha}\langle Bx, x \rangle \leq \langle Ax, x \rangle \leq \bar{\alpha}\langle Bx, x \rangle \quad \text{for all } x \in \mathbb{K}^I. \quad (7.48f)$$

The inequalities (7.47e,f) yield the next lemma.

Lemma 7.55. *Let A , B , C be positive definite. Then*

$$\kappa(C^{-1}A) \leq \kappa(C^{-1}B)\kappa(B^{-1}A). \quad (7.49)$$

Interpreting (7.47e) and (7.49) in the sense of preconditioning yields the following statement. If B is a good preconditioner for A and C is a good preconditioner for B , then C also represents a good preconditioning for A .

The following definition of spectral equivalence does not make sense for a single matrix. Instead we need two infinite families

$$\mathcal{A} = (A_\nu)_{\nu \in F}, \quad \mathcal{B} = (B_\nu)_{\nu \in F} \quad (\#F = \infty)$$

of matrices (cf. §1.4). Usually, $\nu \in F = \mathbb{N}$ is related to a discretisation grid size h_ν with the property $h_\nu \rightarrow 0$. In this case, we prefer the notation $\mathcal{A} = (A_h)_{h \in H}$. Then the size of the matrices is increasing with $\nu \rightarrow \infty$. Another case may be a matrix depending on a parameter ν varying in an interval F .

Definition 7.56 (spectral equivalence). Let $\mathcal{A} = (A_\nu)_{\nu \in F}$ and $\mathcal{B} = (B_\nu)_{\nu \in F}$ be two families of positive semidefinite matrices. Then \mathcal{A} and \mathcal{B} are called *spectrally equivalent* if there is a constant $c > 0$ so that

$$\frac{1}{c}A_\nu \leq B_\nu \leq cA_\nu \quad \text{for all } \nu \in F. \quad (7.50)$$

The explicit notation of the equivalence relation is

$$\mathcal{A} \sim \mathcal{B}.$$

Often, the less precise notation $A_\nu \sim B_\nu$ is used.

The characteristic properties of an equivalence relation are obviously satisfied: the symmetry $\mathcal{A} \sim \mathcal{B} \Leftrightarrow \mathcal{B} \sim \mathcal{A}$ and the transitivity $\mathcal{A} \sim \mathcal{B} \sim \mathcal{C} \Rightarrow \mathcal{A} \sim \mathcal{C}$. Gunn [171] used similar arguments in 1964 without mentioning the term *spectral equivalence*. This term is introduced by D'Yakonov [110] in 1966.

A more general definition of an equivalence relation can be based on $\text{cond}(\cdot)$.⁶

Remark 7.57. (a) Assume that $A_\nu \geq 0$ but not $A_\nu > 0$. Then $A_\nu \sim B_\nu$ implies that B_ν is also semidefinite and that both matrices have coinciding kernels.

(b) If $A_\nu > 0$ and $B_\nu > 0$, then (7.50) is equivalent to $\sup_{\nu \in F} \kappa(A_\nu^{-1} B_\nu) < \infty$.

Proof. Rewriting (7.50) using (7.48f), part (a) is obvious. For part (b), use Lemma 7.53. \square

Proposition 7.58. Let the matrices $A_\nu, B_\nu, C_\nu, D_\nu$ be positive semidefinite. The spectral equivalence relation satisfies the following rules:

$$A_\nu \sim B_\nu \text{ and } \lambda \geq 0 \Rightarrow \lambda A_\nu \sim \lambda B_\nu, \quad (7.51a)$$

$$A_\nu \sim B_\nu \text{ and } C_\nu \sim D_\nu \Rightarrow A_\nu + C_\nu \sim B_\nu + D_\nu, \quad (7.51b)$$

$$A_\nu \sim B_\nu \Rightarrow A_\nu^{-1} \sim B_\nu^{-1} \quad \text{if } A_\nu > 0, \quad (7.51c)$$

$$A_\nu \sim B_\nu \text{ and } C_\nu \in \mathbb{K}^{J \times I} \Rightarrow C_\nu A_\nu C_\nu^H \sim C_\nu B_\nu C_\nu^H, \quad (7.51d)$$

$$A_\nu \sim B_\nu \Rightarrow A_\nu^{-1/2} B_\nu A_\nu^{-1/2} \sim I \quad \text{if } A_\nu > 0. \quad (7.51e)$$

In the cases (7.51a,b,d), the constant c in (7.50) is identical on both sides. The matrix C_ν in (7.51d) may be any rectangular matrix.

Proof. The implications (7.51a,b,d) are an immediate consequence of (7.48f). For (7.51c), use (7.48d,e). Statement (7.51d) implies (7.51e). \square

We recall that the iteration $\Phi(x, b) = x - \vartheta W^{-1}(Ax - b)$ with optimal damping has the convergence rate $\rho(M_{\vartheta_{\text{opt}}}) = \frac{\kappa-1}{\kappa+1}$ with $\kappa = \kappa(W^{-1}A)$.

Conclusion 7.59. (a) Assume $A, W, W' > 0$ and $W \leq c'W'$, $W' \leq cW$. Then the linear iterations $\Phi(x, b) = x - \vartheta W^{-1}(Ax - b)$ and $\Phi'(x, b) = x - \vartheta W'^{-1}(Ax - b)$ with optimal damping have comparable convergence rates determined by

$$\kappa = \kappa(W^{-1}A), \quad \kappa' = \kappa(W'^{-1}A) \quad \text{with} \quad \frac{1}{c'} \kappa' \leq \kappa \leq c \kappa'.$$

⁶ Consider families of regular matrices. Let $\text{cond}(A) = \|A\| \|A^{-1}\|$ be defined with respect to some submultiplicative matrix norm. Analogously to Remark 7.57b, we define

$$\mathcal{A} \sim_{\text{cond}} \mathcal{B} : \Leftrightarrow \sup_{\nu \in F} \text{cond}(A_\nu^{-1} B_\nu) < \infty.$$

Also in this case, the properties (7.47a,e) prove that \sim_{cond} is an equivalence relation.

(b) If the family $\mathcal{A} = (A_h)_{h \in H}$ is indexed by the step size and $\kappa(W_h^{-1} A_h) = \mathcal{O}(h^{-\tau})$ holds with $\tau > 0$, the convergence of $\Phi_{\vartheta_{\text{opt}}}$ is of the order τ . All linear iterations $\Phi'_h(x, b) = x - \vartheta W_h'^{-1}(A_h x - b)$ with $W_h' \sim W_h$ have the same convergence order τ . Hence the convergence order is a property of the equivalence class.

The optimal convergence order is $\tau = 0$ characterised by $\rho(M_h) \leq c < 1$. In the case of linear iterations $\Phi(x, b) = x - \vartheta W^{-1}(Ax - b)$ with $A, W > 0$, the latter inequality can be ensured by the next statement.

Proposition 7.60. *The family of linear iterations $\Phi_h(x, b) = x - \vartheta_{\text{opt}} W_h^{-1}(A_h x - b)$ with $A_h, W_h > 0$ satisfies*

$$\rho(M_h) \leq c < 1 \quad \text{for all } h \in H$$

if and only if

$$A_h \sim W_h.$$

Proof. $A_h \sim W_h$ implies that $\kappa_h = \kappa(W_h^{-1} A_h) = \mathcal{O}(1)$. Hence $\rho(M_{\vartheta_{\text{opt}}}) = \frac{\kappa_h - 1}{\kappa_h + 1} \leq c < 1$ holds with $c := \sup\{2/(1 + \kappa_h) : h \in H\}$. \square

This result shows a way how to obtain optimal convergence, provided that $W_h^{-1}(A_h x - b)$ is easy to evaluate. As in Remark 7.7 we have to ask on what data the choice of W_h could be based. Using only the data of A_h , the traditional techniques do not lead to $A_h \sim W_h$ in general. In §13.4 we shall propose a new technique which is able to satisfy $A_h \sim W_h$.

7.4.5 Equivalent Bilinear Forms

We recall the Definition E.2 of coercive forms.

Definition 7.61. Two symmetric and coercive sesquilinear forms $a, b : V \times V \rightarrow \mathbb{C}$ are called *equivalent* (notation: $a \sim b$) if there is some $c > 0$ with

$$\frac{1}{c} a(u, u) \leq b(u, u) \leq c a(u, u) \quad \text{for all } u \in V. \quad (7.52)$$

For simplicity, we use the term *bilinear* which suits for $\mathbb{K} = \mathbb{R}$. For $\mathbb{K} = \mathbb{C}$, the form must be sesquilinear (cf. Definition E.1).

The Galerkin matrix A_h corresponding to the bilinear form $a(\cdot, \cdot)$ satisfies

$$\langle A_h x, y \rangle = a(P_h x, P_h y) \quad \text{for all } x, y \in \mathbb{K}^I,$$

where $\langle \cdot, \cdot \rangle$ is the Euclidean scalar product (cf. Exercise E.5). The mapping $P_h : \mathbb{K}^I \rightarrow V_n \subset V$ is defined in (E.6).

Applying (7.52) to $u = P_h x$, we obtain

$$\frac{1}{c} \langle A_h x, x \rangle \leq \langle B_h x, x \rangle \leq c \langle A_h x, x \rangle \quad \text{for all } x \in \mathbb{K}^I, \quad (7.53a)$$

where B_h is the Galerkin matrix corresponding to the bilinear form $b(\cdot, \cdot)$. The symmetry of a and b implies that A_h and B_h are also Hermitian (cf. Exercise E.6a). Therefore the property (7.53a) is equivalent to the inequalities

$$\frac{1}{c} A_h \leq B_h \leq c A_h \quad (7.53b)$$

in the sense of §C.1.1.

Note that the constants in (7.52) and (7.53b) coincide. Therefore they hold for all discretisation parameters $h \in H$ which form the families $\mathcal{A} = (A_h)_{h \in H}$ and $\mathcal{B} = (B_h)_{h \in H}$. Using the notion of equivalence, we obtain the following statement.

Proposition 7.62. *Equivalent forms $a \sim b$ produce equivalent Galerkin matrix families $\mathcal{A} \sim \mathcal{B}$.*

A potential practical strategy is the following. Let a and \mathcal{A} correspond to the problem to be solved. If there is a simpler but equivalent form b , it may be that the corresponding matrices B_h are easier to handle. Either $W_h^{-1} \delta = d$ can be solved for $W_h = B_h$ or for another choice $W_h \sim B_h$. By Proposition 7.62, $A_h \sim W_h$ holds as required in Proposition 7.60.

Conclusion 7.63. *Let the symmetric and coercive form $a(\cdot, \cdot)$ correspond to a boundary value problem for $u \in V := H_0^1(\Omega)$. Use the same finite element discretisation for $a(\cdot, \cdot)$ and the standard Poisson problem. Then both discretisation matrices are spectrally equivalent.*

7.5 Time-Stepping Methods

The term of a time-stepping method is used, in particular, in the engineering community. The function $x(t)$, $0 \leq t < \infty$, is introduced as a solution of the system of ordinary differential equations

$$\frac{d}{dt} x(t) = b - Ax \quad \text{with the initial value } x(0) = x^0. \quad (7.54)$$

If A is positive definite (or if $\Re(\lambda) > 0$ holds for all eigenvalues $\lambda \in \sigma(A)$), then $x(t)$ converges for $t \rightarrow \infty$ to the solution $x^* := A^{-1}b$, which is now interpreted as the stationary solution of (7.54). The time-stepping method tries to discretise the differential equation by a grid

$$0 = t_0 < t_1 < \dots$$

and to approximate $x(t)$ for a large $t = t_m$. One explicit Euler step with the time step $\Delta t := t_{m+1} - t_m$ reads as

$$x(t_{m+1}) \approx x^{m+1} = x^m - \Delta t (Ax^m - b) \quad (7.55)$$

(cf. Quarteroni–Sacco–Saleri [314, §11.2]). For a fixed (or variable) step size Δt , recursion (7.55) describes the stationary (or instationary) Richardson method.

Often Runge–Kutta-like methods are proposed. For example, the Heun method becomes

$$x' := x^m - \alpha \Delta t (Ax^m - b), \quad x(t_{m+1}) \approx x^{m+1} = x^m - \beta \Delta t (Ax' - b) \quad (7.56)$$

with $\alpha = \frac{1}{2}$ and $\beta = 1$ (cf. Heun [220]; in the true Runge–Kutta case, there are four coefficients; cf. Runge [327] and Kutta [250]).

While the original discretisation methods try to achieve small discretisation errors $\|x^m - x(t_m)\|$ for all grid points t_m , the coefficients α, β are now chosen such that the convergence $x^m \rightarrow x^*$ is improved. The produced methods (as, e.g., (7.56)) are the semi-iterative variants of the Richardson iteration which will be described in §8.3.7.

In the language of ordinary differential equations, one explains the unfavourably slow convergence of the Richardson variants by the *stiffness* of the system. When preconditioning is introduced to speed up the convergence:

$$x^{m+1} = x^m - \Delta t W^{-1} (Ax^m - b),$$

this is called a *quasi-time stepping* method, which however does no longer approximate the equation (7.54) but only the same stationary solution x^* .

In essence, the interpretation by a time-stepping method is misleading (e.g., since the high consistency order of a Runge–Kutta method is given up for purposes which are not connected with this method). In particular, this concept is of no help for analysing the iteration or for constructing efficient iterations.

7.6 Nested Iteration

Three families of linear iterations, the multigrid iteration, the domain decomposition methods, and the hierarchical LU iteration will be described in Part III. The multigrid method is usually combined with the *nested iteration*. As shown in §11.5, the nested iteration technique can be combined with any linear or nonlinear iteration. It does not change the iteration, but yields advantageous starting values.

Part II

Semi-Iterations and Krylov Methods

While Part I is dedicated to *linear* iterations, we now consider *nonlinear* approaches. These nonlinear methods are not completely new algorithms, but reuse linear iterations. They may be considered as ‘acceleration methods’ which try to speed up linear iterations. We denote the set of nonlinear algorithms by \mathcal{N} . Any method $\Upsilon \in \mathcal{N}$ requires a linear iteration $\Phi \in \mathcal{L}$ as an argument. Then $\Upsilon[\Phi]$ is a nonlinear mapping of a starting value y^0 into a sequence y^1, y^2, \dots . In contrast to linear iterations, it may happen that the sequence $\{y^m\}$ is not infinite, but terminates either since the exact solution is obtained or since the method breaks down.

Often the nonlinear method Υ is identified with $\Upsilon[\Phi_1^{\text{Rich}}]$, i.e., with its application to the Richardson iteration. In that case, $\Upsilon[\Phi]$ for other $\Phi \in \mathcal{L}$ is called the ‘preconditioned version’ of Υ (preconditioned by N_Φ).

We do not fix the domain and range of $\Upsilon[\Phi]$ since these sets depend on the method. In the simplest case, $y^{m+1} = \Upsilon[\Phi](y^m)$ holds as for the linear case (an example is the gradient method). Alternatively, $\Upsilon[\Phi]$ may be a three-term recursion of the form $y^{m+1} = \Upsilon[\Phi](y^m, y^{m-1})$ with a first iterate $y^1 = \Upsilon[\Phi](y^0)$ (see, e.g., the Chebyshev method). In the standard cases, the iterates y^m are enriched by further auxiliary variables that are updated in each step.

The connection of $\Upsilon[\Phi]$ with Φ can be seen from the fact that the nonlinear iterate y^m belongs to the affine space $x^0 + \text{span}\{x^\mu - x^0 : 1 \leq \mu \leq m\}$, where x^μ are the iterates of Φ with the initial value $x^0 = y^0$. Since the corresponding space is called the *Krylov space*, the nonlinear methods are also called *Krylov methods*.

The most general approach is the semi-iterative method defined and analysed in Chapter 8. Roughly speaking, a semi-iterative method (shortly ‘semi-iteration’) is produced by a linear iteration $\Phi \in \mathcal{L}$ with varying damping factors. The discussion leads to the question of optimal polynomials. Using a simplification of the optimisation task, the optimal polynomial turns out to be the (transformed) Chebyshev polynomial. The related nonlinear method is the Chebyshev method $\Upsilon_{\text{Cheb}} \in \mathcal{N}$ described in §§8.3.4–8.3.5.

Besides the practical relevance of semi-iterative methods, this technique is of fundamental theoretical interest since all other methods can be interpreted as certain semi-iterations, so that semi-iterative error estimates can be applied.

In §8.5 we insert the description of the ADI method which is not really of the form described above, but it might be seen as a generalisation of semi-iterations (replacing scalar parameters by matrix-valued ones). Under conditions, which are not so easy to fulfil in practice, the method has very favourable convergence properties.

Chapter 9 is devoted to the gradient method. An optimisation of greedy type determines the damping factors of the underlying iteration $\Phi \in \mathcal{L}$. It turns out that the gradient method converges as fast as the optimally damped version $\Phi_{\vartheta_{\text{opt}}}$ of Φ , but the method can be applied without knowing the spectral values that determine ϑ_{opt} . In §9.3 we discuss the drawback of the gradient directions and introduce the conjugate directions as a transition from the gradient to the conjugate gradient method.

The conjugate gradient method $\Upsilon_{\text{CG}} \in \mathcal{N}$ introduced in Chapter 10 is the most popular acceleration method. It applies to $\Phi \in \mathcal{L}_{\text{pos}}$ with system matrices $A > 0$ and to $\Phi \in \mathcal{L}_{>0}$. Sections 10.3–10.5 are devoted to CG variants that apply to a larger class of problems.

Chapter 8

Semi-Iterative Methods

Abstract The semi-iteration comes in three formulations. The first one in Section 8.1 is the most general and associates each semi-iterate with a polynomial. Using the notion of Krylov spaces, we only require that the errors of the semi-iterates y^m be elements of the Krylov space $x^0 + N\mathcal{K}_m(AN, r^0)$. In the second formulation of Section 8.2, the polynomials p_m associated with y^m are related either by a two-term or by a three-term recursion. Section 8.3 tries to determine the optimal polynomials. Here the result depends on what quantity we want to minimise. Three minimisation problems are discussed. The last formulation is practically solvable and leads to (transformed) Chebyshev polynomials. The corresponding semi-iteration is called the Chebyshev method (cf. §8.3.4). The Chebyshev method improves the order of convergence. Its convergence speed corresponds to the square root of the spectral condition number (cf. §8.3.5). In Section 8.4 the Chebyshev method is applied to the iterations discussed in Part I. In Section 8.5 we describe the ADI method which is not really of the form discussed above, but it might be seen as a generalisation of semi-iterations (replacing scalar parameters by matrix-valued ones).

8.1 First Formulation

8.1.1 Notation

Let $\Phi \in \mathcal{L}$ be a linear and consistent (not necessarily convergent) iteration with an iteration matrix M . In the following Φ is also called the *basic iteration*. Assume that for a starting iterate x^0 , the iterates

$$x^{m+1} = Mx^m + Nb = \Phi(x^m, b)$$

are computed. Up to now, the last computed iterate x^m is regarded as the result of the iterative process. The previously calculated x^j ($0 \leq j \leq m-1$) are ‘forgotten’. The semi-iterative method is based on a different view. Now, the result of m steps

of the basic iteration Φ is the complete sequence

$$X_m := (x^0, x^1, \dots, x^m) \in (\mathbb{K}^I)^{m+1}. \quad (8.1)$$

We shall investigate whether a better result than x^m can be constructed from X_m . A semi-iterative method is a mapping

$$\Sigma : \bigcup_{m=0}^{\infty} (\mathbb{K}^I)^{m+1} \rightarrow \mathbb{K}^I.$$

The results

$$y^m := \Sigma(X_m) \quad (m = 0, 1, 2, \dots)$$

yield a new sequence: the semi-iterative sequence. We shall see that in many cases $\{y^m\}$ converges faster than $\{x^m\}$.

Remark 8.1. The simple example $y^m = \Sigma(x^0, x^1, \dots, x^m) := x^m$ shows that an optimally chosen semi-iterative method cannot be worse than the basic iteration.

To simplify the notation of polynomials, we introduce the following definition.

Definition 8.2. For $m \in \mathbb{N}_0$, \mathcal{P}_m is the linear space of polynomials of degree $\leq m$ with the underlying field \mathbb{K} . $\mathcal{P}_{-1} := \{0\}$ contains the zero polynomial.

8.1.2 Consistency and Asymptotic Convergence Rate

Similar as in Definition 2.5, a semi-iterative method Σ is called *consistent* if equation (8.2) holds for all solutions of $Ax = b$:

$$x = \Sigma(\underbrace{x, x, \dots, x}_{m+1 \text{ arguments}}) \quad (m = 0, 1, 2, \dots). \quad (8.2)$$

The *convergence rate* $\rho = \rho(M)$ can be characterised as the minimal ρ satisfying

$$\lim_{m \rightarrow \infty} (\|x^m - x\| / \|x^0 - x\|)^{1/m} \leq \rho \text{ for all } x^0 \neq x \quad (\text{cf. Remark 2.22b}).$$

This characterisation can be transferred to the semi-iterative case.

Definition 8.3. The semi-iterative method has the asymptotic convergence rate ρ , if ρ is the smallest number with

$$\overline{\lim_{m \rightarrow \infty}} (\|y^m - x\| / \|y^0 - x\|)^{1/m} \leq \rho \quad (x = A^{-1}b)$$

for all semi-iterative sequences $\{y^m\}$ corresponding to arbitrary starting iterates $y^0 = x^0$.

In the following, we restrict our considerations to linear semi-iterations. Σ is called *linear* if $y^m = \Sigma(X_m)$ is a linear combination

$$y^m = \sum_{j=0}^m \alpha_{mj} x^j \quad (8.3)$$

with coefficients $\alpha_{mj} \in \mathbb{K}$ ($m \in \mathbb{N}_0$, $1 \leq j \leq m$). Obviously, a linear semi-iterative method is *consistent* if and only if

$$\sum_{j=0}^m \alpha_{mj} = 1 \quad \text{for all } m = 0, 1, 2, \dots . \quad (8.4)$$

Applying condition (8.4) to $m = 0$, we find that a consistent semi-iterative method satisfies the initial condition

$$y^0 = x^0. \quad (8.5)$$

8.1.3 Error Representation

Theorem 8.4. *Let x be a solution of $Ax = b$, while M denotes the iteration matrix of the basic iteration $\Phi \in \mathcal{L}$. Then the error*

$$\eta^m := y^m - x \quad (x = A^{-1}b) \quad (8.6a)$$

admits the representation

$$\eta^m = p_m(M) e^0 \quad \text{with } e^0 := x^0 - x, \quad (8.6b)$$

where $y^0 = x^0$ (cf. (8.5)) is the starting iterate and p_m is the polynomial

$$p_m(\zeta) = \sum_{j=0}^m \alpha_{mj} \zeta^j \in \mathcal{P}_m \quad (8.6c)$$

with the coefficients α_{mj} in (8.4).

Proof. Let $e^j = x^j - x$ be the iteration errors of the basic iteration. Subtracting $x = \sum_{j=0}^m \alpha_{mj} x$ from $y^m = \sum_{j=0}^m \alpha_{mj} x^j$ (cf. (8.2) and (8.4)), we obtain the semi-iterative error

$$\eta^m := y^m - x = \sum_{j=0}^m \alpha_{mj} (x^j - x) = \sum_{j=0}^m \alpha_{mj} e^j.$$

Inserting the representation $e^j = x^j - x = M^j e^0$ (cf. (2.16b)), we arrive at

$$\eta^m = \sum_{j=0}^m \alpha_{mj} (M^j e^0) = \left(\sum_{j=0}^m \alpha_{mj} M^j \right) e^0 = p_m(M) e^0. \quad \square$$

Theorem 8.4 associates the linear semi-iteration Σ with a family of polynomials

$$\{p_m \in \mathcal{P}_m : m = 0, 1, \dots\}.$$

Vice versa, any sequence $\{p_m \in \mathcal{P}_m\}$ of polynomials defines a semi-iterative method by means of its coefficients α_{mj} .

Remark 8.5. (a) A linear semi-iterative method Σ is uniquely described by the family of associated polynomial sequence $\{p_m \in \mathcal{P}_m\}$. Σ is consistent if and only if

$$p_m(1) = 1 \quad \text{for } m = 0, 1, \dots. \quad (8.6d)$$

(b) Let the basic iteration with iteration matrix M be consistent. Then the semi-iterates y^m have the representation¹

$$y^m = M_m x^0 + N_m b \quad \text{with } M_m := p_m(M), \quad N_m := (I - M_m)A^{-1}. \quad (8.7)$$

(c) The asymptotic convergence rate is equal to

$$\overline{\lim}_{m \rightarrow \infty} \rho(p_m(M))^{1/m}.$$

If M is diagonalisable, the quantity above coincides with $\overline{\lim} \|p_m(M)\|^{1/m}$. The equality

$$\overline{\lim} \rho(p_m(M))^{1/m} = \overline{\lim} \|p_m(M)\|^{1/m}$$

is not valid in general, but holds for many important polynomial sequences p_m (cf. Eiermann–Niethammer–Varga [120]).

From (8.6d), we derive an alternative characterisation of p_m .

Remark 8.6. (a) Any polynomial $p_m \in \mathcal{P}_m$ satisfying the consistency condition (8.6d) is uniquely associated with a polynomial $q_m \in \mathcal{P}_{m-1}$ so that

$$p_m(\zeta) = 1 - (1 - \zeta) q_m(1 - \zeta). \quad (8.8)$$

(b) $M = I - NA$ (cf. (2.9')) yields

$$p_m(M) = I - NA q_m(NA).$$

¹ The expression $I - M_m$ has the form $X_m A$ so that $(I - M_m)A^{-1} = X_m$ is well defined also for singular A .

8.1.4 Krylov Space

Definition 8.7. The Krylov space associated with a matrix $X \in \mathbb{K}^{I \times I}$ and with a vector $v \in \mathbb{K}^I$ is defined by

$$\mathcal{K}_m(X, v) := \text{span}\{v, Xv, \dots, X^{m-1}v\} \quad \text{for } m \in \mathbb{N},$$

while $\mathcal{K}_0(X, v) := \{0\}$ (cf. Aleksey Nikolaevich Krylov [249]).

Exercise 8.8. Let $\mathcal{U} = \text{span}\{u^1, \dots, u^m\}$ be a subspace of \mathbb{K}^I .

(a) Prove that $\text{span}\{\mathcal{U}, x\} = \text{span}\{\mathcal{U}, y\}$ for any x, y with $x - y \in \mathcal{U}$.

(b) Let $A \in \mathbb{K}^{I \times I}$ be any matrix. $A\mathcal{U}$ abbreviates the subspace $\{Ax : x \in \mathcal{U}\}$. Prove that $A\mathcal{U} = \text{span}\{Au^1, \dots, Au^m\}$.

Since the monomials $\{1, x, \dots, x^{m-1}\}$ span the space \mathcal{P}_{m-1} of polynomials of degree $\leq m-1$, we obtain the first statement of the next remark. There we use the notation $v + \mathcal{U} := \{v + u : u \in \mathcal{U}\}$ for the *affine subspace* with a subspace $\mathcal{U} \subset \mathbb{K}^I$ and a vector $v \in \mathbb{K}^I$. The *residual* of an approximation \tilde{x} is defined by $r := b - A\tilde{x}$ and is the negative defect (2.17).

Proposition 8.9. (a) The connection with matrix polynomials is given by

$$\mathcal{K}_m(X, v) = \{p(X)v : p \in \mathcal{P}_{m-1}\}.$$

(b) Assume that the iteration Φ with the iteration matrix $M = I - NA$ yields the iterates x^m with the errors $e^m = x^m - x$ and the residuals $r^m := b - Ax^m$. They satisfy

$$\begin{aligned} x^m &\in x^0 + N\mathcal{K}_m(AN, r^0) = x^0 + NAK_m(NA, e^0) \subset x + \mathcal{K}_{m+1}(NA, e^0), \\ e^m &\in e^0 + N\mathcal{K}_m(AN, r^0) = e^0 + NAK_m(NA, e^0) \subset \mathcal{K}_{m+1}(NA, e^0), \\ r^m &\in r^0 + AN\mathcal{K}_m(AN, r^0) \subset \mathcal{K}_{m+1}(AN, r^0), \end{aligned}$$

and

$$\begin{aligned} \text{span}\{e^0, \dots, e^{m-1}\} &= \mathcal{K}_m(M, e^0) = \mathcal{K}_m(NA, e^0), \\ \text{span}\{r^0, \dots, r^{m-1}\} &= \mathcal{K}_m(M, r^0) = \mathcal{K}_m(NA, r^0). \end{aligned}$$

(c) The following identity holds for regular T :

$$T\mathcal{K}_m(X, v) = \mathcal{K}_m(TXT^{-1}, Tv).$$

(d) For $m \in \mathbb{N}_0$, we have

$$X\mathcal{K}_m(X, v) \subset v + X\mathcal{K}_m(X, v) \subset \text{span}\{v\} + X\mathcal{K}_m(X, v) = \mathcal{K}_{m+1}(X, v). \quad (8.9)$$

Proof. The statements in part (b) follow by induction. Note that $\mathcal{K}_m(M, v) = \mathcal{K}_m(NA, v)$ holds for all v since a polynomial in $M = I - NA$ can be written as a polynomial of same degree in NA . The inclusions use part (d).

Part (c) is a consequence of Exercise A.16a. \square

Definition 8.10. The degree of a vector $v \in \mathbb{K}^I$ (with respect to a matrix $X \in \mathbb{K}^{I \times I}$) is defined by

$$\deg_X(v) := \min \{m \in \mathbb{N}_0 : p(X)v = 0 \text{ for } p \in \mathcal{P}_m \text{ with } \deg(p) = m\}.$$

Exercise 8.11. For $m \in \mathbb{N}$, prove: (a) $\dim(\mathcal{K}_m(X, v)) = \min\{m, \deg_X(v)\} \leq m$.

(b) If $\dim(\mathcal{K}_{m+1}(X, v)) = \dim(\mathcal{K}_m(X, v))$, then $\mathcal{K}_{m+1}(X, v) = \mathcal{K}_m(X, v)$. If, in addition, X is regular, $X\mathcal{K}_m(X, v) = \mathcal{K}_m(X, v)$ also holds.

(c) $\deg_X(v) = 0$ holds if and only if $v = 0$, while $\deg_X(v) = 1$ characterises all eigenvectors of X .

(d) $\deg_X(v) \leq \deg(\mu_X) \leq \#I$, where μ_X is the minimum function (A.16c).

(e) Any $w \in \mathcal{K}_m(X, v)$ is characterised by a polynomial $p \in \mathcal{P}_{m-1}$ via $w = p(X)v$. If $\dim(\mathcal{K}_m(X, v)) = m$, this polynomial is unique.

Lemma 8.12. For any $v \in \mathbb{K}^I$ and any regular matrix X , the polynomial p with $p(X)v = 0$ and $\deg(p) = \deg_X(v)$ satisfies $p(0) \neq 0$.

Proof. If $p(0) = 0$, there is a polynomial $q \in P_{\deg_X(v)-1}$ with $p(\xi) = \xi q(\xi)$. Hence $0 = p(X)v = Xq(X)v$ implies that $q(X)v = 0$ in contradiction to the minimality of $\deg_X(v)$. \square

Combining Proposition 8.9a with Theorem 8.4 and repeating the arguments of Proposition 8.9, we obtain the next statement.

Conclusion 8.13. (a) The first formulation of a semi-iteration is equivalent to

$$y^m \in x^0 + N\mathcal{K}_m(AN, r^0) \subset x + \mathcal{K}_{m+1}(NA, e^0),$$

where $x := A^{-1}b$. The polynomial (8.6c) coincides with the polynomial associated with the error $\eta^m = y^m - x \in \mathcal{K}_{m+1}(NA, e^0)$ in (8.6a) by Exercise 8.11e.

(b) If the polynomials in (8.6c) satisfy $\deg(p_\mu) = \mu$, the errors η^m span

$$\mathcal{K}_m(M, e^0) = \text{span}\{\eta^0, \eta^1, \dots, \eta^{m-1}\}.$$

(c) The residuals $r^m = -A\eta^m = b - Ay^m$ of the semi-iterates span the space $A \text{span}\{\eta^0, \dots, \eta^{m-1}\}$. Under the conditions of part (b), Proposition 8.9c yields

$$\text{span}\{r^0, r^1, \dots, r^{m-1}\} = A\mathcal{K}_m(M, e^0) = \mathcal{K}_m(AN, r^0).$$

8.2 Second Formulation of a Semi-Iterative Method

8.2.1 General Representation

The representation used in §8.1 requires storing all iterates (x^0, x^1, \dots, x^m) , which is not desirable in the case of large m and high-dimensional systems. Since, in §8.1, the definition of $y^m = \Sigma(X_m)$ is completely independent of the previous iterates $y^j = \Sigma(X_j)$ ($0 \leq j \leq m-1$), it is in general not possible to use the semi-iterative results y^0, \dots, y^{m-1} for computing y^m .

This situation changes in the second formulation. Let $\Phi \in \mathcal{L}$ be the basic iteration. After starting with

$$y^0 = x^0 \quad (\text{cf. (8.5)}), \quad (8.10a)$$

we compute the iterates recursively by

$$y^m = \vartheta_m \Phi(y^{m-1}, b) + (1 - \vartheta_m)y^{m-1} \quad (m \geq 1) \quad (8.10b)$$

with *extrapolation factors* $\vartheta_m \in \mathbb{K}$ ($m \in \mathbb{N}$) that may be chosen arbitrarily.

Exploiting the normal forms $\Phi(x, b) = Mx + Nb = x - N(Ax - b)$, equation (8.10b) can be written in the form (8.10b') or (8.10b''r'):

$$y^m = \vartheta_m (My^{m-1} + Nb) + (1 - \vartheta_m)y^{m-1}, \quad (8.10b')$$

$$y^m = y^{m-1} - \vartheta_m N(Ay^{m-1} - b) = \Phi_{\vartheta_m}(y^{m-1}, b). \quad (8.10b''r')$$

Formulae (8.10b',b'') represent one step of the damped version Φ_{ϑ_m} of the basic iteration (cf. §5.2.2), however with a parameter ϑ_m depending on m .

Below we state that recursion (8.10a,b) yields a semi-iterative method.

Theorem 8.14. *For arbitrary factors $\vartheta_m \in \mathbb{K}$ ($m \in \mathbb{N}$), algorithm (8.10a,b) defines a linear and consistent semi-iteration Σ . The polynomials $\{p_m \in \mathcal{P}_m\}$ describing Σ are recursively defined by*

$$p_0(\zeta) = 1, \quad p_m(\zeta) = (\vartheta_m \zeta + 1 - \vartheta_m) p_{m-1}(\zeta) \quad (m \in \mathbb{N}). \quad (8.11)$$

Proof. (i) One shows by induction that the polynomials p_m in (8.11) satisfy the consistency condition (8.6d): $p_m(1) = 1$. Also $\deg(p_m) \leq m$ is obvious.

(ii) The basic iteration Φ is assumed to be consistent. By construction (8.10b'), the first matrix M_m in the representation $y^m = M_m x^0 + N_m b$ has the form $M_m = \vartheta_m M M_{m-1} + (1 - \vartheta_m) M_{m-1}$, where $M_0 = I$. According to (8.7), the polynomials in (8.11) lead to the same matrix $M_m = p_m(M)$. Since these matrices uniquely determine y^m because of $N_m := (I - M_m)A^{-1}$ (using the consistency of Φ), the method (8.10a,b) coincides with the semi-iteration defined by the polynomials (8.11). The case of an inconsistent basic iteration is left to the reader (proof by induction). \square

The case $\vartheta_m = 0$ is uninteresting because of $y^m = y^{m-1}$. Therefore, we assume that $\vartheta_m \neq 0$. The set of all methods representable by (8.10a,b) is characterised next.

Lemma 8.15. Let $\Phi \in \mathcal{L}$ be the basic iteration and assume $\vartheta_m \neq 0$ in (8.10b). Then the second formulation (8.10a,b) represents exactly those linear and consistent semi-iterations for which the associated polynomials p_m satisfy (8.6d) and

$$\text{degree}(p_m) = m, \quad p_{m-1} \text{ is a divisor of } p_m \quad \text{for all } m \geq 1. \quad (8.12a)$$

Given polynomials $\{p_m\}$ with (8.6d) and (8.12a), the extrapolation factors ϑ_m of the equivalent representation (8.10a,b) are determined by

$$\frac{p_m(\zeta)}{p_{m-1}(\zeta)} = 1 + \vartheta_m(\zeta - 1). \quad (8.12b)$$

Proof. In the case of $\vartheta_m \neq 0$, the method (8.10a,b) leads to polynomials (8.11) satisfying $\text{degree}(p_m) = m$; hence, (8.12a) is satisfied. Vice versa, under the assumption (8.12a), p_m/p_{m-1} must be a polynomial of the form (8.12b). \square

The example of recursion (8.10a,b) shows that the mapping $X_m \mapsto y^m = \Sigma(X_m)$ does not need the iterates of X_m explicitly. Since X_m is uniquely determined by x^0 , there is a mapping $\Xi : x^0 \mapsto y^m$ for $y^m = \Sigma(X_m)$. Recursion (8.10a,b) describes such a mapping Ξ .

By Lemma 8.15, the semi-iterate y^m for a fixed m can be produced as follows.

Remark 8.16. y^m is connected with a polynomial p_m . Let

$$p_m(\zeta) = c_m \prod_{\nu=1}^m (\zeta - \zeta_\nu) \quad \text{with} \quad c_m = 1 / \prod_{\nu=1}^m (1 - \zeta_\nu) \quad (8.13a)$$

be a factorisation into linear factors (possibly with complex ζ_ν) and define auxiliary polynomials \hat{p}_μ for $0 \leq \mu \leq m$ by

$$\hat{p}_\mu(\zeta) = \prod_{\nu=1}^\mu \frac{\zeta - \zeta_\nu}{1 - \zeta_\nu}. \quad (8.13b)$$

Set $\vartheta_\mu := \frac{1}{1 - \zeta_\mu}$ for $0 < \mu \leq m$. Then all polynomials \hat{p}_μ satisfy (8.12a,b) and $\hat{p}_m = p_m$. The corresponding semi-iteration

$$\hat{y}^\mu = \Phi_{\vartheta_\mu}(\hat{y}^{\mu-1}, b) \quad (1 \leq \mu \leq m; \text{ cf. (8.10a,b)})$$

is as easy to perform and yields $\hat{y}^m = y^m$ (only for $\mu = m$, not for $\mu < m$). However, this approach has severe disadvantages.

1. To compute the next y^{m+1} , we have to perform (8.10a,b) again from $\mu = 0$ to $\mu = m + 1$, since then other auxiliary polynomials \hat{p}_μ are needed.
2. The second formulation (8.10a,b) may be unstable. For relative small m , the rounding error influence of the iteration errors $y^m - x$ can already predominate. It is possible to avoid instability by a suitable renumbering of the ϑ_ν . Concerning the stability analysis and the choice of an appropriate ordering, we refer to Lebedev–Finogenov [261, 262] (cf. also Samarskii–Nikolaev [330, §6.2.4]).

It will turn out that the three-term recursion described next is the best representation of the polynomials.

8.2.2 Three-Term Recursion

Algorithm (8.10b) determines y^m from y^{m-1} . Alternatively, a three-term recursion connects y^m with y^{m-1} and y^{m-2} (cf. §2.2.8):

$$y^0 = x^0, \quad (8.14a)$$

$$y^1 = (1 - \frac{1}{2}\vartheta_1)x^1 + \frac{1}{2}\vartheta_1x^0 = (1 - \frac{1}{2}\vartheta_1)\Phi(x^0, b) + \frac{1}{2}\vartheta_1x^0, \quad (8.14b)$$

$$y^m = \Theta_m[\Phi(y^{m-1}, b) - y^{m-2}] + \vartheta_m(y^{m-1} - y^{m-2}) + y^{m-2}. \quad (8.14c)$$

From $\Phi(x, b) = Mx + Nb = x - N(Ax - b)$, we obtain the representations

$$\begin{aligned} y^1 &= (1 - \frac{1}{2}\vartheta_1)(Mx^0 + Nb) + \frac{1}{2}\vartheta_1x^0 \\ &= x^0 - (1 - \frac{1}{2}\vartheta_1)N(Ax^0 - b), \end{aligned}$$

$$\begin{aligned} y^m &= \Theta_m(My^{m-1} + Nb - y^{m-2}) + \vartheta_m(y^{m-1} - y^{m-2}) + y^{m-2} \\ &= (1 + \vartheta_m + \Theta_m)y^{m-2} + (\vartheta_m + \Theta_m)(y^{m-1} - y^{m-2}) - \Theta_mN(Ay^{m-1} - b). \end{aligned}$$

Analogous to Theorem 8.14, one proves the next theorem.

Theorem 8.17. For arbitrary factors Θ_m and ϑ_m , algorithm (8.14a–c) defines a linear and consistent semi-iteration Σ . The polynomials $\{p_m\}$ describing Σ are recursively defined by

$$p_0(\zeta) = 1, \quad p_1(\zeta) = (1 - \frac{1}{2}\vartheta_1)\zeta + \frac{1}{2}\vartheta_1, \quad (8.15a)$$

$$p_m(\zeta) = (\Theta_m\zeta + \vartheta_m)p_{m-1}(\zeta) + (1 - \Theta_m - \vartheta_m)p_{m-2}(\zeta). \quad (8.15b)$$

For the particular choice $\vartheta_m = 0$, the recursion becomes

$$p_0(\zeta) = 1, \quad p_1(\zeta) = \zeta, \quad (8.15c)$$

$$p_m(\zeta) = \Theta_m[\zeta p_{m-1}(\zeta) - p_{m-2}(\zeta)] + p_{m-2}(\zeta). \quad (8.15d)$$

We remark that all orthogonal polynomials can be generated by recursion of the form (8.15a,b) (cf. Quarteroni–Sacco–Saleri [314, §10.1]).

Exercise 8.18. Prove that the polynomials q_m in (8.8) associated with p_m and defined either in (8.15a,b) or (8.15c,d) can be determined by the recursion

$$q_0(\xi) = 0, \quad q_1(\xi) = 1 - \frac{1}{2}\vartheta_1,$$

$$q_m(\xi) = \Theta_m + (1 - \Theta_m - \vartheta_m)q_{m-2}(\xi) + (\Theta_m(1 - \xi) + \vartheta_m)q_{m-1}(\xi)$$

or, respectively,

$$q_0(\xi) = 0, \quad q_1(\xi) = 1,$$

$$q_m(\xi) = \Theta_m + (1 - \Theta_m)q_{m-2}(\xi) + \Theta_m(1 - \xi)q_{m-1}(\xi).$$

8.3 Optimal Polynomials

Since the semi-iterates are completely determined by polynomials, we can ask for the best polynomials in the sense that the corresponding semi-iteration is as fast as possible. The quantity to be minimised is still to be specified. It might be a certain norm of the error (cf. Problem 8.19) or the convergence rate (cf. Problem 8.21) or an upper bound of the error (cf. Problem 8.20).

8.3.1 Minimisation Problem

Let Σ be a linear and consistent semi-iteration. By Theorem 8.4, the semi-iteration error $\eta^m = y^m - x$ has the representation (8.6b):

$$\eta^m = p_m(M)e^0.$$

Therefore, it seems reasonable to pose the following problem.

Problem 8.19 (first minimisation problem). Given $m \in \mathbb{N}$, determine a polynomial $p_m \in \mathcal{P}_m$ satisfying (8.6d), i.e.,

$$p_m(1) = 1, \quad (8.16)$$

such that

$$\|p_m(M)e^0\|_2 \stackrel{!}{=} \min, \quad (8.17)$$

i.e., $\|p_m(M)e^0\|_2 \leq \|q_m(M)e^0\|_2$ for all admissible polynomials.

The solution of (8.17) seems hopeless, since the unknown error $e^0 = x^0 - x$ is involved in the problem (if e^0 were known, $x = x^0 - e^0$ already represents the solution). Nevertheless, we shall solve this problem with respect to the energy norm instead of $\|\cdot\|_2$ in §9.3 (cf. Remark 10.12).

Even if e^0 is unknown, $\|p_m(M)e^0\|_2$ can be estimated by

$$\|p_m(M)e^0\|_2 \leq \|p_m(M)\|_2 \|e^0\|_2$$

and the factor $\|p_m(M)\|_2$ can be minimised separately.

Problem 8.20 (second minimisation problem). Given $m \in \mathbb{N}$, determine a polynomial $p_m \in \mathcal{P}_m$ with (8.16) such that

$$\|p_m(M)\|_2 \stackrel{!}{=} \min. \quad (8.18)$$

8.3.2 Discussion of the Second Minimisation Problem

A partial answer to the minimisation problems follows in Theorem A.37 (Cayley–Hamilton). Assume that M has no eigenvalue $\lambda = 1$ ($\rho(M) < 1$ is sufficient). For all $m \geq n := \#I$, the choice $p_m(\lambda) = \chi(\lambda) := \det(\lambda I - M)/\det(I - M)$ leads to a polynomial with the properties (8.16) and $\deg(p_m) \leq m$ solving problems (8.17) and (8.18). In particular, (8.19) holds:

$$p_m(M) = 0 \quad \text{and} \quad \|p_m(M)\| = 0. \quad (8.19)$$

The minimum function $p_m(\lambda) = \mu(\lambda)$ of M (cf. (A.16c)) already satisfies (8.19) for $m \geq m_\mu := \deg(\mu)$.

The solution given in (8.19) is unsatisfactory for two reasons. First, the characteristic polynomial χ (more precisely, its coefficients) is not easy to compute; second, the case $m \geq n$ is rather uninteresting.

Intermediately, we require that

$$M \text{ be normal}, \quad (8.20)$$

i.e., $MM^H = M^HM$ (M being Hermitian would be sufficient). Since then $p_m(M)$ is also normal, Theorem B.25 implies that

$$\|p_m(M)\|_2 = \rho(p_m(M)) = \max\{|p_m(\lambda)| : \lambda \in \sigma(M)\}.$$

Therefore, minimising (8.18) is equivalent to determining a polynomial whose absolute value is minimal on the set $\sigma(M)$. Even if the normality (8.20) does not hold, minimisation of $\max\{|p_m(\lambda)| : \lambda \in \sigma(M)\}$ makes sense. The new minimisation problem is

$$\rho(p_m(M)) = \max\{|p_m(\lambda)| : \lambda \in \sigma(M)\} \stackrel{!}{=} \min, \quad (8.21a)$$

i.e., the spectral radius is minimised over all admissible polynomial in \mathcal{P}_m instead of the spectral norm $\|p_m(M)\|_2$.

For the next interpretation, we assume that $M = T^{-1}DT$ (D diagonal matrix) is diagonalisable. This leads to $p_m(M) = p_m(T^{-1}DT) = T^{-1}p_m(D)T$. Using the norm $\|\cdot\|_T$ defined in Exercise B.13c, we obtain

$$\begin{aligned} \|p_m(M)\|_T &= \|Tp_m(M)T^{-1}\|_2 = \|p_m(D)\|_2 = \rho(p_m(D)) \\ &= \rho(p_m(M)) = \max\{|p_m(\lambda)| : \lambda \in \sigma(M)\}. \end{aligned} \quad (8.21b)$$

Alternatively, we may estimate by

$$\|p_m(M)\|_2 \leq \text{cond}_2(T)\|p_m(D)\| = \text{cond}_2(T)\rho(p_m(M)). \quad (8.21c)$$

Hence minimising the spectral radius $\rho(p_m(M))$ in (8.21a) minimises the upper bound $\text{cond}_2(T)\|p_m(D)\|$ in (8.21c).

According to §5.1.2, symmetric iterations have the property that $A > 0$ implies that $A^{1/2}MA^{-1/2}$ is also Hermitian. Then the energy norm of $p_m(M)$ is well defined and equal to

$$\|p_m(M)\|_A = \max\{|p_m(\lambda)| : \lambda \in \sigma(M)\} = \rho(p_m(M)). \quad (8.21d)$$

The minimisation of $\max\{|p_m(\lambda)| : \lambda \in \sigma(M)\}$ can only be solved with the knowledge of the spectrum $\sigma(M)$. Computing the complete spectrum, however, would be by far more expensive than the solution of the system.

As a remedy, we assume that there is an a priori known superset

$$\sigma_M \supset \sigma(M)$$

containing the spectrum. Then $\sigma(M)$ can be replaced with σ_M . An example for the larger set σ_M is the complex circle

$$\sigma_M = \{\lambda \in \mathbb{C} : |\lambda| \leq \bar{\rho}\} \quad \text{with } \bar{\rho} \geq \rho(M).$$

Unfortunately, this circle is inappropriate for our purposes as we shall see in Theorem 8.32. If, however, M has only real eigenvalues, the interval

$$\sigma_M = [-\bar{\rho}, \bar{\rho}] \quad \text{with } \bar{\rho} \geq \rho(M) \quad (8.22a)$$

is a candidate. In some cases, it is known that M has only nonnegative eigenvalues (cf. Theorem 3.34c). Then one may choose

$$\sigma_M = [0, \bar{\rho}] \quad \text{with } \bar{\rho} \geq \rho(M). \quad (8.22b)$$

In all cases, it is sufficient to know an upper bound $\bar{\rho}$ of $\rho(M)$, where $\bar{\rho} = \rho(M)$ would be optimal and $\bar{\rho} < 1$ must hold. For instance, we may choose $\bar{\rho}$ as $\rho_{m+k,k}$ in (2.23b) for suitable m and k (cf. Remark 2.32).

Accordingly, the minimisation of $\|p_m(M)\|_2$ in Problem 8.20 is replaced with the following minimisation.

Problem 8.21 (third minimisation problem). Given $m \in \mathbb{N}$ and σ_M , determine a polynomial $p_m \in \mathcal{P}_m$ with (8.16) such that

$$\max\{|p_m(\lambda)| : \lambda \in \sigma_M\} \stackrel{!}{=} \min. \quad (8.23)$$

Finally, we briefly discuss the choice of alternative norms in (8.17) and (8.18). A non-Hilbert norm (as, e.g., the maximum or row-sum norm $\|\cdot\|_\infty$) leads to a considerably more complicated minimisation problem. It would be possible to replace the Euclidean norm $\|\cdot\|_2$ by $\|\cdot\|_T = \|Tx\|_2$ or $\|\cdot\|_K = \|K^{1/2}x\|_2$ (K positive definite) as already done in (8.21b,d). Examples for K would be A and the matrix W of the third normal form (cf. (3.35e) and (8.21d)).

8.3.3 Chebyshev Polynomials

As a preparation for the next section we discuss the Chebyshev polynomials.

Definition 8.22. The Chebyshev polynomials T_m are defined by

$$T_m(x) := \cos(m \arccos x) \quad \text{for } m \in \mathbb{N}_0, -1 \leq x \leq 1. \quad (8.24)$$

Part (a) of the following theorem summarising all properties needed later shows that the functions T_m are in fact polynomials of degree m .

Lemma 8.23. (a) The functions T_m in (8.24) fulfil the recursion

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_{m+1}(x) = 2xT_m(x) - T_{m-1}(x). \quad (8.25a)$$

(b) For $x \geq 1$, the polynomials T_m have the representation

$$T_m(x) = \cosh(m \operatorname{arcosh} x) \quad \text{for } m \in \mathbb{N}_0, x > 1, \quad (8.25b)$$

where $\cosh(x) = \frac{e^x + e^{-x}}{2}$ is the hyperbolic cosine, while arcosh (area-hyperbolic cosine) is its inverse function.

(c) For all $x \in \mathbb{C}$, the representation (8.25c) holds:

$$T_m(x) = \frac{1}{2} \left[\left(x + \sqrt{x^2 - 1} \right)^m + \left(x + \sqrt{x^2 - 1} \right)^{-m} \right]. \quad (8.25c)$$

Proof. Eqs. (8.25a) follows from the cosine addition theorem. For (8.25b), it is sufficient to prove that the functions defined there also satisfy recursion (8.25a). Substituting $x = \cos \zeta$, we see that (8.25c) coincides with $\cos(m\zeta) = T_m(x)$. \square

$\{T_m\}$ are orthogonal polynomials with respect to the weight function $\frac{1}{\sqrt{1-x^2}}$, i.e., $\int_{-1}^1 \frac{T_m(x) T_n(x)}{\sqrt{1-x^2}} dx = 0$ for $n \neq m$ (cf. Quarteroni–Sacco–Saleri [314, §10.1.1]).

8.3.4 Chebyshev Method (Solution of the Third Minimisation Problem)

As in the examples (8.22a,b), we assume that σ_M is a real interval. The solution to the third minimisation problem (8.23) is given below.

Notation 8.24. In the following, the real numbers a, b with $-\infty < a \leq b < 1$ define an interval with the property

$$\sigma_M = [a, b] \supset \sigma(M). \quad (8.26a)$$

Because $M = I - NA = I - W^{-1}A$ (cf. (2.9)), inclusion (8.26a) is equivalent to

$$[\gamma, \Gamma] \supset \sigma(NA) = \sigma(W^{-1}A) \quad (8.26b)$$

with

$$\gamma = 1 - b, \quad \Gamma = 1 - a. \quad (8.26c)$$

Note that $0 < \gamma \leq \Gamma < \infty$. Often, the use of γ and Γ leads to simpler formulae. In particular, the ratio

$$\kappa = \Gamma/\gamma \quad (8.26d)$$

is of interest. If the inclusion (8.26a) is strict, i.e., $a, b \in \sigma(M)$, $[\gamma, \Gamma] \supset \sigma(NA)$ is also strict and $\kappa = \kappa(NA)$ is the spectral number defined in (B.13).

Lemma 8.25. *Let $[a, b]$ be an interval with $-\infty < a \leq b < 1$. The problem*

$$\begin{aligned} &\text{minimise } \max\{|p_m(\lambda)| : a \leq \lambda \leq b\} \\ &\text{with respect to all polynomials } p_m \in \mathcal{P}_m \text{ and } p_m(1) = 1 \end{aligned}$$

has the unique solution

$$p_m(\zeta) = T_m\left(\frac{2\zeta-a-b}{b-a}\right)/C_m \quad \text{with } C_m := T_m\left(\frac{2-a-b}{b-a}\right) = T_m\left(\frac{\Gamma+\gamma}{\Gamma-\gamma}\right). \quad (8.27a)$$

Here, γ, Γ are as in (8.26c) and T_m is the Chebyshev polynomial defined in (8.24). The minimising polynomial p_m has the degree m and leads to the minimum

$$\max\{|p_m(\lambda)| : a \leq \lambda \leq b\} = 1/C_m \quad \text{for } p_m \text{ in (8.27a).} \quad (8.27b)$$

Proof. (i) The constant C_m does not vanish, since the argument $\frac{2-a-b}{b-a}$ lies outside of $[-1, 1]$ and the representation (8.25b) applies. By construction, $p_m(1) = 1$ and $\deg(p_m) = m$ hold. For $a \leq \zeta \leq b$, the argument $\frac{2\zeta-a-b}{b-a}$ belongs to $[-1, 1]$. Definition (8.24) shows that $|T_m| \leq 1$ in $[-1, 1]$. Since T_m attains the bounds ± 1 , the statement (8.27b) follows.

(ii) It remains to show that for any other polynomial the maximum in (8.27b) is larger than $1/C_m$. Let $q_m \in \mathcal{P}_m$ be a polynomial with $q_m(1) = 1$ and $\max\{|q_m(\lambda)| : \gamma \leq \lambda \leq \Gamma\} \leq 1/C_m$. The Chebyshev polynomial $T_m(x) = \cos(m \arccos x)$ meets the values ± 1 in alternating ordering at $x = \cos \frac{n\pi}{m}$ for $n = -m, 1-m, \dots, 0$. The function p_m obtained from T_m by transforming $x \mapsto \zeta = \frac{1}{2}[a + b + x(b-a)]$ is $p_m(\frac{1}{2}[a + b + x(b-a)]) := T_m(x)$ and has the values

$$p_m(\zeta_\nu) = (-1)^\nu / C_m \quad (-m \leq \nu \leq 0)$$

at $\zeta_\nu = \frac{1}{2}[a + b + (b-a) \cos \frac{\nu\pi}{m}]$. From $|q_m(\zeta_\nu)| \leq 1/C_m = |p_m(\zeta_\nu)|$, we conclude that the difference $r := p_m - q_m$ satisfies

$$r(\zeta_\nu) \geq 0 \quad \text{for even } \nu, \quad r(\zeta_\nu) \leq 0 \quad \text{for odd } \nu.$$

By the intermediate value theorem, there exists at least one zero of r in each subinterval $[\zeta_{\nu-1}, \zeta_\nu]$ ($1-m \leq \nu \leq 0$). If the zeros in $[\zeta_{\nu-1}, \zeta_\nu]$ and $[\zeta_\nu, \zeta_{\nu+1}]$

coincide at the common point ζ_ν , this is a double zero. Hence, counted with respect to multiplicity, r has at least m zeros in $[a, b]$. By $p_m(1) = q_m(1) = 1$, the value $1 \notin [a, b]$ represents the $(m+1)$ -th zero of r . Hence, $r=0$ follows from $\text{degree}(r) \leq m$, proving that $p_m = q_m$ is unique. \square

Exercise 8.26. (a) Prove by means of (8.25a) that the polynomials p_m in (8.27a) can be obtained by the recursion

$$p_0(\zeta) = 1, \quad p_1(\zeta) = \frac{2\zeta - a - b}{2 - a - b}, \quad (8.28a)$$

$$C_{m+1} p_{m+1}(\zeta) = 2 \frac{2\zeta - a - b}{2 - a - b} C_m p_m(\zeta) - C_{m-1} p_{m-1}(\zeta). \quad (8.28b)$$

(b) Let $\vartheta_{\text{opt}} = \frac{2}{\Gamma + \gamma}$ (cf. (6.6a)). Prove that

$$p_m(I - NA) = \frac{1}{C_m} T_m \left(\frac{\Gamma + \gamma}{\Gamma - \gamma} I + \frac{2}{\Gamma - \gamma} NA \right) = \frac{1}{C_m} T_m \left(\frac{\Gamma + \gamma}{\Gamma - \gamma} [I + \vartheta_{\text{opt}} NA] \right).$$

To investigate the minimum $1/C_m = 1/T_m(\frac{2-a-b}{b-a})$ reached in (8.27b), we have to evaluate (8.25c) at

$$x_0 := \frac{2 - a - b}{b - a} = \frac{\Gamma + \gamma}{\Gamma - \gamma}$$

with γ and Γ defined in (8.26c). We use that $x_0^2 - 1 = 4\gamma\Gamma/(\Gamma - \gamma)^2 > 0$ and $x_0 + \sqrt{x_0^2 - 1} = (\sqrt{\Gamma} + \sqrt{\gamma})^2/(\Gamma - \gamma)$. The representation (8.25c) shows that

$$C_m = \frac{1}{2} \left\{ \left(\frac{(\sqrt{\Gamma} + \sqrt{\gamma})^2}{\Gamma - \gamma} \right)^m + \left(\frac{(\sqrt{\Gamma} + \sqrt{\gamma})^2}{\Gamma - \gamma} \right)^{-m} \right\}.$$

The bracket $\frac{(\sqrt{\Gamma} + \sqrt{\gamma})^2}{\Gamma - \gamma}$ can be rewritten as $\frac{\Gamma}{\Gamma - \gamma} (1 + \sqrt{\frac{\gamma}{\Gamma}})^2$. To simplify the expression, we introduce

$$\kappa := \frac{\Gamma}{\gamma} \quad \text{and} \quad c := \left(1 - \frac{1}{\sqrt{\kappa}} \right) / \left(1 + \frac{1}{\sqrt{\kappa}} \right) \quad (\text{cf. (8.26d)}).$$

Since $\frac{\Gamma - \gamma}{\Gamma} = 1 - \frac{1}{\kappa} = (1 - \frac{1}{\sqrt{\kappa}})(1 + \frac{1}{\sqrt{\kappa}})$ and $1 + \sqrt{\frac{\gamma}{\Gamma}} = 1 + \frac{1}{\sqrt{\kappa}}$, we arrive at

$$\frac{(\sqrt{\Gamma} + \sqrt{\gamma})^2}{\Gamma - \gamma} = \frac{1 + \frac{1}{\sqrt{\kappa}}}{1 - \frac{1}{\sqrt{\kappa}}} = \frac{1}{c}.$$

Hence, the expression for $1/C_m$ reduces to

$$\frac{1}{C_m} = \frac{2c^m}{1 + c^{2m}} \quad \text{with} \quad c = \frac{1 - 1/\sqrt{\kappa}}{1 + 1/\sqrt{\kappa}} = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}, \quad \kappa = \frac{\Gamma}{\gamma}. \quad (8.28c)$$

For the interpretation of κ as a spectral condition number, compare with Notation 8.24.

Conclusion 8.27. (a) For the case (8.22a), i.e., $\sigma_M = [-\bar{\rho}, \bar{\rho}]$ with $0 < \bar{\rho} < 1$, the solution to the third minimisation problem (8.23) is:

$$p_m(\zeta) = T_m(\zeta/\bar{\rho}) / C_m \quad \text{with } C_m := T_m(1/\bar{\rho}). \quad (8.29a)$$

(b) For the case (8.22b): $\sigma_M = [0, \bar{\rho}]$ with $0 < \bar{\rho} < 1$, the respective solution becomes

$$p_m(\zeta) = T_m\left(\frac{2\zeta - \bar{\rho}}{\bar{\rho}}\right) / C_m \quad \text{with } C_m := T_m\left(\frac{2 - \bar{\rho}}{\bar{\rho}}\right). \quad (8.29b)$$

(c) The respective attained minima are

$$\frac{1}{C_m} = \frac{2c^m}{1 + c^{2m}} \quad \text{with} \quad c = \begin{cases} \frac{2\bar{\rho}}{(\sqrt{1+\bar{\rho}} + \sqrt{1-\bar{\rho}})^2} & \text{for (8.29a),} \\ \frac{\bar{\rho}}{(1 + \sqrt{1-\bar{\rho}})^2} & \text{for (8.29b).} \end{cases}$$

(d) If NA is diagonalisable by a transformation T (cf. (8.21c)), the semi-iterates y^m satisfy the error estimate

$$\begin{aligned} \|y^m - x\|_2 &\leq \eta_m \operatorname{cond}_2(T) \|x^0 - x\|_2 \quad \text{with} \\ \eta_m &= 2 \left(1 - \frac{1}{\kappa}\right)^m / \left[\left(1 + \frac{1}{\sqrt{\kappa}}\right)^{2m} + \left(1 - \frac{1}{\sqrt{\kappa}}\right)^{2m} \right], \end{aligned} \quad (8.29c)$$

where κ is defined by (8.28c). In the case of a symmetric iteration applied to $A > 0$ (cf. §3.5.2), an estimate analogous to (8.29c) holds with respect to the energy norm:

$$\|y^m - x\|_A \leq \eta_m \|x^0 - x\|_A.$$

Proof of (d). Use $c = (1 - 1/\kappa) / (1 + 1/\sqrt{\kappa})^2$. \square

For the implementation of the Chebyshev method, one could in principle apply Remark 8.16 and use the second formulation. The Chebyshev polynomial T_m has the zeros

$$x_\nu = \cos\left([\nu + \frac{1}{2}]\pi/m\right) \quad (1 \leq \nu \leq m).$$

Hence, the transformed polynomial p_m in (8.27a) admits the factorisation (8.13a) with $\zeta_\nu = \frac{1}{2}[a + b + (b - a)x_\nu]$. The auxiliary polynomials p_μ in (8.13b) lead to the damping factors $\vartheta_\mu := 1/(1 - \zeta_\mu)$ in (8.10b) and (8.12b). However, this approach suffers from numerical instabilities (cf. Lebedev–Finogenov [261, 262]).

The only elegant and practical implementation is the use of the *three-term recursion* (8.14a–c), since recursion (8.28a,b) is a particular case of (8.15a,b). The coefficients Θ_m and ϑ_m required in (8.14a–c) are provided by the next exercise.

Exercise 8.28. Prove: (a) For the case of $\sigma_M = [a, b]$ with $a < b < 1$, recursion (8.15a,b) for p_m in (8.28a,b) uses the factors

$$\Theta_m = 4C_{m-1}/[(b - a)C_m], \quad (8.30a)$$

$$\vartheta_m = -2(a + b)C_{m-1}/[(b - a)C_m]. \quad (8.30b)$$

(b) In the case of $\sigma_M = [-\rho, \rho]$ with $\rho > 0$, (8.28b) leads to recursion (8.15c,d) with

$$\Theta_m = 2C_{m-1}/(\rho C_m) = 1 + C_{m-2}/C_m.$$

(c) Which coefficients correspond to the case of $\sigma_M = [0, \rho]$?

(d) Use Eq. (8.28b) at $\zeta = 1$: $C_{m+1} = AC_m - C_{m-1}$ with $A := 2(2-a-b)/(b-a)$ and prove for the general case of $\sigma_M = [a, b]$ that

$$\Theta_m = \frac{16}{8(2-a-b) - (b-a)^2\Theta_{m-1}}, \quad \Theta_1 = \frac{4}{2-a-b}, \quad (8.30c)$$

$$\vartheta_m = -\frac{1}{2}(a+b)\Theta_m. \quad (8.30d)$$

(e) The coefficients converge monotonically to

$$\lim \Theta_m = \frac{4c}{b-a} \quad \text{and} \quad \lim \vartheta_m = \frac{-2c(a+b)}{b-a}$$

with c in (8.28c).

(f) The assumptions $a < b$ in (a) and $\rho > 0$ in (b) avoid the division by zero. Show that $a = b$ or $\rho = 0$ lead to a direct solution: the semi-iterate y^1 is already the exact solution.

Hint for (a): For $m > 2$, compare the coefficients in (8.15b) and (8.28b). For $m = 1$, compare (8.15a) with (8.28a), taking notice of $C_0 = 1$ and $C_1 = \frac{2-a-b}{b-a}$ according to (8.28c). Part (e): Insert (8.28c) into (8.30a,b).

Instead of Θ_m and ϑ_m , one can also compute the sum $\sigma_m := \Theta_m + \vartheta_m$ recursively from

$$\sigma_1 = 2, \quad \sigma_m = 4 / \left\{ 4 - \left(\frac{1 - 1/\kappa}{1 + 1/\kappa} \right)^2 \sigma_{m-1} \right\}$$

(derived from (8.30c,d) with κ in (8.26d)). Equation (8.30d) yields the values

$$\Theta_m = 2\sigma_m/(2-a-b), \quad \vartheta_m = -(a+b)\sigma_m/(2-a-b).$$

The coefficients σ_m can also be used directly for the three-term recursion. Given the matrix N of the second normal form of Φ , the formulae (8.14a-c) with the coefficients (8.30a,b) are equivalent to

$$\begin{aligned} y^0 &= x^0, & y^1 &= y^0 - \frac{2}{2-a-b} N(Ay^0 - b), \\ y^m &= \sigma_m \left\{ y^{m-1} - \frac{2}{2-a-b} N(Ay^{m-1} - b) \right\} + (1 - \sigma_m)y^{m-2}. \end{aligned} \quad (8.31)$$

The factor $\frac{2}{2-a-b}$ may also be written as $\frac{2}{\gamma+I}$ (cf. (8.26c)).

We recall the set \mathcal{N} of nonlinear acceleration methods mentioned on page 173. The Chebyshev method is a first example.

Notation 8.29. We denote the Chebyshev method based on $\sigma_M = [a, b]$ by

$$\Upsilon_{a,b}^{\text{Cheb}} \in \mathcal{N}.$$

In principle, the Chebyshev method is well defined for all iterations $\Phi \in \mathcal{L}$. However, the convergence statements only refer to algorithm $\Upsilon_{a,b}^{\text{Cheb}}[\Phi]$ and matrices A such that $\sigma(M) \subset [a, b]$ holds for the iteration matrix $M = M_\Phi[A]$ of Φ .

8.3.5 Order Improvement by the Chebyshev Method

Theorem 8.30. (a) Assume that $\sigma(M) \subset \sigma_M = [a, b]$ holds with $a < b < 1$. The Chebyshev method has the asymptotic convergence rate $c = \lim_{m \rightarrow \infty} \sqrt[m]{\frac{1}{C_m}}$ with

$$c = \frac{b - a}{2 - a - b + 2\sqrt{(1-a)(1-b)}} = \frac{\Gamma - \gamma}{(\sqrt{\Gamma} + \sqrt{\gamma})^2} = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}, \quad (8.32a)$$

where $\kappa = \Gamma/\gamma$ (cf. (8.26c)). Particular cases are

$$\lim_{m \rightarrow \infty} \sqrt[m]{\frac{1}{C_m}} = \frac{\rho}{1 + \sqrt{1 - \rho^2}} \quad \text{for } \sigma_M = [-\rho, \rho], \rho < 1, \quad (8.32b)$$

$$\lim_{m \rightarrow \infty} \sqrt[m]{\frac{1}{C_m}} = \frac{\rho}{(1 + \sqrt{1 - \rho})^2} \quad \text{for } \sigma_M = [0, \rho], \rho < 1. \quad (8.32c)$$

(b) Let τ be the order of the basic iteration: $\rho(M) = 1 - Ch^\tau + \mathcal{O}(h^{2\tau})$. Then the Chebyshev method is of order $\tau/2$. The asymptotic convergence rate equals

$$\begin{aligned} 1 - 2\sqrt{\frac{C}{1-\gamma}} h^{\tau/2} + \mathcal{O}(h^\tau) &\quad \text{for (8.32a) with } \Gamma = \rho(M), \\ 1 - \sqrt{2C} h^{\tau/2} + \mathcal{O}(h^\tau) &\quad \text{for } \sigma_M = [-\rho(M), \rho(M)], \\ 1 - 2\sqrt{C} h^{\tau/2} + \mathcal{O}(h^\tau) &\quad \text{for } \sigma_M = [0, \rho(M)]. \end{aligned}$$

Proof. Since $0 \leq c \leq 1$, (8.28c) shows that $\sqrt[m]{\frac{1}{C_m}} = c \sqrt[m]{2/(1+c^{2m})} \rightarrow c$. \square

Therefore, the Chebyshev method achieves a halving of the order similar to the SOR iteration. Concerning the connection of both methods, we refer to §8.4.3 and Varga [375, §5.2].

8.3.6 Optimisation Over Other Sets

Up to now, we considered an interval $[a, b]$ with $a < b < 1$. If, for instance, no eigenvalue of M lies in $(a', b') \subset [a, b]$, we may replace σ_M by the smaller set

$$\sigma_M = [a, a'] \cup [b', b] \quad (a \leq a' < b' \leq b). \quad (8.33)$$

Obviously, the minimum of $\{\max_{\zeta \in \sigma_M} |p_m(\zeta)| : p_m \in \mathcal{P}_m\}$ can only become smaller. In the case of $a' - a = b - b'$, it is easy to describe the optimal polynomial (cf. Axelsson–Barker [13, p. 26f]). Concerning the determination of optimal polynomials, we refer to de Boor–Rice [102] and Fischer [133, §3.3]. In particular, the case $\sigma_M = [a, a'] \cup [b', b]$ for $a \leq a' < 1 < b' \leq b$ is interesting. The latter situation occurs for indefinite matrices.

Remark 8.31. Consider discretisation of Helmholtz' equation $-\Delta u - cu = f$ with positive c , which leads to $A = A_\Delta - cI$, where A_Δ is the matrix of the Poisson model problem. Let $0 < \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$ be the eigenvalues of A_Δ . Assume for a suitable k that $\lambda_k < c < \lambda_{k+1}$. Then the spectrum of $A = A^H$ is contained in

$$\sigma_A = [-\beta_-, -\alpha_-] \cup [\alpha_+, \beta_+] \quad \text{with} \quad -\beta_- \leq -\alpha_- < 0 < \alpha_+ < \beta_+,$$

where $\beta_- := c - \lambda_1$, $\alpha_- := c - \lambda_k$, $\alpha_+ := \lambda_{k+1} - c$, $\beta_+ := \lambda_n - c$. The Richardson iteration with $0 < \Theta < 1/\beta_+$ leads to the iteration matrix $M = M_\Theta^{\text{Rich}}$ whose spectrum is contained in the set σ_M described in (8.33), where

$$a = 1 - \Theta\beta_+ < a' = 1 - \Theta\alpha_+ < 1 < b' = 1 + \Theta\alpha_- \leq b = 1 + \Theta\beta_-.$$

If one extreme eigenvalue b of M is known and the others are enclosed by $[a, b']$, we arrive at

$$\sigma_M = [a, b'] \cup \{b\} \quad \text{with} \quad b' < b, \quad b' < 1, \quad b \neq 1.$$

Let $q_{m-1} \in \mathcal{P}_{m-1}$ with $q_{m-1}(1) = 1$ be optimal for $[a, b']$. A simple but not optimal proposal for a polynomial p_m suited to σ_M is

$$p_m(\zeta) := q_{m-1}(\zeta)(\zeta - b)/(1 - b).$$

Concerning the construction of asymptotically optimal polynomials for arbitrary compact sets σ_M with $1 \notin \sigma_M$, we refer to Niethammer–Varga [294] and Eiermann–Niethammer–Varga [119]. The simplest set σ_M that is more general than the interval $[a, b]$ is the ellipse (cf. Fischer–Freund [134, 135], Niethammer–Varga [294], and Manteuffel [272]). Since, in general, a suitable ellipse enclosing the eigenvalues of M is not known a priori, one has to improve its parameters adaptively (cf. Manteuffel [272]). The fact that the ellipse lies in the complex plane does not imply that the optimal polynomial has also complex parameters. As long as $\sigma(M)$ is symmetric with respect to the real axis (i.e., all complex eigenvalues belong to conjugate pairs), one can find an optimal polynomial with real coefficients (cf. Opfer–Schober [297]).

In any case, the spectrum $\sigma(M)$ is enclosed by the complex circle

$$\sigma_M = \{z = x + iy \in \mathbb{C} : x^2 + y^2 \leq \rho(M)^2\}.$$

Unfortunately, this choice does not lead to an interesting solution (cf. [297]).

Theorem 8.32. Let σ_M be a circle around $z_0 \in \mathbb{C} \setminus \{1\}$ with radius $r < |1 - z_0|$. The optimal polynomial for σ_M is $p_m(\zeta) = [(\zeta - z_0)/(1 - z_0)]^m$. In particular, for $z_0 = 0$, the corresponding semi-iteration coincides with the basic iteration Φ . In the general case, the semi-iteration corresponds to the damped iteration Φ_ϑ with $\vartheta := 1/|1 - z_0|$.

Proof. The absolute value of the polynomial p_m defined above takes its maximum

$$\rho := \max\{|p_m(\zeta)| : \zeta \in \sigma_M\} = r/|1 - z_0|$$

at all boundary points $\zeta \in \partial\sigma_M$. If p_m is not optimal, there is some polynomial $q_m \in \mathcal{P}_m$ with $q_m(1) = 1$ and $\max\{|q_m(\zeta)| : \zeta \in \sigma_M\} < \rho$. $q_m(\zeta) < \rho = p_m(\zeta)$ holds for all boundary values $\zeta \in \partial\sigma_M$, so that the theorem of Rouché is applicable; i.e., the holomorphic functions p_m and $p_m - q_m$ have the same number of zeros in σ_M . Since p_m has an m -fold zero at z_0 , $p_m - q_m$ has also m zeros in σ_M . Since $(p_m - q_m)(1) = p_m(1) - q_m(1) = 1 - 1 = 0$, the polynomial $p_m - q_m \in \mathcal{P}_m$ has even $m + 1$ zeros, implying $p_m = q_m$. Hence, p_m is already optimal. \square

8.3.7 Cyclic Iteration

Following Conclusion 8.27, it has been mentioned that in principle it would be possible to apply the second formulation (8.10b) with the factors $\vartheta_\nu := 1/(1 - \zeta_\nu)$, $\zeta_\nu = \cos([\nu + \frac{1}{2}]\pi/m)$ for $\nu = 1, \dots, m$. The result y^m (only for this fixed m) is the desired Chebyshev semi-iterate. However, by this approach the Chebyshev method cannot be continued. To obtain an infinite iterative process, we may repeat the extrapolation factors m -periodically:

$$\begin{aligned} \vartheta_1, \vartheta_2, \dots, \vartheta_m &\quad \text{given,} \\ \vartheta_i &:= \vartheta_{i-m} \quad \text{for } i > m. \end{aligned}$$

A semi-iterative method (8.10a,b) with these parameters is called a *cyclic iteration*. The restriction to the iterates $y^0, y^m, y^{2m}, y^{3m}, \dots$ produces a proper linear iteration. The related iteration matrix is $p_m(M)$ with p_m generated by $\{\vartheta_i : 1 \leq i \leq m\}$. The convergence rate of the cyclic iteration is not described by $\rho(p_m(M))$ but by $\rho(p_m(M))^{1/m}$, since one cycle $y^0 \mapsto y^m$ is thought to consist of m and not of one step. The cyclic iteration also runs the risk of numerical instabilities as already discussed after Conclusion 8.27.

Exercise 8.33. Prove: Viewing the cyclic iteration as a semi-iteration $\{y^0, y^1, \dots\}$ of all iterates, the asymptotic convergence rate in Definition 8.3 also coincides with $\sqrt[m]{\rho(p_m(M))}$.

8.3.8 Two- and Multi-Step Iterations

Exercise 8.28e yields the limits $\Theta = \lim \Theta_m$ and $\vartheta = \lim \vartheta_m$. Hence, the three-term recursion (8.14c) converges to the (stationary) two-step iteration (2.27):

$$y^m = \Theta [\Phi(y^{m-1}, b) - y^{m-2}] + \vartheta (y^{m-1} - y^{m-2}) + y^{m-2}. \quad (8.35)$$

As described in §2.2.8, the convergence of iteration (2.27) can be reduced to the convergence of a one-step iteration with the iteration matrix

$$\mathbf{M} = \begin{bmatrix} \mu_0 M + \mu_1 I & \mu_2 I \\ I & 0 \end{bmatrix}, \quad \mu_0 = \frac{4c}{b-a}, \quad \mu_1 = -2c \frac{a+b}{b-a}, \quad \mu_2 = 1 - \mu_0 - \mu_1$$

(c defined in (8.28c)). From these coefficients, assuming that $\sigma(M) \subset \sigma_M$ and using Exercise 2.25, we obtain the value $\rho(\mathbf{M}) = c$, i.e., the (stationary) two-step iteration (8.35) achieves the same convergence rate as the semi-iterative method. Hence, the two-step iteration (8.35) also yield an improvement of the order of convergence.

More generally, one can consider the k -step iteration

$$x^m = \mu_0 \Phi(x^m, b) + \sum_{i=1}^k \mu_i x^{m-i} \quad \text{with } \sum_{i=1}^k \mu_i = 1.$$

The connection between k -step iterations and semi-iterative methods is described by Niethammer–Varga [294].

8.3.9 Amount of Work of the Semi-Iterative Method

We consider the realisation of the Chebyshev method by (8.31). There the call of the basic iteration $\Phi(x, b) = x - W^{-1}(Ax - b)$ is replaced by the call of $W^{-1}(b - Ay) = \Phi(x, b) - x$. Besides the call of the basic iteration Φ , the implementation (8.31) (for $m \geq 2$) requires six operations per grid point:

$$\text{semi-iterative Work}(\Phi) \leq \text{Work}(\Phi) + 6n$$

(cf. §2.3 and §3.4). Hence, the cost factor amounts to

$$C_{\Phi, \text{semi}} = C_\Phi + \frac{6}{C_A},$$

where C_{An} is defined in §2.3 as the number of nonzero elements of A .

Replacing in (2.31a) the convergence rate by the asymptotic value c in (8.32a), we obtain the effective amount of work

$$\text{Eff}_{\text{semi}}(\Phi) = -(C_\Phi + \frac{6}{C_A}) / \log c.$$

If $\gamma/\Gamma \ll 1$ holds as in the examples discussed in §8.4, we can exploit the asymptotic behaviour $\log c = -2\sqrt{\gamma/\Gamma} + \mathcal{O}(\gamma/\Gamma)$:

$$\text{Eff}_{\text{semi}}(\Phi) \approx \left(\frac{C_\Phi}{2} + \frac{3}{C_A} \right) \sqrt{\frac{\Gamma}{\gamma}}. \quad (8.36)$$

Exercise 8.34. Assume that the iteration matrix of Φ fulfils $\sigma(M) \subset [a, b]$ with $b = 1 - \mathcal{O}(h^{-\tau})$ and $\tau > 0$. Prove the following comparison of Eff_{semi} and Eff :

$$\text{Eff}_{\text{semi}}(\Phi) \approx \left(C_\Phi + \frac{6}{C_A} \right) \sqrt{\text{Eff}(\Phi) / [(1-a) C_\Phi]}.$$

8.4 Application to Iterations Discussed Above

8.4.1 Preliminaries

The essential condition for the applicability² of the Chebyshev method is that the spectrum $\sigma(M)$ be real. This excludes the SOR method. Semi-iterative variants based on other supersets $\sigma_M \supset \sigma(M)$ are also not successful for the SOR method with $\omega \geq \omega_{\text{opt}}$ (cf. §8.3.6)). The reason for this is statement (e) of Theorem 4.27. For $\omega \geq \omega_{\text{opt}}$, all eigenvalues $\lambda \in \sigma(M_\omega^{\text{SOR}})$ are situated on the boundary of the complex circle $|\zeta| = \omega - 1$, for which no convergence acceleration is possible, as stated in Theorem 8.32.

If A is positive definite, the following already mentioned iterations lead to a real spectrum: the Richardson, (block-)Jacobi, and (block-)SSOR methods. Numerical results for these choices of basic iterations will be presented for the Poisson model problem in the following sections.

Besides the iterations mentioned above, in §5.2 we constructed their damped variants. However, for a discussion of semi-iterative methods the damped variants are without any interest as stated next.

Lemma 8.35. *Let the iteration Φ have a real spectrum $\sigma(M)$. Then Φ and the corresponding damped iterations Φ_ϑ with $\vartheta \neq 0$ generate identical semi-iterative results y^m .*

Proof. By (8.6a,b), the semi-iterate y^m generated by Φ has the representation $y^m = x^0 + p_m(M)(x^0 - x)$. The damped iteration has the iteration matrix

$$M_\vartheta = I - \vartheta N A = I - N_\vartheta A \quad \text{with } N_\vartheta := \vartheta N.$$

For N_ϑ , inclusion (8.26b) can be written as $\sigma(N_\vartheta A) \subset [\gamma', \Gamma']$ with $\gamma' := \vartheta \gamma$ and $\Gamma' := \vartheta \Gamma$ (possibly a complex interval). $p_m(M) = T_m \left(\frac{\Gamma+\gamma}{\Gamma-\gamma} I + \frac{2}{\Gamma-\gamma} N A \right)$ (cf. Exercise 8.26b) is invariant with respect to the replacement of γ , Γ , N by γ' , Γ' , N_ϑ . Hence $p_{m,\vartheta}(M_\vartheta) = p_m(M)$, where $p_{m,\vartheta}$ is the polynomial adapted to the interval $[\gamma', \Gamma'] \supset \sigma(N_\vartheta A)$. The iterates $y_\vartheta^m = x^0 + p_{m,\vartheta}(M_\vartheta)(x^0 - x)$ of Φ_ϑ coincide with those of Φ . \square

² Here ‘applicability of the Chebyshev method’ means that also the assumptions of the convergence statements hold. Otherwise, the Chebyshev method can be applied to any $A \in \mathfrak{D}(\Phi)$.

8.4.2 Semi-Iterative Richardson Method

According to Lemma 8.35, we may fix the factor of Richardson's method (3.4) by $\Theta = 1$, i.e., $x^{m+1} = x^m - (Ax - b)$. Then the matrix $N = I$ of the second normal form is as simple as possible and condition (8.26b) becomes $\sigma(A) \subset [\gamma, \Gamma]$.

Remark 8.36. (a) The Chebyshev method is applicable if A has only positive eigenvalues. For the estimation of γ and Γ in (8.26b), one has to use the respective bounds for the extreme eigenvalues of A .

(b) In particular, the assumptions are satisfied if A is positive definite. In this case, one has to choose $\gamma = 1/\|A^{-1}\|_2$ and $\Gamma = \|A\|_2$ (optimal choice) or at least $\gamma \leq 1/\|A^{-1}\|_2$ and $\Gamma \geq \|A\|_2$.

For the Poisson model problem, we obtain

$$\gamma = \lambda_{\min} = 8h^{-2} \sin^2(\pi h/2), \quad \Gamma = \lambda_{\max} = 8h^{-2} \cos^2(\pi h/2)$$

according to (3.1b,c). Inserting these values into the asymptotic convergence rate (8.32a), we arrive at

$$\lim_{m \rightarrow \infty} \sqrt[m]{\frac{1}{C_m}} = c = \cos(\pi h)/(1 + \sin(\pi h)) = 1 - \pi h + \mathcal{O}(h^2).$$

For $h = 1/16$ and $h = 1/32$, we obtain $c = 0.82$ and $c = 0.906$. The numerical results in Table 8.1 show that the reduction factor approximates the convergence rate only for sufficiently large m . The ratios

$$\rho_m := \|y^m - x\|_2 / \|y^{m-1} - x\|_2, \quad \hat{\rho}_m := (\|y^m - x\|_2 / \|y^0 - x\|_2)^{1/m}$$

tend to c from above.

m	$\ y^m - x\ _2$	ρ_m	$\hat{\rho}_m$
1	$6.44_{10^{-1}}$	$9.09_{10^{-1}}$	$9.09_{10^{-1}}$
10	$2.44_{10^{-1}}$	$8.91_{10^{-1}}$	$8.99_{10^{-1}}$
20	$6.35_{10^{-2}}$	$8.59_{10^{-1}}$	$8.86_{10^{-1}}$
30	$1.29_{10^{-2}}$	$8.48_{10^{-1}}$	$8.75_{10^{-1}}$
40	$2.36_{10^{-3}}$	$8.41_{10^{-1}}$	$8.67_{10^{-1}}$
50	$4.07_{10^{-4}}$	$8.36_{10^{-1}}$	$8.61_{10^{-1}}$
60	$6.75_{10^{-5}}$	$8.34_{10^{-1}}$	$8.57_{10^{-1}}$
70	$1.08_{10^{-5}}$	$8.32_{10^{-1}}$	$8.53_{10^{-1}}$
80	$1.72_{10^{-6}}$	$8.31_{10^{-1}}$	$8.50_{10^{-1}}$
90	$2.67_{10^{-7}}$	$8.29_{10^{-1}}$	$8.48_{10^{-1}}$
100	$4.11_{10^{-8}}$	$8.28_{10^{-1}}$	$8.46_{10^{-1}}$

m	$\ y^m - x\ _2$	ρ_m	$\hat{\rho}_m$
1	$7.14_{10^{-1}}$	$9.54_{10^{-1}}$	$9.54_{10^{-1}}$
10	$4.47_{10^{-1}}$	$9.48_{10^{-1}}$	$9.49_{10^{-1}}$
30	$1.40_{10^{-1}}$	$9.36_{10^{-1}}$	$9.45_{10^{-1}}$
50	$3.21_{10^{-2}}$	$9.24_{10^{-1}}$	$9.38_{10^{-1}}$
70	$6.26_{10^{-3}}$	$9.19_{10^{-1}}$	$9.33_{10^{-1}}$
80	$2.66_{10^{-3}}$	$9.17_{10^{-1}}$	$9.31_{10^{-1}}$
100	$4.65_{10^{-4}}$	$9.15_{10^{-1}}$	$9.28_{10^{-1}}$
120	$7.80_{10^{-5}}$	$9.13_{10^{-1}}$	$9.26_{10^{-1}}$
130	$3.15_{10^{-5}}$	$9.13_{10^{-1}}$	$9.25_{10^{-1}}$
140	$1.27_{10^{-5}}$	$9.12_{10^{-1}}$	$9.24_{10^{-1}}$
150	$5.09_{10^{-6}}$	$9.12_{10^{-1}}$	$9.23_{10^{-1}}$

Table 8.1 Semi-iterative Richardson method for $h = 1/16$ (left) und $h = 1/32$ (right).

8.4.3 Semi-Iterative Jacobi and Block-Jacobi Method

Numerical examples are unnecessary, since in the Poisson model case, the Jacobi method coincides with the damped Richardson method and, according to Lemma 8.35, reproduces the results in Table 8.1.

Concerning the lower bound a of the spectrum $\sigma(M^{\text{Jac}})$, Lemma 4.8 proves that $a = -b$ holds for a particular case.

Lemma 8.37. *If (A, D) is weakly 2-cyclic (cf. Definition 4.2), the Jacobi iteration matrix M^{Jac} has a symmetric spectrum: $\sigma(M^{\text{Jac}}) = -\sigma(M^{\text{Jac}})$. The smallest enclosing interval is $[a, b] = [-\rho(M^{\text{Jac}}), \rho(M^{\text{Jac}})]$.*

A comparison of the semi-iterative Jacobi iteration with the SOR method is possible. In the weakly 2-cyclic case, (8.32b) is applicable because of Lemma 8.37 and yields the asymptotic semi-iterative convergence rate

$$\beta/[1 + \sqrt{1 - \beta^2}] \quad \text{with } \beta := \rho(M^{\text{Jac}}).$$

This quantity coincides with the square root of the optimal SOR convergence rate $\omega_{\text{opt}} - 1$; hence, the semi-iterative Jacobi iteration is half as fast as the SOR method. The order improvement by an optimal choice ω_{opt} in the SOR case and the order improvement by the Chebyshev method (cf. Theorem 8.30b) lead to very similar results.

m	$\ y^m - x\ _2$	ρ_m	$\hat{\rho}_m$
1	$6.09_{10^{-1}}$	$8.60_{10^{-1}}$	$8.60_{10^{-1}}$
20	$1.62_{10^{-2}}$	$7.95_{10^{-1}}$	$8.27_{10^{-1}}$
40	$1.19_{10^{-4}}$	$7.75_{10^{-1}}$	$8.04_{10^{-1}}$
60	$6.68_{10^{-7}}$	$7.69_{10^{-1}}$	$7.93_{10^{-1}}$
80	$3.33_{10^{-9}}$	$7.65_{10^{-1}}$	$7.86_{10^{-1}}$
90	$2.1_{10^{-10}}$	$7.55_{10^{-1}}$	$7.84_{10^{-1}}$

m	$\ y^m - x\ _2$	ρ_m	$\hat{\rho}_m$
1	$6.94_{10^{-1}}$	$9.28_{10^{-1}}$	$9.28_{10^{-1}}$
20	$1.53_{10^{-1}}$	$9.12_{10^{-1}}$	$9.23_{10^{-1}}$
40	$1.84_{10^{-2}}$	$8.92_{10^{-1}}$	$9.11_{10^{-1}}$
60	$1.70_{10^{-3}}$	$8.85_{10^{-1}}$	$9.03_{10^{-1}}$
80	$1.40_{10^{-4}}$	$8.81_{10^{-1}}$	$8.98_{10^{-1}}$
100	$1.08_{10^{-5}}$	$8.78_{10^{-1}}$	$8.94_{10^{-1}}$

Table 8.2 Semi-iterative column-block-Jacobi iteration for $h = 1/16$ (left) and $h = 1/32$ (right).

The block variants of the Jacobi iteration converge faster than the pointwise version. Correspondingly, the results of the semi-iterative column-block-Jacobi method in Table 8.2 are better than those in Table 8.1. The factors should tend to the asymptotic value 0.7565 for $h = 1/16$ and to 0.8702 for $h = 1/32$.

8.4.4 Semi-Iterative SSOR and Block-SSOR Iteration

As already mentioned in §8.4.1, the Gauss–Seidel and SOR methods are not suited for semi-iterative purposes, since, in general, the spectrum is not real. A remedy is offered by the symmetric Gauss–Seidel and SSOR iteration. Theorem 6.26 states that the spectrum of the SSOR method is real for Hermitian matrices A .

Theorem 6.28 gives an upper bound for the spectral radius $\rho(M_\omega^{\text{SSOR}})$. Hence, under conditions (6.18a,b), the spectrum can be enclosed by the interval $[a, b]$ with

$$a = 0, \quad b = 1 - 2\Omega / \left[\frac{\Omega^2}{\gamma} + \Omega + \frac{\Gamma}{4} \right], \quad \text{where } \Omega := \frac{2 - \omega}{2\omega}, \quad 0 < \omega < 2. \quad (8.37)$$

Here, Γ is defined by (6.18b). Corollary 3.45 helps to determine Γ . For the Poisson model problem, Lemma 3.62 yields the value $\Gamma = 2$. Inequality (6.18a) states that γ coincides with λ in (3.35c) applied to the (block-)Jacobi method. In the Poisson model case, $\gamma = 2 \sin^2(\pi h/2)$ holds.

Theorem 8.38. *Let $A = D - E - E^H > 0$ and γ, Γ satisfy the assumptions (6.18a,b). Assume, in addition, that $0 < \omega \leq 2/(\Gamma + 1)$. Then*

$$a = \left(\frac{1 - \xi}{1 + \xi} \right)^2 \quad \text{with} \quad \xi := \frac{2 - \omega}{\Gamma \omega} \quad (8.38)$$

is a lower bound of the spectrum $\sigma(M_\omega^{\text{SSOR}})$.

Proof. Using the parameter Ω in (3.46c), we can rewrite

$$W_\omega^{\text{SSOR}} = \left(\frac{1}{\omega} D - E \right) \left[\left(\frac{2}{\omega} - 1 \right) D \right]^{-1} \left(\frac{1}{\omega} D - E \right)^H$$

as

$$W_\omega^{\text{SSOR}} = [\Omega D + \Delta](2\Omega D)^{-1}[\Omega D + \Delta]^H \quad \text{with} \quad \Delta := \frac{1}{2}D - E.$$

Defining $X := \Omega D + (1 - \alpha)\Delta$ for some real α , we have $[\Omega D + \Delta] = X + \alpha\Delta$. The expansion of $[X + \alpha\Delta](2\Omega D)^{-1}[X + \alpha\Delta]^H$ yields

$$W_\omega^{\text{SSOR}} = \frac{1}{2\Omega} X D^{-1} X^H + \frac{\alpha}{2} A + \frac{1}{2\Omega} (2\alpha - \alpha^2) \Delta D^{-1} \Delta^H.$$

because of $\Delta + \Delta^H = A$. The factor $(2\alpha - \alpha^2)$ is negative for $\alpha > 2$. Hence,

$$W_\omega^{\text{SSOR}} \geq g(\alpha)A \quad \text{with} \quad g(\alpha) := \frac{\alpha}{2} \left(1 + \frac{\Gamma}{4} \frac{2 - \alpha}{\Omega} \right) \quad \text{for } \alpha \geq 2.$$

The assumption $\omega \leq 2/(\Gamma + 1)$ implies $\alpha_0 := 1 + 2\Omega/\Gamma \geq 2$. Theorem 3.34a with $1 - a = 1/g(\alpha_0)$ yields the value (8.38). \square

The statement is less interesting, since (because of $\Gamma = 2$ for the Poisson model case) Theorem 8.38 only applies to strong underrelaxation: $\omega \leq 2/3$.

There are two possibilities in improving (halving) the convergence order. First, this can be achieved by the optimal choice of ω in the SOR or SSOR method (cf. Conclusions 3.46 and 6.29). Second, the semi-iterative method leads to halving of the order compared with the basic iteration. In the case of SSOR as the basic iteration, both techniques can be applied simultaneously. First, the optimal SSOR relaxation parameter ω' is chosen as described in (3.47b). The hereby defined iteration $\Phi_{\omega'}^{\text{SSOR}}$ is chosen as the basic iteration of the Chebyshev method. Together, we succeed in quartering the order. In the Poisson model case, we obtain the asymptotic convergence rate $1 - \mathcal{O}(h^{1/2})$.

The bound b in (8.37) becomes minimal for

$$\omega' = 2/(1 + \sqrt{\gamma/\Gamma}).$$

The corresponding value is

$$b = \frac{\sqrt{\Gamma} - \sqrt{\gamma}}{\sqrt{\Gamma} + \sqrt{\gamma}} = \frac{1 - \sqrt{\gamma/\Gamma}}{1 + \sqrt{\gamma/\Gamma}}. \quad (8.39a)$$

Inserting this value into (8.32c) yields the asymptotic convergence rate

$$\lim_{m \rightarrow \infty} \sqrt[m]{\frac{1}{C_m}} = c = \frac{1 - \sqrt{1 - b}}{1 + \sqrt{1 - b}} \quad \text{with } b \text{ in (8.39a).} \quad (8.39b)$$

The spectral condition number $\kappa = \kappa((W^{\text{SSOR}})^{-1}A)$ is equal to $\frac{1}{2}(1 + \sqrt{\gamma/\Gamma})$. Using the inequality $\gamma \geq 1/\kappa(A)$ in Exercise 5.20, we end up with the result

$$\kappa((W^{\text{SSOR}})^{-1}A) \leq \frac{1}{2} \left(1 + \sqrt{\Gamma \kappa(A)} \right). \quad (8.39c)$$

m	$\ y^m - x\ _2$	ρ_m	$\hat{\rho}_m$	N	ω'	c
1	$4.673_{10^{-1}}$	$6.24_{10^{-1}}$	$6.24_{10^{-1}}$	2	0.8284	0.0470
2	$2.761_{10^{-1}}$	$5.90_{10^{-1}}$	$6.07_{10^{-1}}$	4	1.1329	0.1467
3	$1.359_{10^{-1}}$	$4.92_{10^{-1}}$	$5.66_{10^{-1}}$	8	1.4386	0.2727
4	$7.681_{10^{-2}}$	$5.65_{10^{-1}}$	$5.66_{10^{-1}}$	16	1.6721	0.4059
5	$3.801_{10^{-2}}$	$4.94_{10^{-1}}$	$5.51_{10^{-1}}$	32	1.8212	0.5315
				64	1.9064	0.6408
20	$2.080_{10^{-6}}$	$5.08_{10^{-1}}$	$5.27_{10^{-1}}$	128	1.9520	0.7305
21	$1.007_{10^{-6}}$	$4.84_{10^{-1}}$	$5.25_{10^{-1}}$	256	1.9757	0.8010
22	$5.195_{10^{-7}}$	$5.15_{10^{-1}}$	$5.24_{10^{-1}}$	512	1.9878	0.8549
23	$2.541_{10^{-7}}$	$4.89_{10^{-1}}$	$5.23_{10^{-1}}$	1028	1.9939	0.8953
				5000	1.9987	0.9511
29	$3.395_{10^{-9}}$	$4.82_{10^{-1}}$	$5.15_{10^{-1}}$	10000	1.9993	0.9651
30	$1.628_{10^{-9}}$	$4.79_{10^{-1}}$	$5.14_{10^{-1}}$			

Table 8.3 Left: semi-iterative lexicographical SSOR for the parameters in (8.40); concerning ρ_m and $\hat{\rho}_m$ see Table 8.1. Right: optimal ω' and asymptotic rate c for $h = 1/N$.

For the values γ and Γ in Lemma 3.62 (Poisson model case), the convergence rate (8.39b) is asymptotically equal to the value

$$c = 1 - Ch^{1/2} + \mathcal{O}(h) \quad \text{with } C = 2\sqrt{\pi}.$$

The results in Table 8.3 refer to the parameters

$$h = 1/32, \quad \omega = 1.8455, \quad a = 0, \quad b = 0.878. \quad (8.40)$$

In §6.3.5 the value ω is proved to be optimal (note that ω' is optimal only for the bound in (6.18c)). We learn from Table 6.1 that $b = 0.878$ is an upper bound of the convergence rate. From (8.39b) with $b = 0.878$, one calculates the rate $c = 0.482$, which is numerically well confirmed (cf. Table 8.3). From $C_{\Phi}^{\text{SSOR}} = 2 + 6/C_A = 3.2$ (according to Remark 6.27 and because of $C_A = 5$ for five-point formulae), we obtain the effective amount of work

$$\text{Eff}_{\text{semi}}(\Phi^{\text{SSOR}}) = -3.2 / \log c = 4.38 \quad (8.41)$$

for the semi-iterative SSOR method with $h = 1/32$, which can be compared, e.g., with $\text{Eff}(\Phi^{\text{SOR}}) = 7.05$ in Example 2.28.

If we use the values ω' in (3.47b), Eq. (8.39b) yields the asymptotic convergence rates c reported in Table 8.3. These values give an impression of the asymptotic value $c = 1 - \mathcal{O}(h^{1/2})$.

8.5 Method of Alternating Directions (ADI)

The *alternating-direction-implicit iteration* or shortly *ADI method* was first described in 1955 by Peaceman–Rachford [308] in connection with parabolic differential equations. ADI is not a semi-iterative method in the sense of the previous sections, but it can be considered as a generalisation using rational functions instead of polynomials (see also §8.5.4).

Further material can be found in Marchuk [274] and Wachspress [383].

8.5.1 Application to the Model Problem

For the model problem in §1.2, the matrix A can be split into

$$A = B + C, \quad \text{where} \quad (8.42a)$$

$$(Bu)(x, y) = h^{-2} [-u(x-h, y) + 2u(x, y) - u(x+h, y)], \quad (8.42b)$$

$$(Cu)(x, y) = h^{-2} [-u(x, y-h) + 2u(x, y) - u(x, y+h)] \quad (8.42c)$$

for $(x, y) \in \Omega_h$ are the second differences of u with respect to the x and y direction. If we choose the rows (x direction) of Ω_h as blocks, $B + 2h^{-2}I$ represents the block diagonal of A . Similarly, $C + 2h^{-2}I$ is the block diagonal of A if the columns (y direction) are chosen as blocks.

Remark 8.39. For A , B , and C in (8.42a–c), the statements (8.43a,b) hold:

$$B > 0 \text{ and } C > 0, \quad (8.43a)$$

$$A, B, C \text{ are pairwise commutative.} \quad (8.43b)$$

The last statement is equivalent to

$$A, B, C \text{ can simultaneously be transformed to diagonal form.} \quad (8.43b')$$

Proof. Lemma 3.58 analyses the block diagonal of A (with respect to the row-block structure). Because of the $x-y$ symmetry, the same result holds for the column-block structure. Therefore, the spectrum of $B + 2h^{-2}I$ and $C + 2h^{-2}I$ is equal to

$$\left\{ h^{-2} \left[2 + 4 \sin^2 \frac{jh\pi}{2} \right] : 1 \leq j \leq N-1 \right\},$$

i.e., $4h^{-2} \sin^2 \frac{jh\pi}{2}$ are the eigenvalues of B and C . Since these values are positive, (8.43a) is proved. By Lemma 3.58 the eigenvectors e^{ij} of A (cf. Lemma 3.2) are also the eigenvectors of $B + 2h^{-2}I$, $C + 2h^{-2}I$ and hence of B and C . This proves (8.43b') and (8.43b). \square

The first half-step of the ADI method corresponds to the additive splitting

$$A = W - R \quad \text{with } W = \omega I + B \text{ and } R = \omega I - C \quad (8.44a)$$

and reads

$$x^{m+1/2} := \Phi_\omega^B(x^m, b) := (\omega I + B)^{-1}(b + \omega x^m - Cx^m), \quad (8.45a)$$

where ω is a (real) parameter. Interchanging the roles of B and C , i.e., *alternating* the directions, we generate the splitting (8.44b) of the second half-step (8.45b):

$$A = W - R \quad \text{with } W = \omega I + C, R = \omega I - B, \quad (8.44b)$$

$$x^{m+1} := \Phi_\omega^C(x^{m+\frac{1}{2}}, b) := (\omega I + C)^{-1}(b + \omega x^{m+\frac{1}{2}} - Bx^{m+\frac{1}{2}}). \quad (8.45b)$$

Remark 8.40. Each single half-step (8.45a,b) resembles a block-Jacobi method. For $\omega = 2h^{-2}$, iteration (8.45a) represents the row- and (8.45b) the column-block-Jacobi iteration. Because of (8.43a), the matrices $\omega I + B$ and $\omega I + C$ with $\omega \geq 0$ are positive definite and therefore regular; hence, the steps (8.45a,b) are well defined. Since, furthermore, $\omega I + B$ and $\omega I + C$ are tridiagonal matrices, the solution of $(\omega I + B)z = c$ or $(\omega I + C)z = c$ required in (8.45a,b) is easy to perform.

The complete ADI step $x^m \mapsto x^{m+1}$ is the product iteration

$$\Phi_\omega^{\text{ADI}} := \Phi_\omega^C \circ \Phi_\omega^B. \quad (8.45c)$$

8.5.2 General Representation

In the general case, we start from a splitting (8.42a): $A = B + C$ and assume (8.43a) in a weakened form. One of the matrices B or C may be only positive semidefinite. Without loss of generality, this might be C :

$$B > 0, \quad C \geq 0. \quad (8.46a)$$

Therefore, for

$$\omega > 0, \quad (8.46b)$$

the matrices $\omega I + B$ and $\omega I + C$ are positive definite and, in particular, regular. Hence, ADI iteration (8.45c) can be defined by (8.45a,b). To ensure practicability, we assume (8.46c):

$$\text{equations with } \omega I + B \text{ or } \omega I + C \text{ are easy to solve.} \quad (8.46c)$$

Theorem 8.41 (convergence). (a) *The iteration matrix of the ADI method is*

$$M_\omega^{\text{ADI}} = (\omega I + C)^{-1}(\omega I - B)(\omega I + B)^{-1}(\omega I - C). \quad (8.47a)$$

(b) *If (8.46a,b) holds, the ADI iteration converges.*

Proof. M_ω^{ADI} is the product of the iteration matrices $(\omega I + C)^{-1}(\omega I - B)$ and $(\omega I + B)^{-1}(\omega I - C)$ of the respective half-steps Φ_ω^C and Φ_ω^B (cf. §5.4). Lemma A.20 allows a cyclic permutation of the factors in the argument of the spectral radius:

$$\begin{aligned} \rho(M_\omega^{\text{ADI}}) &= \rho((\omega I - B)(\omega I + B)^{-1}(\omega I - C)(\omega I + C)^{-1}) \\ &\leq \|(\omega I - B)(\omega I + B)^{-1}(\omega I - C)(\omega I + C)^{-1}\|_2 \\ &\leq \|(\omega I - B)(\omega I + B)^{-1}\|_2 \|(\omega I - C)(\omega I + C)^{-1}\|_2. \end{aligned} \quad (8.47b)$$

As B is Hermitian, $B_\omega := (\omega I - B)(\omega I + B)^{-1}$ is also. In particular, it is a normal matrix, implying that $\rho(B_\omega) = \|B_\omega\|_2$ (cf. Theorem B.25). Therefore, (8.47b) becomes

$$\rho(M_\omega^{\text{ADI}}) \leq \rho(B_\omega) \rho(C_\omega) \quad (8.47c)$$

since analogous considerations also apply to $C_\omega := (\omega I - C)(\omega I + C)^{-1}$. By Remark A.15b, the spectrum of B_ω is equal to

$$\sigma(B_\omega) = \left\{ \frac{\omega - \beta}{\omega + \beta} : \beta \in \sigma(B) \right\}, \quad \rho(B_\omega) = \max_{\beta \in \sigma(B)} \left| \frac{\omega - \beta}{\omega + \beta} \right|. \quad (8.47d)$$

By assumption (8.46a), β is positive. This fact implies that $|\omega - \beta| < |\omega + \beta|$ for all $\omega > 0$. This proves $\rho(B_\omega) < 1$. Since C is only positive semidefinite, a similar argument leads to $\rho(C_\omega) \leq 1$. (8.47c) proves $\rho(M_\omega^{\text{ADI}}) < 1$. \square

Exercise 8.42. Formulate a convergence statement in the case of normal matrices B and C under the condition that the splittings (8.44a,b) are regular. For which ω are (8.44a,b) regular splittings in the model case?

Next, we want to determine the optimal value ω_{opt} of the ADI method. Here, we restrict ourselves to the minimisation of $\rho(B_\omega)$. If, as for the model problem, $\rho(C_\omega) = \rho(B_\omega)$ holds, minimisation of $\rho(B_\omega)$ is equivalent to the minimisation of the bound $\rho(B_\omega)\rho(C_\omega)$ in (8.47c).

The extreme eigenvalues of B (or their bounds) are assumed to be

$$0 < \beta_{\min} \leq \beta_{\max} \quad \text{with } \sigma(B) \subset [\beta_{\min}, \beta_{\max}]. \quad (8.48a)$$

In the model case, as seen in the proof of Remark 8.39, the eigenvalues of B are $4h^{-2} \sin^2(jh\pi/2)$ for $1 \leq j \leq N - 1$. This implies that

$$\beta_{\min} = 4h^{-2} \sin^2(h\pi/2), \quad \beta_{\max} = 4h^{-2} \cos^2(h\pi/2).$$

For any $\beta \in [\beta_{\min}, \beta_{\max}]$ and therefore for any $\beta \in \sigma(B)$, we have

$$\left| \frac{\omega - \beta}{\omega + \beta} \right| \leq \max \left\{ \left| \frac{\omega - \beta_{\min}}{\omega + \beta_{\min}} \right|, \left| \frac{\omega - \beta_{\max}}{\omega + \beta_{\max}} \right| \right\} \quad (\omega > 0) \quad (8.48b)$$

since $|\omega - \beta| / |\omega + \beta|$ as a function of β is decreasing in $[0, \omega]$ and increasing in $[\omega, \infty)$. To minimise the right-hand side in (8.48b), one has to determine ω from $\left| \frac{\omega - \beta_{\min}}{\omega + \beta_{\min}} \right| = \left| \frac{\omega - \beta_{\max}}{\omega + \beta_{\max}} \right|$. The result is given by

$$\omega_{\text{opt}} = \sqrt{\beta_{\min}\beta_{\max}}. \quad (8.48c)$$

Inserting this value into (8.47d), we obtain

$$\rho(B_{\omega_{\text{opt}}}) = \left(\sqrt{\beta_{\max}} - \sqrt{\beta_{\min}} \right) / \left(\sqrt{\beta_{\max}} + \sqrt{\beta_{\min}} \right).$$

Exercise 8.43. Prove for the Poisson model problem: (a) The following holds:

$$\begin{aligned} \omega_{\text{opt}} &= 2h^{-2} \sin h\pi, \\ \rho(B_{\omega_{\text{opt}}}) &= \left[\cos \frac{\pi h}{2} - \sin \frac{\pi h}{2} \right] / \left[\cos \frac{\pi h}{2} + \sin \frac{\pi h}{2} \right], \\ \rho(M_{\omega_{\text{opt}}}^{\text{ADI}}) &= [1 - \sin(\pi h)] / [1 + \sin(\pi h)]. \end{aligned}$$

(b) The convergence speed $\rho(M_{\omega_{\text{opt}}}^{\text{ADI}})$ coincides exactly with the optimal convergence rate (4.33) of the SOR iteration.

If we replace the definiteness in assumption (8.46a) by the M-matrix property, the convergence proof becomes much more difficult. A general convergence result of this kind (also for instationary ADI methods) is due to Alefeld [1]. Here, we call the method stationary if ω is constant during the iteration and instationary if it varies (as, e.g., it is assumed throughout the following section).

8.5.3 ADI in the Commutative Case

In addition to the assumptions (8.46a–c), we require that

$$BC = CB. \quad (8.49a)$$

Commutativity is equivalent to the simultaneous diagonalisability:

$$\begin{aligned} Q^H B Q &= D_B = \text{diag}\{\beta_\alpha : \alpha \in I\}, \\ Q^H C Q &= D_C = \text{diag}\{\gamma_\alpha : \alpha \in I\} \end{aligned} \quad (8.49b)$$

(cf. Theorem A.43), which here can be achieved by a unitary transformation Q , since B and C are Hermitian. Assumption (8.49b) implies that B_ω , C_ω , and the iteration matrix M_ω^{ADI} built from these matrices can also be transformed by Q to diagonal form (cf. (8.47a)):

$$Q^H M_\omega^{\text{ADI}} Q = \text{diag} \left\{ \frac{\omega - \gamma_\alpha}{\omega + \gamma_\alpha}, \frac{\omega - \beta_\alpha}{\omega + \beta_\alpha} : \alpha \in I \right\}. \quad (8.49c)$$

In the following, we apply the ADI method with varying parameters $\omega = \omega_m$:

$$y^{m+1} = \Phi_{\omega_m}^{\text{ADI}}(y^m, b) \quad (m \in \mathbb{N}).$$

Exercise 8.44. Let x be the solution of $Ax = b$. Prove that the error $\eta^m = y^m - x$ has the representation

$$\eta^m = M_{\omega_m}^{\text{ADI}} \cdot \dots \cdot M_{\omega_1}^{\text{ADI}} \eta^0.$$

We would like to choose the parameters $\omega_1, \omega_2, \dots, \omega_m \geq 0$ such that the spectral norm of the matrix $M_{\omega_m}^{\text{ADI}} \cdot \dots \cdot M_{\omega_1}^{\text{ADI}}$ becomes as small as possible:

$$\|M_{\omega_m}^{\text{ADI}} \cdot \dots \cdot M_{\omega_1}^{\text{ADI}}\|_2 \stackrel{!}{=} \min. \quad (8.50a)$$

Multiplications by unitary matrices do not change the spectral norm:

$$\begin{aligned} \|Q^H M_{\omega_m}^{\text{ADI}} \cdot \dots \cdot M_{\omega_1}^{\text{ADI}} Q\|_2 &= \|Q^H M_{\omega_m}^{\text{ADI}} Q \cdot \dots \cdot Q^H M_{\omega_1}^{\text{ADI}} Q\|_2 \\ &= \left\| \prod_{i=1}^m \text{diag} \left\{ \frac{\omega_i - \gamma_\alpha}{\omega_i + \gamma_\alpha}, \frac{\omega_i - \beta_\alpha}{\omega_i + \beta_\alpha} : \alpha \in I \right\} \right\|_2. \end{aligned}$$

Together with (8.49c), we obtain

$$\left\| \text{diag}_{\alpha \in I} \left\{ \prod_{i=1}^m \frac{\omega_i - \gamma_\alpha}{\omega_i + \gamma_\alpha} \frac{\omega_i - \beta_\alpha}{\omega_i + \beta_\alpha} \right\} \right\|_2 = \max_{\alpha \in I} \left| \prod_{i=1}^m \frac{\omega_i - \gamma_\alpha}{\omega_i + \gamma_\alpha} \frac{\omega_i - \beta_\alpha}{\omega_i + \beta_\alpha} \right|.$$

Hence, the minimisation problem (8.50a) is equivalent to

$$\max_{\alpha \in I} \left| \prod_{i=1}^m \frac{\omega_i - \gamma_\alpha}{\omega_i + \gamma_\alpha} \frac{\omega_i - \beta_\alpha}{\omega_i + \beta_\alpha} \right| \stackrel{!}{=} \min. \quad (8.50b)$$

Remark 8.45. For $m \geq n := \#I$, as in §8.3.2, one finds parameters ω_i bringing the left-hand side in (8.50b) to the minimum 0. For this purpose, the values ω_i must be an enumeration of the eigenvalues $\{\gamma_\alpha : \alpha \in I\} \cup \{\beta_\alpha : \alpha \in I\}$.

Since, in general, γ_α or β_α are not known, we optimise over a larger set $[a, b]$ containing the spectra of B and C , as we did in the third minimisation problem (8.23):

$$0 < a \leq \gamma_\alpha, \beta_\alpha \leq b \quad \text{for all } \alpha \in I.$$

Then, the minimisation problem takes the following form. Let

$$r_m(\zeta) := \prod_{i=1}^m \frac{\omega_i - \zeta}{\omega_i + \zeta}$$

be a rational function with a numerator and denominator of degree m replacing the previous polynomials. Substituting the discrete eigenvalues in (8.50b) by the interval $[a, b]$, we arrive at the problem

$$\begin{aligned} &\text{determine parameters } \{\omega_i : 1 \leq i \leq m\} \text{ so that} \\ &\max\{|r_m(\beta)r_m(\gamma)| : a \leq \beta, \gamma \leq b\} = \min. \end{aligned} \quad (8.51a)$$

Because of $\max_{\beta, \gamma} \{|r_m(\beta)r_m(\gamma)|\} = \max_\beta \{|r_m(\beta)|\} \max_\gamma \{|r_m(\gamma)|\}$, we may optimise each factor separately. Hence, problem (8.51a) simplifies to

$$\begin{aligned} &\text{determine parameters } \{\omega_i : 1 \leq i \leq m\} \text{ so that} \\ &\max\{|r_m(\zeta)| : a \leq \zeta \leq b\} = \min. \end{aligned} \quad (8.51b)$$

The following results are due to Wachspress [382] (see also Wachspress–Habetler [384] from 1960). We omit these proofs, since the derivation of Eqs. (8.52a–c) is presented in detail in the book of Varga [375, S. 224f].

Theorem 8.46 (optimal ADI parameters). (a) For any $m \in \mathbb{N}$, the problem (8.51b) has a unique solution $\{\omega_1, \dots, \omega_m\}$. The parameters ω_i are disjoint numbers in (a, b) .

(b) The increasingly ordered parameters $\omega_1 < \omega_2 < \dots < \omega_m$ satisfy

$$\omega_{m+1-i} = ab/\omega_i \quad \text{for } 1 \leq i \leq m. \quad (8.52a)$$

(c) Denote the parameters $\omega_1 < \omega_2 < \dots < \omega_m$ belonging to $m \in \mathbb{N}$ and the interval $[a, b]$ with $0 < a < b$ by $\omega_i(a, b, m)$ ($1 \leq i \leq m$). Then we have

$$\omega_{2m+1-i}(a, b, 2m) = \omega_i\left(\sqrt{ab}, \frac{a+b}{2}, m\right) + \sqrt{\omega_i\left(\sqrt{ab}, \frac{a+b}{2}, m\right)^2 - ab} \quad (8.52b)$$

for $i = 1, \dots, m$.

(d) The minimised quantities $\delta_m := \max\{|r_m(\zeta)| : a \leq \zeta \leq b\}$ for $m = 2^p$ are

$$\delta_m = \left(\sqrt{b_p} - \sqrt{a_p} \right) / \left(\sqrt{b_p} + \sqrt{a_p} \right), \quad (8.52c)$$

where $a_0 = a$, $b_0 = b$, $a_{i+1} = \sqrt{a_i b_i}$, $b_{i+1} = \frac{1}{2}(a_i + b_i)$ for $0 \leq i \leq p-1$.

Determining the ADI parameters ω_i is very easy for binary powers $m = 2^p$. For $p = 0$ (i.e., $m = 1$), we conclude from (8.52a) that

$$\omega_1(a, b, 1) = \sqrt{ab}, \quad (8.52d)$$

repeating the result in (8.48c). As soon as the parameters for $m = 2^{p-1}$ are known, those for $2m = 2^p$ can be obtained from formula (8.52b) for the indices $2m+1-i \in [m+1, \dots, 2m]$. The parameters ω_i for $1 \leq i \leq m$ result from (8.48a).

Evidently, one may apply the calculated parameters ω_i in a cyclic manner: $\omega_{i+km} := \omega_i$ ($1 \leq i \leq m$, $k \in \mathbb{N}$). Different from the case in §8.3.7, the cyclic ADI process does not lead to stability problems.

δ_m in (8.52c) is the bound for $r_m(B_\omega)$ and $r_m(C_\omega)$. Therefore, the asymptotic rate is bounded by $\rho_m := \delta_m^{2/m}$. One recognises from (8.52c) that ρ_m depends only on the ratio a/b , which in the model case has the size $\mathcal{O}(h^2)$. The recursions $a_{i+1} = \sqrt{a_i b_i}$ and $b_{i+1} = \frac{1}{2}(a_i + b_i)$ prove the following remark.

Remark 8.47. Let $a/b = \mathcal{O}(h^\tau)$ and assume (8.49a). For the optimal choice of the parameters, the cyclic ADI method with m parameters has the order τ/m :

$$\rho_m = 1 - \mathcal{O}(h^{\tau/2m}) = 1 - C_m h^{\tau/2m} + \mathcal{O}(h^{\tau/m}).$$

Hence, the instationary ADI method permits not only halving of the order (for the case $m = 1$, compare also with Exercise 8.43b), but any arbitrarily small (and hence very favourable) order can be reached for sufficiently large m . However, we will see in §8.5.5 that the obvious conclusion of choosing a rather large number m leads to practical difficulties.

The construction of the parameters ω_i in Theorem 8.46d is restricted to $m = 2^p$. For other m , the description of ω_i requires elliptic integrals (cf. Wachpress [383], Samarskii–Nikolaev [330, page 276]). Lebedev [260] was the first suggesting that the solution to the approximation problem (8.51b) could be reformulated into another one for rational functions that is already solved in 1877 by Zolotarev. In this connection, we refer to the review paper of Todd [363] concerning the

'legacy of Zolotarev' (see also Todd [364]). Approximation problems appearing here also play an important role in the iterative solution of the Sylvester matrix equation $AX - XB = C$ (A, B, C given, X unknown; cf. Starke [350] and Wachspress [383, §5]). Concerning the determination of the parameters in the case of nonsymmetric matrices, we refer to Starke–Niethammer [351].

Although the asymptotic convergence rates ρ_m in Remark 8.47 and the following Table 8.4 look quite favourable, the effective amount of work is less favourable because of the relatively expensive iteration (8.45a,b) (cf. Remark 8.49). Moreover, the assumption of commutativity (8.49a) is rarely satisfied in practice. As soon as it is violated, one is not able to achieve good convergence acceleration.

8.5.4 ADI Method and Semi-Iterative Methods

After choosing the Richardson method as the basic iteration, the half-steps (8.45a,b) have the representation (8.10b):

$$\begin{aligned} y^{m+\frac{1}{2}} &= \Theta_{m+\frac{1}{2}} (M_1^{\text{Rich}} y^m + N_1^{\text{Rich}} b) + (1 - \Theta_{m+\frac{1}{2}}) y^m, \\ y^{m+1} &= \Theta_{m+1} (M_1^{\text{Rich}} y^{m+\frac{1}{2}} + N_1^{\text{Rich}} b) + (1 - \Theta_{m+1}) y^{m+\frac{1}{2}} \end{aligned}$$

with $M_1^{\text{Rich}} = I - A$ and $N_1^{\text{Rich}} = I$, if we allow the matrix-valued factors

$$\Theta_{m+\frac{1}{2}} = (\omega I + B)^{-1}, \quad \Theta_{m+1} = (\omega I + C)^{-1}.$$

These equations correspond to the second formulation in §8.2. If, as in the case of §8.5.3, B and C commute with A , we obtain the first formulation (8.3): $y^m = \sum \alpha_{mj} x^j$, where x^j are the Richardson iterates and α_{mj} are matrices commuting with A . In this sense, one might view the ADI method as a semi-iteration with matrix-valued coefficients.

On the other hand, the ADI method can function as a basic iteration of the Chebyshev method, as shown in the next exercise.

Exercise 8.48. Assume that B , C , and ω satisfy (8.46a,b) and (8.49a). Prove:
(a) The matrix of the third normal form of Φ_ω^{ADI} is

$$W_\omega = \frac{1}{2\omega} (\omega I + C)(\omega I + B) \quad (\text{hint: (5.12c)}).$$

(b) Φ_ω^{ADI} is a positive definite iteration.

(c) Products $\Phi := \Phi_{\omega_1}^{\text{ADI}} \circ \Phi_{\omega_2}^{\text{ADI}} \circ \dots \circ \Phi_{\omega_m}^{\text{ADI}}$ with $\omega_j > 0$ form a positive definite iteration. Hint: Determine N_Φ in $\Phi(x, b) = x - N_\Phi(Ax - b)$ and show that $N_\Phi > 0$.

(d) In the stationary case, choose ω according to (8.52d). Determine the bounds in $\gamma W \leq A \leq \Gamma W$. What is the optimal damping factor ϑ_{opt} for Φ_ω^{ADI} (cf. Theorem 6.7)?

8.5.5 Amount of Work and Numerical Examples

The ADI method was already applied in §5.5.6 as a secondary iteration. In the following, we consider a general five-point formula ($C_A = 5$). The amount of work for solving the equations with the tridiagonal matrices $\omega I + C$, $\omega I + B$ amounts to $5n$ operations. Evaluating $b + \omega x - Cx$ and $b + \omega x - Bx$ requires $6n$ operations each. Because of $C_A = 5$, this leads to

$$C_{\Phi}^{\text{ADI}} = 4.4$$

and, in the Poisson model case, even to $C_{\Phi}^{\text{ADI}} = 4$.

The asymptotic rates $\rho_m = \delta_m^{1/m}$ attainable by (8.52c) are reported in Table 8.4. We observe that for small step sizes, good rates are achieved. The concrete results for $h = \frac{1}{128}$ with $m = 4$ different parameters from Table 8.5 confirm that the factor 0.5365 in Table 8.4 is reached. The convergence behaves regularly modulo m . Each second ratio $\|e^k\|_2 / \|e^{k-1}\|_2$ is ≈ 1 . However, since one cannot achieve the accuracy of $\|e^k\|_2 \approx \delta_m^k \|e^0\|_2$ with fewer than m iteration steps, the following dilemma arises:

(i) To exploit the good (asymptotic) convergence rate δ_m for large m , one must perform at least m iterations.

(ii) On the other hand, one would like to stop the iteration as soon as, e.g., the error becomes $\|e^k\|_2 \approx 1/1000$ (cf. Remark 2.34). The better the convergence rate, the fewer iterations one is willing to perform.

In the example of Table 8.5, about eight steps would be sufficient. Hence, one could still enlarge the cycle length from 4 to 8 (the corresponding result is $\|e^8\|_2 = 1.05_{10^{-3}}$); a further increase to 16 or more parameters would not help. The last two columns in Table 8.5 correspond to $\rho_{k,k-1}$ and $\rho_{k,0}$.

Remark 8.49. Good convergence rates are combined with a relatively high cost factor $C_{\Phi}^{\text{ADI}} = 4$ in the Poisson model case. For the example in Table 8.5, the effective amount of work is equal to $\text{Eff}(\Phi^{\text{ADI}}) = \frac{-4}{\log 0.5365} = 6.42$. For $h = 1/32$ and four parameters, we obtain $\text{Eff}(\Phi^{\text{ADI}}) = 4.06$ (for comparison: $\text{Eff}(\Phi^{\text{SOR}}) = 7.05$ in Example 2.28 and $\text{Eff}(\Phi_{\text{semi}}^{\text{SSOR}}) = 4.38$ in (8.41)).

m	$h = 1/32$	$1/64$	$1/128$
1	0.8215	0.9065	0.9521
2	0.5231	0.6373	0.7291
4	0.3735	0.4607	0.5365
8	0.3141	0.3874	0.4513
16	0.2880	0.3553	0.4139

Table 8.4 Asymptotic convergence rates ρ_m for ADI-cycle length m .

k	value in the middle	$\ e^k\ _2$	$\frac{\ e^k\ _2}{\ e^{k-1}\ _2}$	$\sqrt[m]{\frac{\ e^k\ _2}{\ e^0\ _2}}$
1	-0.0320913257	$5.02_{10^{-1}}$	0.6446	0.6446
2	-0.0342861024	$4.62_{10^{-1}}$	0.9106	0.7699
3	0.3534506991	$5.98_{10^{-2}}$	0.1295	0.4250
4	0.3538351873	$5.49_{10^{-2}}$	0.9187	0.5154
5	0.4031600829	$3.87_{10^{-2}}$	0.7055	0.5488
6	0.4063222547	$3.68_{10^{-2}}$	0.9487	0.6012
7	0.4976831847	$4.47_{10^{-3}}$	0.1215	0.4784
8	0.4976688610	$4.26_{10^{-3}}$	0.9536	0.5215
9	0.4961625617	$3.10_{10^{-3}}$	0.7284	0.5412
10	0.4961175712	$2.97_{10^{-3}}$	0.9568	0.5729
11	0.4990489164	$3.50_{10^{-4}}$	0.1179	0.4962
12	0.4990525844	$3.37_{10^{-4}}$	0.9625	0.5244
13	0.4993912351	$2.51_{10^{-4}}$	0.7454	0.5388
14	0.4994041549	$2.41_{10^{-4}}$	0.9612	0.5615
15	0.4999776724	$2.79_{10^{-5}}$	0.1154	0.5053
16	0.4999776614	$2.69_{10^{-5}}$	0.9671	0.5262

Table 8.5 ADI results (Poisson model problem, four parameters ω_i , $h = 1/128$).

Chapter 9

Gradient Method

Abstract The gradient method is an optimisation method of greedy type. For this purpose, the system of equations has to be rewritten as a minimisation problem (see Section 9.1). The gradient method $\Upsilon_{\text{grad}}[\Phi]$ derived in Section 9.2 determines the damping factors of the underlying iteration $\Phi \in \mathcal{L}$. It turns out that the convergence is not faster than the optimally damped version $\Phi_{\vartheta_{\text{opt}}}$ of Φ , but the method can be applied without knowing the spectral values determining ϑ_{opt} . In Section 9.3 we discuss the drawback of the gradient directions and introduce the conjugate directions in preparation for the conjugate gradient method in the next chapter. The final Section 9.4 mentions a variant of the gradient method: the minimal residual iteration which can be applied to any regular matrix A .

9.1 Reformulation as Minimisation Problem

9.1.1 Minimisation Problem

In the following, $A \in \mathbb{R}^{I \times I}$ and $b \in \mathbb{R}^I$ are real. We consider a system

$$Ax = b$$

and assume that

$$A \text{ is positive definite.} \quad (9.1)$$

System $Ax = b$ is associated with the function

$$F(x) := \frac{1}{2} \langle Ax, x \rangle - \langle b, x \rangle. \quad (9.2)$$

The derivative (gradient) of F is $F'(x) = \frac{1}{2}(A + A^\top)x - b$. Since $A = A^\top$ by assumption¹ (9.1), the derivative is equal to

¹ Under the weaker assumption (C.2), the function F can also be minimised, but the minimiser would not be the solution of $Ax = b$; i.e., the method would be inconsistent.

$$F'(x) = \text{grad } F(x) = Ax - b.$$

A necessary condition for a *minimum* of F is the vanishing of the gradient: $Ax = b$. Since the Hessian matrix $F''(x) = (F_{x_i x_j})_{i,j \in I} = A$ is positive definite, the solution of $Ax = b$ (in the following denoted by x^*) in fact leads to a minimum. This proves the next lemma.

Lemma 9.1. *Let $A \in \mathbb{R}^{I \times I}$ be positive definite. The solution of the system $Ax = b$ is equivalent to the solution to the minimisation problem*

$$F(x) \stackrel{!}{=} \min.$$

A second proof of Lemma 9.1 results from the representation

$$F(x) = F(x^*) + \frac{1}{2} \langle A(x - x^*), x - x^* \rangle \quad \text{with } x^* := A^{-1}b. \quad (9.3)$$

This equation proves $F(x) > F(x^*)$ for $x \neq x^*$, i.e., $x^* = A^{-1}b$ is the unique minimiser of F . The representation (9.3) is a particular case of the following Taylor expansion of F around an arbitrary value $\tilde{x} \in \mathbb{R}^I$:

$$F(x) = F(\tilde{x}) + \langle A\tilde{x} - b, x - \tilde{x} \rangle + \frac{1}{2} \langle A(x - \tilde{x}), x - \tilde{x} \rangle. \quad (9.4)$$

9.1.2 Search Directions

In the following, the minimisation of F with respect to a particular direction $p \in \mathbb{R}^I \setminus \{0\}$ plays a central role. Optimisation over all $x \in \mathbb{R}^I$ is replaced by the one-dimensional minimisation problem (9.5a,b):

$$f(\lambda) \stackrel{!}{=} \min \quad \text{for the function} \quad (9.5a)$$

$$f(\lambda) := F(x + \lambda p) \quad (x, p \in \mathbb{R}^I \text{ fixed}, \lambda \in \mathbb{R}). \quad (9.5b)$$

Replacing the variables x and \tilde{x} in (9.4) by $x + \lambda p$ and x , we obtain that

$$f(\lambda) = F(x) + \lambda \langle Ax - b, p \rangle + \frac{\lambda^2}{2} \langle Ap, p \rangle. \quad (9.5c)$$

$p \neq 0$ implies that $\langle Ap, p \rangle > 0$ (cf. (9.1)); hence, the minimum of the parabola f can be determined from $f'(\lambda) = 0$.

Lemma 9.2. *Assume $p \neq 0$ and (9.1): $A > 0$. The unique minimum of problem (9.5a,b) is attained at*

$$\lambda = \lambda_{\text{opt}}(r, p, A) := \frac{\langle r, p \rangle}{\langle Ap, p \rangle}, \quad (9.6a)$$

where

$$r := b - Ax.$$

In the following, the letter r always denotes the residual (residue) $b - Ax$ of the actual x . It is the negative defect $Ax - b$ and also the negative gradient $F' = Ax - b$.

The *optimal* search direction is evidently $p = x^* - x$ (or a nonvanishing multiple) because $f(\lambda_{\text{opt}}) = F(x^*)$ yields the global minimum. However, since $p = x^* - x$ requires knowledge of the solution, another proposal is needed. Let p be normalised by $\|p\|_2 = 1$. The directional derivative $f'(0) = -\langle r, p \rangle = \langle \text{grad } F(x), p \rangle$ at $\lambda = 0$ is maximal for the gradient direction $p = -r/\|r\|_2$ and minimal for the reverse direction $p = r/\|r\|_2$. The vector $\text{grad } F(x) = -r$ is the direction of the steepest ascent, while the residual r is the direction of the *steepest descent*. This consideration shows the optimality of $p = r$ from a *local* point of view. For $p = r$, the expression (9.6a) becomes

$$\lambda = \lambda_{\text{opt}}(r, r, A) = \frac{\|r\|_2^2}{\langle Ar, r \rangle} \quad \text{for } r := b - Ax \neq 0. \quad (9.6b)$$

The definition

$$\lambda_{\text{opt}}(r, 0, A) := 0 \quad (9.6c)$$

is added for formal reasons only: now $\lambda_{\text{opt}}(\cdot, \cdot, A)$ is defined for all arguments. As soon as $r = 0$ occurs, x is already the exact solution x^* .

9.1.3 Other Quadratic Functionals

The function F in (9.2) is not the only quadratic function having $x^* := A^{-1}b$ as the minimiser.

Lemma 9.3. (a) Any quadratic form with a unique minimum at $x^* = A^{-1}b$ has the form

$$F(x) = \frac{1}{2} \langle HA(x - x^*), A(x - x^*) \rangle + c = \frac{1}{2} \langle H(Ax - b), Ax - b \rangle + c \quad (9.7a)$$

with an arbitrary constant c and

$$H > 0. \quad (9.7b)$$

Here, in contrast to (9.1), A may be any regular matrix.

(b) To ensure that the calculation of $\text{grad } F(x) = A^H H A(x - x^*) = -A^H H r$ from the residual $r = Ax - b$ be practical, the matrix H must be such that the matrix-vector multiplication $r \mapsto A^H H r$ is feasible.

(c) Under assumption (9.1), $H := A^{-1}$ and $c := \frac{1}{2} \langle b, x^* \rangle$ may be chosen. Then F in (9.7a) coincides with F in (9.2).

Proof of (c). By (9.1), $H = A^{-1}$ satisfies (9.7b) (cf. Lemma C.4b). A comparison of (9.7a) and (9.3) shows that $c = F(x^*) = \frac{1}{2} \langle Ax^*, x^* \rangle - \langle b, x^* \rangle = -\frac{1}{2} \langle b, x^* \rangle$. \square

Conclusion 9.4. Let A be positive definite. The (energy) scalar product $\langle \cdot, \cdot \rangle_A$ and (energy) norm $\|\cdot\|_A$ are defined by (9.8a):

$$\langle x, y \rangle_A := \langle Ax, y \rangle, \quad \|x\|_A := \|A^{1/2}x\|_2 = \sqrt{\langle x, x \rangle_A}. \quad (9.8a)$$

The minimisation of F in (9.2) is equivalent to the minimisation problem

$$\|x - x^*\|_A \stackrel{!}{=} \min \quad \text{with } x^* := A^{-1}b. \quad (9.8b)$$

Proof. Problem (9.8b) may be replaced with $\|x - x^*\|_A^2 \stackrel{!}{=} \min$. The identity

$$\|x - x^*\|_A^2 = 2[F(x) - F(x^*)] \quad (\text{cf. (9.3)}) \quad (9.8c)$$

completes the proof. \square

- Remark 9.5.** (a) For the choice $H = I$ and $c = 0$, equation (9.7a) becomes $F(x) = \frac{1}{2}\|Ax - b\|_2^2$ and describes the *least-squares minimisation*.
(b) For $H = A^{-H}A^{-1} > 0$ and $c = 0$, the identity $F(x) = \frac{1}{2}\|x - x^*\|_2^2$ holds.
(c) For a positive definite K , the minimisation of the norm

$$\|x - x^*\|_K^2 = \|K^{1/2}(x - x^*)\|_2^2$$

corresponds to problem (9.7a) with

$$H = \frac{1}{2}A^{-H}KA^{-1}, \quad c = 0.$$

According to Lemma 9.3b, multiplying by KA^{-1} must be feasible.

Remark 9.6. Any iteration converging (weakly) monotonically with respect to the norm $\|\cdot\|_A$ leads to a descent sequence

$$F(x^0) \geq F(x^1) \geq \dots .$$

9.1.4 Complex Case

In the complex case of $A \in \mathbb{C}^{I \times I}$ and $b \in \mathbb{C}^I$, the function F can again be defined by (9.7a,b), provided that c in (9.7a) is real. Definition (9.2) cannot be generalised without change, since only real functions F can be minimised and, in general, F is not real because of the term $\langle b, x \rangle$. One has to replace F in (9.2) by

$$F(x) := \frac{1}{2} \langle Ax, x \rangle - \Re e \langle b, x \rangle \quad \text{for } x \in \mathbb{C}^I. \quad (9.9a)$$

Exercise 9.7. Assume (9.1) and let F be defined by (9.9a). Prove the following:

(a) F is real, and $\Re \langle b, x^* \rangle = \langle b, x^* \rangle$ holds for $x^* = A^{-1}b$.

(b) Equations (9.9b,c) hold for $x, y \in \mathbb{K}^I$:

$$F(x) = \frac{1}{2} [\langle A(x - x^*), x - x^* \rangle - \langle b, x^* \rangle], \quad (9.9b)$$

$$F(x) = F(y) + \Re \langle Ay - b, x - y \rangle + \frac{1}{2} \langle A(x - y), x - y \rangle. \quad (9.9c)$$

(c) The minimum of $f(\lambda) = F(x + \lambda p)$ over $\lambda \in \mathbb{C}$ with F in (9.9a) is attained for the value $\lambda_{\text{opt}}(r, p, A)$ in (9.6a), which in general is complex.

9.2 Gradient Method

Another name for the gradient method is the *method of steepest descent*.

9.2.1 Construction

In general, the gradient method is an algorithm for solving a minimisation problem $F(x) = \min$ with a differentiable function $F: \mathbb{R}^I \rightarrow \mathbb{R}$ (cf., e.g., Kosmol [240, §4], Quarteroni–Sacco–Saleri [314, §7.2.2]). We apply the gradient method only to the quadratic function F in (9.2) or (9.7a).

The gradient method minimises F iteratively in the direction of the *steepest descent*:

$$x^0 \in \mathbb{R}^I: \text{ arbitrary starting iterate,} \quad (9.10a)$$

iteration $m = 0, 1, \dots$:

$$r^m := b - Ax^m, \quad (9.10b)$$

$$x^{m+1} := x^m + \lambda_{\text{opt}}(r^m, r^m, A)r^m. \quad (9.10c)$$

The representation

$$r^{m+1} = b - Ax^{m+1} = b - A(x^m + \lambda_{\text{opt}}r^m) = r^m - \lambda_{\text{opt}}A r^m$$

allows the following update of the residual:

start:	$x^0: \text{ arbitrary}, \quad r^0 := b - Ax^0,$	$(9.11a)$
---------------	--	-----------

iteration $m = 0, 1, \dots :$	$x^{m+1} := x^m + \lambda_{\text{opt}}(r^m, r^m, A)r^m,$	$(9.11b)$
--------------------------------------	--	-----------

	$r^{m+1} := r^m - \lambda_{\text{opt}}(r^m, r^m, A)Ar^m$	$(9.11c)$
--	--	-----------

with $\lambda_{\text{opt}}(r^m, r^m, A)$ in (9.6b,c). The advantage of (9.11c) over (9.10b) is the fact that the product Ar^m is already calculated in (9.6b) when λ_{opt} is determined.

The gradient method (9.11a–c) is denoted by $\mathcal{Y}_{\text{grad}}[\varPhi_1^{\text{Rich}}]$ (cf. §9.2.4).

9.2.2 Properties of the Gradient Method

Remark 9.8. Assume (9.1). (a) In contrast to the previous methods, the iteration $x^m \mapsto \Phi(x^m, b)$ defined in (9.11a–c) is not linear.

(b) $\Phi(\cdot, \cdot)$ is continuous with respect to both of its arguments.

(c) The gradient method is consistent and convergent.

Proof. (a) $\lambda_{\text{opt}}(r^m, r^m) = \lambda_{\text{opt}}(b - Ax^m, b - Ax^m)$ is a nonconstant function of x^m and b . Hence, $\Phi(x, b) = x + \lambda_{\text{opt}}(b - Ax, b - Ax, A)(b - Ax)$ is not linear.

(c) Convergence will be proved in Theorem 9.10. If x^* is a solution of $Ax = b$, the residual r vanishes. Together with (9.6c), we conclude that $\Phi(x^*, b) = x^*$, i.e., F is consistent. \square

Although the gradient method is not linear, it can be interpreted as a semi-iterative method applied to a linear basic iteration.

Remark 9.9. The sequence $\{x^m\}$ of the gradient method (9.11a–c) is identical to the sequence $\{y^m\}$ of the semi-iterative Richardson method

$$y^{m+1} = y^m - \Theta_{m+1}(Ay^m - b) = \Phi_{\Theta_{m+1}}^{\text{Rich}}(y^m, b) \quad (9.12)$$

(cf. (8.10b)), if one chooses $y^0 = x^0$ and fixes the factors Θ_{m+1} by

$$\Theta_{m+1} := \lambda_{\text{opt}}(r^m, r^m, A) \quad \text{with } r^m := b - Ax^m.$$

Theorem 9.10 (convergence). Let A be positive definite and denote the extreme eigenvalues of A by $\lambda = \lambda_{\min}(A)$ and $\Lambda = \lambda_{\max}(A)$. Let F be defined by (9.2). Then, for any starting iterate x^0 , the sequence $\{x^m\}$ of the gradient method converges to the solution $x^* = A^{-1}b$ and satisfies the error estimates

$$F(x^m) - F(x^*) \leq \left(\frac{\Lambda - \lambda}{\Lambda + \lambda} \right)^{2m} [F(x^0) - F(x^*)], \quad (9.13a)$$

$$\|x^m - x^*\|_A \leq \left(\frac{\Lambda - \lambda}{\Lambda + \lambda} \right)^m \|x^0 - x^*\|_A. \quad (9.13b)$$

Proof. (i) By (9.8c), the estimates (9.13a) and (9.13b) are equivalent.

(ii) For proving (9.13b), it is sufficient to consider the case $m = 1$. The Richardson iteration

$$x_{\text{Rich}}^1 = x^0 - \Theta_{\text{Rich}}(Ax^0 - b) \quad \text{with } \Theta_{\text{Rich}} = 2/(\Lambda + \lambda)$$

yields the error $e_{\text{Rich}}^1 = M e^0$. The iteration matrix $M = M_{\Theta_{\text{Rich}}}^{\text{Rich}} = I - \Theta_{\text{Rich}} A$ has the norm $\|M\|_2 \leq \eta$, where

$$\eta = \frac{\Lambda - \lambda}{\Lambda + \lambda} \quad (9.13c)$$

(cf. Theorem 3.23). Since M commutes with A and $A^{1/2}$, we have

$$\tilde{e}_{\text{Rich}}^1 = M \tilde{e}^0 \quad \text{for } \tilde{e}_{\text{Rich}}^1 := A^{1/2} e_{\text{Rich}}^1, \quad \tilde{e}^0 := A^{1/2} e^0.$$

By $\|\tilde{e}^0\|_2 = \|e^0\|_A$ and $\|\tilde{e}_{\text{Rich}}^1\|_2 = \|e_{\text{Rich}}^1\|_A$, we can estimate e_{Rich}^1 by

$$\|e_{\text{Rich}}^1\|_A = \|\tilde{e}_{\text{Rich}}^1\|_2 \leq \|M\|_2 \|\tilde{e}^0\|_2 \leq \eta \|e^0\|_A.$$

Both x_{Rich}^1 and x^1 are of the form $x^0 + \Theta r^0$. Since the iterate x^1 of the gradient method minimises the error $\|x^1 - x^*\|_A$ (cf. Conclusion 9.4), the assertion follows for $m = 1$: $\|x^1 - x^*\|_A \leq \|e_{\text{Rich}}^1\|_A \leq \eta \|e^0\|_A$. \square

Corollary 9.11. (a) The factor η in (9.13c) is the minimal one in (9.13a,b).

(b) The asymptotic convergence rate of the gradient method is η .

(c) η depends only on the condition $\kappa(A) = \text{cond}_2(A) = \Lambda/\lambda$:

$$\eta = \frac{\kappa - 1}{\kappa + 1} \quad \text{with } \kappa = \kappa(A). \quad (9.14)$$

Proof. Let v_1 and v_2 with $\|v_1\|_2 = \Lambda$ and $\|v_2\|_2 = \lambda$ be the eigenvectors corresponding to Λ and λ . For $x^0 := x^* + e^0$ with $e^0 := v_1 \pm v_2$, one obtains $e^1 = \eta(v_1 \mp v_2)$ and $e^2 = \eta^2(v_1 \pm v_2) = \eta^2 e^0$. $\|e^2\|_A/\|e^0\|_A = \eta^2$ proves part (a). Analogously, $e^{2k} = \eta^{2k} e^0$ shows part (b). \square

Usually, the values $\lambda = \lambda_{\min}(A)$ and $\Lambda = \lambda_{\max}(A)$ are not known. Their numerical approximation is discussed below.

Remark 9.12 (approximation of λ and Λ). (a) Let $\langle e^0, v_i \rangle \neq 0$ hold for the eigenvectors v_1 and v_2 of A corresponding to Λ and λ . Then

$$\rho_{m+1,m} := \|x^{m+1} - x^*\|_A / \|x^m - x^*\|_A \quad (x^m \text{ defined by (9.11a-c)})$$

converges to $\eta = (\kappa - 1)/(\kappa + 1)$ in (9.14).

(b) Using $\rho(M_{\Theta}^{\text{Rich}}) = 1 - \Theta\lambda$ (e.g., for $\Theta = 1/\|A\|_{\infty}$), we can approximate λ from the convergence behaviour of the Richardson method. The approximation of η yields an approximation of $\kappa = (1 + \eta)/(1 - \eta)$ which allows us to determine the other extreme eigenvalue Λ by $\Lambda/\lambda = \kappa$.

Finally, we describe the relation of the gradient method with the Krylov space (cf. §8.1.4).

Proposition 9.13. *The errors $e^m = x^m - x^*$ of the gradient method satisfy*

$$\mathcal{K}_m(A, e^0) = \text{span}\{e^0, e^1, \dots, e^{m-1}\}$$

for all $m \in \mathbb{N}$. The residuals $r^{\mu} = -Ae^{\mu}$ ($0 \leq \mu \leq m-1$) span the space $\mathcal{K}_m(A, r^0) = A\mathcal{K}_m(A, e^0)$.

Proof. As long as $r^m \neq 0$, the equivalent semi-iteration corresponds to polynomials p_m of degree m , so that Conclusion 8.13b applies. Otherwise, there is a first m' with $r^{m'} = -Ae^{m'} = 0$. Since $e^{m'} = p_{m'}(A)e^0$, $\deg_A(e^0) \leq m'$ follows (cf. Definition 8.10). Exercise 8.11a states that $\mathcal{K}_m(A, e^0) = \mathcal{K}_{m'}(A, e^0)$ for all $m \geq m'$. Therefore, the statement also holds in the degenerate case of $r^m = 0$. \square

9.2.3 Numerical Examples

At first view, the gradient method seems to surpass the semi-iterative method because, in the latter case, the parameters Θ_k have to be chosen *a priori* (cf. (9.12)), whereas the gradient method determines these values *a posteriori* in an optimal way. However, the opposite is the case. While the Chebyshev method leads to an improvement of the order, Corollary 9.11a yields the convergence rate η in (9.13c), which is as slow as the stationary Richardson method with $\Theta = \Theta_{\text{opt}}$ (cf. Theorem 3.23).

In the model case, λ and A in (3.1b,c) are known and lead to

$$\eta = \frac{\cos^2(\pi h/2) - \sin^2(\pi h/2)}{\cos^2(\pi h/2) + \sin^2(\pi h/2)} = \cos(\pi h) = 1 - \frac{\pi^2 h^2}{2} + \mathcal{O}(h^4).$$

The low convergence speed of the gradient method is confirmed by the following numerical example (Poisson model problem (2.33a,b)). Table 9.1 contains the results for the step size $h = 1/32$ and the starting iterate $x^0 = 0$. The ratios $\|x^{m+1} - x^*\|_A / \|x^m - x^*\|_A$ in the last column of Table 9.1 clearly approximate the asymptotic convergence rate $h = \cos \frac{\pi}{32} = 0.9951847$. Even after 300 iterations, the value $u_{16,16}$ at the midpoint is wrong by 50%: 0.2778 instead of 0.5. The error measured in the scaled energy norm $h^2 \|x^m - x^*\|_A$ deviates very little from the maximum norm $\|e^m\|_\infty$. However, the error with respect to the energy norm $\|\cdot\|_A$ decreases uniformly, whereas the ratios of $\|e^m\|_\infty$ oscillate. Because $\eta = \rho(M^{\text{Jac}})$, the results in Table 9.1 and Table 3.1 prove to be very similar.

m	value in the middle	$\frac{\ e^m\ _A}{\ e^{m-1}\ _A}$	m	value in the middle	$\frac{\ e^m\ _A}{\ e^{m-1}\ _A}$
1	-1.86560 ₁₀₋₃		100	-1.89771 ₁₀₋₂	0.993444
2	-3.52293 ₁₀₋₃	0.844824	110	-5.13520 ₁₀₋₃	0.993749
3	-4.84034 ₁₀₋₃	0.907804	120	1.01805 ₁₀₋₂	0.993990
4	-5.97611 ₁₀₋₃	0.935293	200	1.45146 ₁₀₋₁	0.994852
5	-7.10198 ₁₀₋₃	0.946906	250	2.18301 ₁₀₋₁	0.995024
6	-8.16295 ₁₀₋₃	0.953838	296	2.73548 ₁₀₋₁	0.995102
7	-9.23998 ₁₀₋₃	0.958895	297	2.75710 ₁₀₋₁	0.995103
8	-1.02699 ₁₀₋₂	0.962711	298	2.75702 ₁₀₋₁	0.995104
9	-1.13230 ₁₀₋₂	0.965778	299	2.77844 ₁₀₋₁	0.995105
10	-1.23360 ₁₀₋₂	0.968271	300	2.77836 ₁₀₋₁	0.995106

Table 9.1 Result of the gradient method $\Upsilon_{\text{grad}}[\Phi_1^{\text{Rich}}]$ for $h = 1/32$ (Poisson model problem).

9.2.4 Gradient Method Based on Other Basic Iterations

Let $\mathcal{Y}_{\text{grad}} \in \mathcal{N}$ be the notation of the gradient method. Above we applied the gradient method to the Richardson iteration Φ_1^{Rich} resulting in the nonlinear iteration $\mathcal{Y}_{\text{grad}}[\Phi_1^{\text{Rich}}]$. Now we discuss $\mathcal{Y}_{\text{grad}}[\Phi]$ for other iterations.

9.2.4.1 Standard Version

By Remark 9.9, the gradient method is a particular semi-iterative method with Richardson's iteration as the basic iteration. From the analysis of semi-iterative methods, we know that other basic iterations Φ may better suit because of a smaller spectral condition number $\kappa(NA)$ with the matrix $N = N^\Phi[A]$ of the second normal form. This suggests replacing Richardson's iteration by another one (e.g., the SSOR iteration; cf. §8.4.4). For this purpose, the matrix A has to be replaced formally with $\hat{A} := NA$, because the Richardson method applied to the left-transformed (preconditioned) system $\hat{A}\hat{x} = \hat{b} := Nb$ is equivalent to Φ applied to A (cf. Proposition 5.44).

Let A and N be positive definite. Since, in general, the matrix $\hat{A} = NA$ is no longer symmetric, \hat{A} does not satisfy the assumption (9.1), which is necessary for the applicability of the gradient method. A remedy is offered in §5.6.6: the iteration $\check{\Phi}$ defined by

$$\check{x}^{m+1} = \check{x}^m - N^{1/2}(AN^{1/2}\check{x}^m - b) = \check{x}^m - (\check{A}\check{x}^m - \check{b}), \quad (9.15a)$$

$$\check{A} := N^{1/2}AN^{1/2}, \quad \check{b} := N^{1/2}b, \quad (9.15b)$$

is equivalent to the basic iteration $\Phi(x^m, b) = x^m - N(Ax^m - b)$ via the transformation

$$\check{x}^m = N^{-1/2}x^m$$

(multiplying by $N^{\pm 1/2}$ is of course not practically feasible²) and represents the Richardson iteration for the system $\check{A}\check{x} = \check{b}$ with the positive definite matrix \check{A} . Therefore, the gradient method has to be applied not to F in (9.2) but to

$$\check{F}(\check{x}) := \frac{1}{2} \langle \check{A}\check{x}, \check{x} \rangle - \langle \check{b}, \check{x} \rangle. \quad (9.15c)$$

Its negative gradient is the new residual

$$\check{r} := \check{b} - \check{A}\check{x} = N^{1/2}r \quad (r = b - Ax).$$

The gradient method (9.11b,c) associated with \check{A} yields the iterates

$$\mathcal{Y}_{\text{grad}}[\check{\Phi}] : \quad \left. \begin{aligned} \check{x}^{m+1} &:= \check{x}^m + \check{\lambda}_{\text{opt}} \check{r}^m, \\ \check{r}^{m+1} &:= \check{r}^m - \check{\lambda}_{\text{opt}} \check{A} \check{r}^m \end{aligned} \right\} \quad \text{with} \quad \check{\lambda}_{\text{opt}} := \frac{\|\check{r}^m\|_2^2}{\langle \check{A}\check{r}^m, \check{r}^m \rangle}.$$

² In principle, we may replace the factorisation $N = N^{\frac{1}{2}} \cdot N^{\frac{1}{2}}$ with the Cholesky decomposition $N = VV^H$ and introduce $\check{A} = V^H AV$ (cf. Exercise 5.63).

Inserting $\check{A} = N^{1/2}AN^{1/2}$, $\check{x}^m = N^{-1/2}x^m$, $\check{r}^m = N^{1/2}r^m$, and solving the defining equations for x^{m+1} and r^{m+1} , we obtain the following algorithm for the iterates $\{x^m\}$:

$$x^{m+1} := x^m + \check{\lambda}_{\text{opt}}Nr^m \quad \text{with } N = N_{\Phi}[A], \quad (9.16a)$$

$$\Upsilon_{\text{grad}}[\Phi] : \quad r^{m+1} := r^m - \check{\lambda}_{\text{opt}}ANr^m \quad \text{with} \quad (9.16b)$$

$$\check{\lambda}_{\text{opt}} := \lambda_{\text{opt}}(r^m, Nr^m, A) = \frac{\langle Nr^m, r^m \rangle}{\langle ANr^m, Nr^m \rangle}. \quad (9.16c)$$

The quantities $N^{\pm\frac{1}{2}}$ do no longer appear, so that $\Upsilon_{\text{grad}}[\Phi]$ defined by (9.16a–c) is a practical algorithm. We call $\Upsilon_{\text{grad}}[\Phi]$ defined in (9.16a–c) the *gradient method applied to the basic iteration $\Phi(\cdot, \cdot, A)$* . The term ‘preconditioned gradient method’ is also used. Note that not the method but the gradient is ‘preconditioned’. While the method (9.11a–c) takes the (negative) gradient r^m as search direction, this is replaced in (9.17a–e) with the ‘preconditioned’ gradient $q = Nr$.

The derivation of (9.16a–c) requires $N > 0$. Nevertheless, $\Upsilon_{\text{grad}}[\Phi]$ is well defined as long as $A > 0$ and N is regular since this guarantees $\langle ANr^m, Nr^m \rangle > 0$ for $r^m \neq 0$ ($r^m = 0$ is a ‘lucky breakdown’ since the exact solution $x = x^m$ is found). However, the convergence statements are restricted to the case $A > 0$ and $N > 0$. Since $A > 0$ implies $N > 0$ for $\Phi \in \mathcal{L}_{\text{sym}}$, symmetric iterations Φ are the natural basic iterations of the gradient method.

In analogy to Remark 9.9, equation (9.16a) proves the next remark.

Remark 9.14. The sequence $\{x^m\}$ of the gradient method (9.16a–c) applied to the positive definite iteration Φ is identical to the sequence $\{y^m\}$ of the semi-iterative method

$$y^{m+1} = y^m - \Theta_{m+1}N(Ay^m - b) = \Theta_{m+1}\Phi(y^m, b, A) + (1 - \Theta_{m+1})y^m$$

with Φ as the basic iteration when the factors Θ_{m+1} are defined by $\check{\lambda}_{\text{opt}}$ in (9.16c).

The amount of work needed by the algorithm (9.16a–c) can be reduced by introducing $q^m := Nr^m$ and $a^m := Aq^m$. Note that q^m and a^m need not be saved for the next iteration step.

start:	x^0 arbitrary; $r^0 := b - Ax^0;$	(9.17a)
iteration $m = 0, 1, \dots :$	$q^m := Nr^m;$ $a^m := Aq^m;$	(9.17b)
	$\lambda_{\text{opt}} := \lambda_{\text{opt}}(r^m, q^m, A) = \frac{\langle q^m, r^m \rangle}{\langle a^m, q^m \rangle};$	(9.17c)
	$x^{m+1} := x^m + \lambda_{\text{opt}}q^m;$	(9.17d)
	$r^{m+1} := r^m - \lambda_{\text{opt}}a^m;$	(9.17e)

Remark 9.15. The representation (9.17a–e) shows that for each iteration step only one multiplication by N and one by A are necessary. For $N = I$, we regain the algorithm (9.11a–c).

The convergence of the method (9.17a–e) follows by applying the convergence statement of Theorem 9.10 to the transformed problem (9.15c): $\tilde{\Phi}(\check{x}) = \min$. First, we obtain an error estimate for \check{x}^m with respect to the corresponding energy norm $\|\cdot\|_{\check{A}}$. Because of

$$\begin{aligned}\|\check{x}^m - \check{x}^*\|_{\check{A}}^2 &= \langle \check{A}(\check{x}^m - \check{x}^*), \check{x}^m - \check{x}^* \rangle \\ &= \langle A(x^m - x^*), x^m - x^* \rangle \\ &= \|x^m - x^*\|_A^2 \quad (\check{x}^* = N^{-1/2}x^*)\end{aligned}$$

the \check{A} -estimates of $\check{x}^m - \check{x}^*$ carry over to the A -norm of the error $x^m - x^*$.

Theorem 9.16 (convergence). *Let A and $N = W^{-1}$ be positive definite. If*

$$\gamma W \leq A \leq \Gamma W, \quad (9.18a)$$

the iterates in (9.17a–e) satisfy the error estimate

$$\|x^m - x^*\|_A \leq \left(\frac{\Gamma - \gamma}{\Gamma + \gamma} \right)^m \|x^0 - x^*\|_A. \quad (9.18b)$$

Remark 9.17. Under an assumption analogous to that in Remark 9.12a, we conclude for algorithm (9.17a–e) that the convergence factors converge to $\eta = \frac{\kappa-1}{\kappa+1}$ with $\kappa := \kappa(NA) = \Gamma/\gamma$ (here, γ and Γ are the optimal bounds in (9.18a)). Therefore, the gradient method (9.17a–e) can be used to determine the spectral condition number Γ/γ .

We regard the gradient method as a general technique that can be applied to all positive definite iterations Φ and problems with $A > 0$. This is the same situation as the Chebyshev method which also requires specifying the basic iteration.

Theorem 9.18. *Let Φ be a positive definite iteration and assume that $A > 0$. The gradient method applied to Φ converges as fast as the optimally damped iteration $\Phi_{\vartheta_{\text{opt}}}$ for $\vartheta_{\text{opt}} = \frac{2}{\Gamma+\gamma}$. However, the explicit knowledge of the optimal bounds γ and Γ in (9.18a) is not necessary.*

Proof. Compare the results of Theorem 6.7 and (9.18b), and use $\gamma = \lambda_{\min}(NA)$ and $\Gamma = \lambda_{\max}(NA)$. \square

Now the statement of Proposition 9.13 reads as follows.

Remark 9.19. The errors $e^m = x^m - x^*$ of the gradient method (9.17a–e) satisfy

$$\mathcal{K}_m(NA, e^0) = \text{span}\{e^0, e^1, \dots, e^{m-1}\}$$

for all $m \in \mathbb{N}$. The residuals $r^\mu = -Ae^\mu$ ($0 \leq \mu \leq m-1$) span the space $A\mathcal{K}_m(NA, e^0) = \mathcal{K}_m(AN, r^0)$.

9.2.4.2 Residual Oriented Version

In (9.15a) we interpreted the iteration Φ as Richardson's iteration applied to the positive definite matrix \bar{A} . This is not the only possibility. Φ is also equivalent to Φ_1^{Rich} applied to \bar{A} :

$$\begin{aligned}\bar{x}^{m+1} &:= \bar{x}^m - (\bar{A}\bar{x}^m - \bar{b}) \quad \text{with} \\ \bar{A} &:= A^{1/2}NA^{1/2} > 0, \quad \bar{b} := A^{1/2}Nb, \quad \bar{x}^m := A^{1/2}x^m.\end{aligned}\quad (9.19)$$

Exercise 9.20. Prove the following: (a) The application of the gradient method to the minimisation of $\bar{F}(\bar{x}) := \frac{1}{2} \langle \bar{A}\bar{x}, \bar{x} \rangle - \langle \bar{b}, \bar{x} \rangle$ yields (9.20a–c)—denoted by $\Upsilon_{\text{grad}}^{\text{res}}[\Phi]$ —after a reformulation using the x -quantities:

$$\text{start: } x^0 \text{ arbitrary, } q^0 := N(b - Ax^0), \quad (9.20a)$$

$$x^{m+1} := x^m + \lambda_{\text{opt}} q^m \quad (9.20b)$$

$$\begin{aligned}\text{with } \lambda_{\text{opt}} &:= \lambda_{\text{opt}}(q^m, Aq^m, N) = \frac{\langle q^m, Aq^m \rangle}{\langle NAq^m, Aq^m \rangle}, \\ q^{m+1} &:= q^m - \lambda_{\text{opt}} NA q^m.\end{aligned}\quad (9.20c)$$

(b) The methods $\Upsilon_{\text{grad}}[\Phi]$ in (9.17a–e) and $\Upsilon_{\text{grad}}^{\text{res}}[\Phi]$ in (9.20a–c) are different. Choosing $N = I$, we do not regain the gradient method (9.11a–c).

(c) Let γ and Γ be the bounds in (9.18a). Then the error estimate (9.20d) holds:

$$\|N^{1/2}A(x^m - x^*)\|_2 \leq \left(\frac{\Gamma - \gamma}{\Gamma + \gamma} \right)^m \|N^{1/2}A(x^0 - x^*)\|_2. \quad (9.20d)$$

Note that both versions (9.17a–e) and (9.20a–c) lead to the same convergence rate, but the involved norms are different. If $W \sim A$, the norms $\|\cdot\|_A$ and $\|\cdot\|_{ANA}$ are equivalent. On the other hand, for $N = I$ the residual $A(x^m - x^*) = r^m$ is the subject of minimisation and for $N \sim I$ the norms $\|N^{1/2}r^m\|_2$ and $\|r^m\|_2$ are equivalent.

Remark 9.21. The statements of Remark 9.19 also hold for the errors $e^m = x^m - x^*$ of the gradient method (9.20a–c) as well as for the results of the following variant (9.21a–c).

9.2.4.3 Directly Positive Definite Case

Assume $\Phi \in \mathcal{L}_{>0}$, i.e., the iteration $\Phi(\cdot, \cdot, A)$ is directly positive definite:

$$N[A]A > 0 \quad (\text{cf. Definition 5.14}).$$

Then the original method (9.10a–c) can be applied with A replaced with the matrix $N[A]A$. Note that in this case the matrix $A \in \mathfrak{D}(\Phi)$ is only required to be regular.

The quadratic function

$$F(x) := \frac{1}{2} \langle N Ax, x \rangle - \langle Nb, x \rangle$$

replaces F in (9.2). The corresponding gradient method reads as follows:

$$\text{start: } x^0 \text{ arbitrary, } q^0 := N(b - Ax^0), \quad (9.21a)$$

$$x^{m+1} := x^m + \lambda q^m \quad \text{with } \lambda := \frac{\|q^m\|_2^2}{\langle NAq^m, q^m \rangle}, \quad (9.21b)$$

$$q^{m+1} := q^m - \lambda NA q^m. \quad (9.21c)$$

Theorem 9.22. Let NA be positive definite with

$$\gamma := \lambda_{\min}(NA) \quad \text{and} \quad \Gamma := \lambda_{\max}(NA).$$

The iterates in (9.21a–c) satisfy the error estimate

$$\|x^m - x^*\|_{NA} \leq \left(\frac{\Gamma - \gamma}{\Gamma + \gamma} \right)^m \|x^0 - x^*\|_{NA}.$$

9.2.5 Numerical Examples

The SSOR iteration is used as a basic iteration of the gradient method for the Poisson model case. As in Table 6.1, we choose the relaxation parameter

$$\omega = 1.82126912$$

for the step size $h = 1/32$. The results given in Table 9.2 suggest the convergence rate $\eta \approx 0.769$. From (9.14), we conclude the spectral condition number

$$\Gamma/\gamma = \kappa = (1 + \eta)/(1 - \eta) = 7.66.$$

According to Table 6.1, the convergence rate of the SSOR iteration equals 0.8796. From $\rho(M^{\text{SSOR}}) = 1 - \lambda$, we deduce $\lambda = 0.1204$, implying $\Gamma = 7.66$ and $\gamma = 0.922$. Hence,

$$\vartheta_{\text{opt}} = 2/(\Gamma + \gamma) \approx 1.92$$

is the optimal damping or (more precisely) extrapolation factor for $\Phi_{\omega=1.82}^{\text{SSOR}}$ in the Poisson model case with $h = 1/32$.

m	value in the middle	$\frac{\ e^m\ _A}{\ e^{m-1}\ _A}$
1	0.2851075107	0.4576
2	0.9245177570	0.5192
3	0.1780816984	0.5886
4	0.2274720552	0.6454
5	0.2956906889	0.6858
10	0.4381492069	0.7577
20	0.4954559469	0.7672
30	0.4996724015	0.7682
40	0.4999764630	0.7685
50	0.4999983084	0.7687
60	0.4999998782	0.7688
70	0.4999999912	0.7689

Table 9.2 Gradient method $\Upsilon_{\text{grad}}[\Phi_{\omega=1.82}^{\text{SSOR}}]$ applied to the SSOR iteration for $h = 1/32$.

9.3 Method of the Conjugate Directions

9.3.1 Optimality with Respect to a Direction

The slowness of the gradient method is demonstrated in Theorem 9.10 by the two-dimensional subspace spanned by the two extreme eigenvectors. Therefore, a system of two equations is able illustrate this situation. The matrix

$$A = \text{diag}\{\lambda_1, \lambda_2\} \quad \text{with } 0 < \lambda_1 \leq \lambda_2$$

has the condition $\text{cond}_2(A) = \lambda_2/\lambda_1$. The corresponding function F in (9.2) leads to ellipses as level curves

$$N_c := \{x \in \mathbb{R}^2 : \Phi(x) = c\}, \quad \text{where } c \in \mathbb{R}.$$

In the two-dimensional case, the gradient method can be illustrated graphically as follows: The point x^m [x^{m+1}] lies on the ellipse $E^{(m)} := N_c$ with $c = F(x^m)$ [or $E^{(m+1)} := N_c$ with $c = F(x^{m+1})$, respectively]. The straight line $x^m x^{m+1}$ is vertical to $E^{(m)}$ and tangential to $E^{(m+1)}$. Therefore, succeeding straight lines (i.e., the corrections $x^{m+1} - x^m$) form right angles. Figure 9.1 shows the case of an elongated ellipse, where the iteration path forms a zigzag line. This illustrates that the approximation to the centre requires many iteration steps. Note that the ellipses are more elongated the larger the condition is. In the case of a circle ($\lambda_1 = \lambda_2$), the first correction would already yield the exact solution x^* .

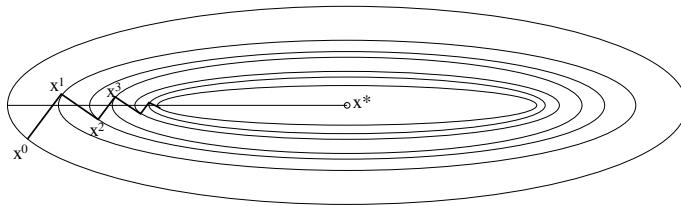


Fig. 9.1 The iterates x^m and the corresponding level lines of the function F .

From the fact that the corrections $x^{m+3} - x^{m+2}$ and $x^{m+1} - x^m$ are parallel, one understands that the iterate x^{m+2} must be corrected in exactly the same direction in which x^m has been corrected previously. Hence, x^{m+2} has lost the property of x^{m+1} being optimal with respect to the direction $x^{m+1} - x^m$. We define:

$$\begin{aligned} x \text{ is optimal with respect to a direction } p \neq 0, \text{ if} \\ F(x) \leq F(x + \lambda p) \quad \text{for all } \lambda \in \mathbb{K}. \end{aligned}$$

Lemma 9.23. *The optimality of x with respect to p is equivalent to*

$$p \perp r := b - Ax.$$

Proof. A necessary condition for $f(\lambda) = F(x + \lambda p)$ in (9.5c) to be minimal for $\lambda = 0$ is $\langle Ax - b, p \rangle = -\langle r, p \rangle = 0$. As (9.5c) is restricted to the field \mathbb{R} , use (9.9c) for the complex case. \square

Exercise 9.24. x is called optimal with respect to a subspace \mathcal{U} if $F(x) \leq F(x + \xi)$ for all $\xi \in \mathcal{U}$. Prove that x is optimal with respect to \mathcal{U} if and only if

$$r = b - Ax \perp \mathcal{U}.$$

Remark 9.25. The iterates x^m of the gradient method satisfy (9.22a,b):

$$x^{m+1} \text{ is optimal with respect to } r^m = b - Ax^m, \quad (m \geq 0) \quad (9.22a)$$

$$r^{m+1} \perp r^m. \quad (9.22b)$$

Proof. By Lemma 9.23, (9.22a) and (9.22b) are equivalent. $r^m \perp r^{m+1} = r^m - \lambda_{\text{opt}}(r^m, r^m) A r^m$ follows from the definition (9.6b,c) of λ_{opt} . \square

The principal deficit of the gradient method can be stated as follows. The relation $r^{m+1} \perp r^m$ is not transitive, i.e., $r^m \perp r^{m+1}$ and $r^{m+1} \perp r^{m+2}$ do not imply $r^m \perp r^{m+2}$. Therefore, in general, x^{m+2} has lost its optimality with respect to r^m .

9.3.2 Conjugate Directions

The change of x into $x' := x + q$ ($q \neq 0$) transforms the residual $r = b - Ax$ of x into the residual

$$r' = b - Ax' = b - A(x + q) = b - Ax - Aq = r - Aq$$

of x' . Let x be optimal with respect to the direction p :

$$r \perp p.$$

The new value x' remains optimal with respect to p if and only if $r' \perp p$, i.e., $Aq \perp p$, because the latter property is equivalent to

$$-\langle Aq, p \rangle = \langle r - Aq, p \rangle = \langle r', p \rangle = 0.$$

This proves the next statement.

Lemma 9.26. *The optimality of x with respect to $p \neq 0$ implies the optimality of $x' = x + q$ with respect to the same $p \neq 0$ if and only if*

$$Aq \perp p. \quad (9.23)$$

Vectors p, q with the property (9.23) are called *conjugate*. The term ‘conjugate’ can also be replaced with ‘ A -orthogonal’, abbreviated as

$$q \perp_A p,$$

where \perp_A denotes orthogonality with respect to the scalar product $\langle \cdot, \cdot \rangle_A$ in (9.8a). Note that the latter definitions only make sense if $A > 0$.

Condition (9.23) leads us to the following method of conjugate directions.

<i>method of conjugate directions</i>		(9.24)
start:	x^0 arbitrary, $r^0 := b - Ax^0$;	
loop:	for $m = 0, 1, \dots, n-1$: $(n := \#I)$	
	choose a direction $p^m \neq 0$ which is conjugate to all preceding directions $p^\ell (\ell < m)$;	(9.24a)
	$x^{m+1} := x^m + \lambda_{\text{opt}}(r^m, p^m, A) A p^m$ with	(9.24b)
	$\lambda_{\text{opt}}(r^m, p^m, A) := \langle r^m, p^m \rangle / \langle A p^m, p^m \rangle$;	(9.24c)
	$r^{m+1} := r^m - \lambda_{\text{opt}}(r^m, p^m, A) A p^m$;	(9.24d)

The lines (9.24b,c) show that x^{m+1} is optimal with respect to the direction p^m : $F(x^{m+1}) = \min\{F(x^m + \lambda p^m) : \lambda \in \mathbb{K}\}$ or equivalently

$$r^{m+1} \perp p^m. \quad (9.24e)$$

Definition (9.24d) is equivalent to $r^{m+1} := b - Ax^{m+1}$.

The properties of this method are collected below.

Theorem 9.27. (a) The directions $\{p^m : 0 \leq m \leq n-1\}$ form a basis of pairwise conjugate vectors, i.e., an A -orthogonal basis.

(b) The algorithm terminates at $m = n-1$ with the exact solution $x^{m+1} = x^n = x^*$.

(c) The iterate x^m is optimal with respect to all directions p^0, \dots, p^{m-1} , i.e., it is optimal with respect to the subspace $\mathcal{U}_m := \text{span}\{p^0, p^1, \dots, p^{m-1}\}$. The residuals r^m satisfy

$$r^m \perp p^\ell \quad (0 \leq \ell \leq m-1), \quad (9.25a)$$

$$r^m \perp \mathcal{U}_\ell \quad (1 \leq \ell \leq m). \quad (9.25b)$$

(d) The error $e^m = x^m - x^*$ fulfills the conditions

$$e^m \perp_A p^\ell \quad (0 \leq \ell \leq m-1). \quad (9.25c)$$

(e) x^m solves the minimisation problem

$$F(x^m) = \min_{\lambda_\ell \in \mathbb{K}} \left\{ F(\xi) : \xi = x^0 + \sum_{\ell=0}^{m-1} \lambda_\ell p^\ell \right\} = \min_{\xi - x^0 \in \mathcal{U}_m} F(\xi), \quad (9.25d)$$

where the minimum in (9.25d) is taken at $\lambda_\ell = \lambda_{\text{opt}}(r^\ell, p^\ell, A)$.

Proof. (a) First, we note that the division by $\langle Ap^m, p^m \rangle$ in (9.24c) is well defined because of $p^m \neq 0$, as long as an additional conjugate direction exists, i.e., as long as $m < n$. As soon as $m = n-1$, the vectors p^0, \dots, p^{n-1} span the whole space \mathbb{K}^I and the process cannot be continued.

(c) The statement (9.25a) is true for $m = 0$ since $\{\ell : 0 \leq \ell \leq m-1\}$ is the empty set. Suppose that (9.25a) holds for m . By Lemma 9.23, x^m is optimal with respect to all directions p^ℓ ($0 \leq \ell \leq m-1$). According to Lemma 9.26, this property is inherited by x^{m+1} because of $p^m \perp_A p^\ell$ ($0 \leq \ell \leq m-1$); hence $r^{m+1} \perp p^\ell$ holds for all $0 \leq \ell \leq m-1$. The missing condition $r^{m+1} \perp p^m$ follows from (9.24e).

(d) (9.25c) follows from (9.25a), as $Ae^m = A(x^m - x^*) = Ax^m - b = -r^m$.

(b) (9.25b) proves that $r^n \perp \mathcal{U}_n$. Since $\mathcal{U}_n = \mathbb{K}^I$ (cf. part (a)), $r_n = 0$ follows, i.e., $x^n = x^*$.

(e) Inserting Eqs. (9.24b) one into another, we obtain

$$x^m = x^0 + \sum_{\ell=0}^{m-1} a_\ell p^\ell \quad \text{with } a_\ell = \lambda_{\text{opt}}(r^\ell, p^\ell, A).$$

From (9.9c) with $\tilde{x} := x^m$, $x := \xi$, and from $r^m \perp \mathcal{U}_m$, we deduce that

$$\begin{aligned} F(\xi) - F(x^m) &= \Re \left\langle r^m, \sum_{\ell=0}^{m-1} (\lambda_\ell - a_\ell) p^\ell \right\rangle + \frac{1}{2} \langle A(\xi - x^m), \xi - x^m \rangle \\ &= \frac{1}{2} \|\xi - x^m\|_A^2 \geq 0 \end{aligned}$$

with an equal sign only for $\xi = x^m$, i.e., for $\lambda_\ell = a_\ell$. This proves (9.25d). \square

The method of conjugate directions is not interesting in practice, unless the directions p^m in (9.24) are suitably selected. If, for instance, one chooses a fixed conjugate system $\{p^0, \dots, p^{n-1}\}$, the starting value $x^0 := x^* - p^{n-1}$ with the residual $r^0 = Ap^{n-1}$ leads to a sequence $x^0 = x^1 = \dots = x^{n-1}$ which only in the last step changes to the exact solution $x^n = x^*$. This explains why, in general, no convergence estimate as in (9.13b) can be given.

9.4 Minimal Residual Iteration

For general matrices A , the function (9.7a) with $H = I$ can be minimised, i.e., the residual $r = b - Ax$ is minimised: $F(x) := \|A(x - x^*)\|_2^2 = \|r\|_2^2 = \min$. Choosing the gradient of F as search direction, we would regain the gradient method in §9.2 applied to the equation $A^\mathsf{H}Ax = A^\mathsf{H}b$. Instead of this gradient, one can use the residual $r = b - Ax$ of the original system as search direction. This yields the *minimal residual iteration*

$$x^{m+1} = x^m - \frac{\Re \langle Ar^m, r^m \rangle}{\langle Ar^m, Ar^m \rangle} r^m, \quad r^m = b - Ax^m.$$

For general matrices A , the method cannot converge since $r^0 \neq 0$ may lead to $\langle Ar^0, r^0 \rangle = 0$ so that $x^m = x^0 \neq A^{-1}b$ for all m . To avoid this problem, we need the following assumptions.

Theorem 9.28. Assume $A + A^\mathsf{H} > 0$. Then the minimal residual iteration converges with the rate

$$c := \sqrt{\frac{\lambda_{\min}(A + A^\mathsf{H})}{2 \|A\|_2}}. \quad (9.26)$$

The convergence is uniform with respect to the residual: $\|r^{m+1}\|_2 \leq c \|r^m\|_2$.

Proof. See Saad [328, Theorem 5.10]. □

Chapter 10

Conjugate Gradient Methods and Generalisations

Abstract The conjugate gradient method is the best-known semi-iteration. Consuming only a small computational overhead, it is able to accelerate the underlying iteration. However, its use is restricted to positive definite matrices and positive definite iterations. There are several generalisations to the Hermitian and to the general case. In Section 10.1 we introduce the general concept of the required orthogonality conditions and the possible connection to minimisation principles. The standard conjugate gradient method is discussed in Section 10.2. The method of conjugate residuals introduced in Section 10.3 applies to Hermitian but possibly indefinite matrices. The method of orthogonal directions described in Section 10.4 also applies to general Hermitian matrices. General nonsymmetric problems are treated in Section 10.5. The generalised minimal residual method (GMRES; cf. §10.5.1), the full orthogonalisation method (cf. §10.5.2), and the biconjugate gradient method and its variants (cf. §10.5.3) are discussed.

10.1 Preparatory Considerations

In the following $x^* := A^{-1}b$ denotes the exact solution, while $x \in \mathbb{K}^I$ may be used as a variable. The iterate x^m is associated with the error $e^m = x^m - x^*$ and the residual $r^m = b - Ax^m = -Ae^m$.

10.1.1 Characterisation by Orthogonality

As seen in Conclusion 8.13a, the semi-iterates (8.3) belong to the affine space $x^0 + N\mathcal{K}_m(AN, r^0) = x^0 + \mathcal{K}_m(NA, Nr^0)$. In the following, we replace the Krylov space $\mathcal{K}_m(NA, Nr^0)$ by a general subspace

$$\mathcal{U}_m \subset \mathbb{K}^I \quad \text{with} \quad \dim(\mathcal{U}_m) = m. \quad (10.1a)$$

The reason for using \mathcal{U}_m is that the following arguments are independent of the special nature of the Krylov space. We are looking for candidates

$$x^m \in x^0 + \mathcal{U}_m. \quad (10.1b)$$

The second space

$$\mathcal{V}_m \subset \mathbb{K}^I \quad \text{with} \quad \dim(\mathcal{V}_m) = m$$

may coincide with \mathcal{U}_m .

For the practical implementation, we use bases

$$\mathcal{U}_m = \text{span}\{u^1, \dots, u^m\}, \quad \mathcal{V}_m = \text{span}\{v^1, \dots, v^m\}. \quad (10.1c)$$

Remark 10.1. (a) The spaces are *nested* if

$$\mathcal{U}_1 \subset \dots \subset \mathcal{U}_m \subset \mathcal{U}_{m+1} \subset \dots . \quad (10.2)$$

In this case, it is advantageous if the basis vectors u^1, \dots, u^m of \mathcal{U}_m coincide with the first m basis vectors of \mathcal{U}_{m+1} .

(b) For stability reasons, orthonormal bases are a good choice. In the case of $\mathcal{U}_m \subset \mathcal{U}_{m+1}$, $u^{m+1} \in \mathcal{U}_{m+1}$ is the normalised vector with $u^{m+1} \perp \mathcal{U}_m$.

The following methods are directly or indirectly characterised by the condition that the m -th iterate x^m fulfills an orthogonality condition:

$$x^m \text{ satisfies (10.1b) and } r^m := b - Ax^m \perp \mathcal{V}_m. \quad (10.3)$$

The questions that arise are:

1. Is (10.3) uniquely solvable?
2. Can we derive estimates for the error e^m in some norm?
3. How costly is the solution of (10.3)?

The first question will be answered in §10.1.2 and the second in §10.1.5. The cost is discussed later for the concrete choice of spaces.

Remark 10.2. Condition (10.3) is equivalent to

$$\langle r^m, v^i \rangle = 0 \quad \text{for all } 1 \leq i \leq m \quad (10.4)$$

(cf. (10.1c)). A generalisation of (10.3) could be $\langle r^m, v^i \rangle_X = 0$ using another scalar product $\langle u, v \rangle_X := \langle Xu, v \rangle$ for some $X > 0$ (cf. Remark C.10). However, this approach is identical to (10.3) with \mathcal{V}_m replaced with $X\mathcal{V}_m$.

10.1.2 Solvability

As required in (2.2), we always assume that the underlying matrix A of the system $Ax = b$ is regular.

The basis (10.1c) of \mathcal{U}_m allows us to make the ansatz $x^m = x^0 + \sum_{j=1}^m a_j u^j$. This implies that $r^m = r^0 - \sum_{j=1}^m a_j A u^j$. The conditions in (10.4) yield the system

$$Za = z \quad \text{with} \quad Z_{ij} := \langle Au^j, v^i \rangle, \quad z_i := \langle r^0, v^i \rangle. \quad (10.5)$$

In general, there is no guarantee that Z is regular. If $m \leq \#I/2$, the orthogonal situation $A\mathcal{U}_m \perp \mathcal{V}_m$ is possible and yields the extreme case of $Z = 0$.

Remark 10.3. The regularity of Z is equivalent to either of the conditions

$$(A\mathcal{U}_m)^\perp \cap \mathcal{V}_m = \mathcal{U}_m^\perp \cap A^\mathsf{H} \mathcal{V}_m = A\mathcal{U}_m \cap \mathcal{V}_m^\perp = \mathcal{U}_m \cap A^\mathsf{H} \mathcal{V}_m^\perp = \{0\}.$$

It remains to formulate sufficient conditions ensuring the regularity of Z .

- Criterion 10.4.** (a) Let $\mathcal{U}_m = \mathcal{V}_m$ and assume $A + A^\mathsf{H} > 0$. Then Z is regular.
 (b) If $\mathbb{K} = \mathbb{C}$, the previous condition may be replaced with $\mathfrak{i}(A^\mathsf{H} - A) > 0$.
 (c) $\mathcal{U}_m = \mathcal{V}_m$ and $A > 0$ are sufficient.
 (d) For $N > 0$ and a general regular matrix A , the choice of $\mathcal{V}_m = NA\mathcal{U}_m$ ensures regularity of Z .

Proof. (a) If Z is singular, there is some $0 \neq a \in \mathbb{K}^m$ with $Za = 0$, i.e., $Au \perp \mathcal{V}_m$ for $u := \sum_{j=1}^m a_j u^j \neq 0$. This is a contradiction to $0 < \langle (A + A^\mathsf{H})u, u \rangle = 2\Re\langle Au, u \rangle$, since $u \in \mathcal{V}_m$.

(b) $\frac{1}{\mathfrak{i}}(A - A^\mathsf{H}) > 0$ implies that $0 < 2\Im\langle Au, u \rangle$. Part (c) is trivial.

(d) Without loss of generality, the basis of \mathcal{V}_m can be defined by $v^i = NAu^i$. Then $Z_{ij} = \langle Au^j, v^i \rangle = \langle A^\mathsf{H} NAu^j, v^i \rangle$ corresponds to case (c) with A replaced by $A^\mathsf{H} NA > 0$. \square

10.1.3 Galerkin and Petrov–Galerkin Methods

Appendix E describes the discretisation of boundary value problems by the Galerkin method. This method can also be applied to finite-dimensional problems. The system $Ax = b$ ($x, b \in \mathbb{K}^I$) can be rewritten as the variation problem

$$\langle Ax, v \rangle = \langle b, v \rangle \quad \text{for all } v \in \mathbb{K}^I.$$

Using the initial value x^0 , we write $x = x^0 + u$ so that

$$\langle Au, v \rangle = \langle r^0, v \rangle \quad \text{for all } v \in \mathbb{K}^I \text{ with } r^0 = b - Ax^0.$$

The Galerkin method replaces this problem by a system of lower dimension m . Let \mathcal{U}_m be the subspace in (10.1a). Then the Galerkin solution $x^m \in x^0 + \mathcal{U}_m$ is defined by $x^m = x^0 + u$ with

$$u \in \mathcal{U}_m \text{ satisfying } \langle Au, v \rangle = \langle r^0, v \rangle \quad \text{for all } v \in \mathcal{U}_m.$$

Obviously this problem is equivalent to $r^m \perp \mathcal{U}_m$ and therefore to the condition (10.3) with $\mathcal{V}_m = \mathcal{U}_m$.

The coercivity formulated in (E.3) requires $A + A^H \geq \frac{1}{C}I$ for some $C > 0$. In the finite-dimensional case, this is equivalent to $A + A^H > 0$ as in Criterion 10.4a.

The more general Petrov–Galerkin method in Definition E.7 yields the problem

$$\text{find } x^m \in x^0 + u, \quad u \in \mathcal{U}_m \text{ with } \langle Au, v \rangle = \langle r^0, v \rangle \quad \text{for all } v \in \mathcal{V}_m,$$

where now \mathcal{V}_m may be different from \mathcal{U}_m . This yields the general condition (10.3).

10.1.4 Minimisation

The orthogonality condition (10.3) may be a consequence of another formulation. If $A > 0$, the Galerkin formulation is the first variation of the minimisation problem (9.2): $F(x) = \frac{1}{2} \langle Ax, x \rangle - \langle b, x \rangle = \min$.

The most general quadratic form whose minimum is the solution of $Ax = b$, is described in Lemma 9.3: $F(x) = \frac{1}{2} \|H^{1/2}r\|_2^2 + c$ with $r = b - Ax$ and $H > 0$. Its first variation leads us to the case (d) in Criterion 10.4 with $N = H$.

10.1.5 Error Statements

Even if the auxiliary system in (10.5) is solvable, there is no guarantee that the quality of x^m improves with increasing m . Nevertheless, if the method makes sense for all $m \leq \#I$, we reach the exact solution, provided that the arithmetic is exact.

Remark 10.5. Let $n := \#I$. (a) The iterate x^n is the exact solution: $x^n = A^{-1}b$. (b) If $r^m \in \mathcal{V}_m$ for some $m \leq n$, then x^m is the exact solution.

Proof. By definition, $r^m \perp \mathcal{V}_m$ holds. Combining this statement with $r^m \in \mathcal{V}_m$ yields $r^m = 0$, i.e., $x^m = A^{-1}b$. This proves part (b). Since $\mathcal{V}_n = \mathbb{K}^I$ because of $\dim(\mathcal{V}_n) = n$, part (b) applies. \square

Error estimates can be based on an underlying minimisation problem, provided that condition (10.3) is the result of an optimisation problem. The formulation of the optimisation problem also defines the norm for measuring the error.

10.1.5.1 Energy Norm

Assume a positive definite matrix $A > 0$ so that the energy norm $\|\cdot\|_A$ can be defined (cf. (C.5a)). The Galerkin formulation in a subspace $\mathcal{U}_m = \mathcal{V}_m$ determines the minimiser of $F(x) = \frac{1}{2} \langle Ax, x \rangle - \langle b, x \rangle$ in $x^0 + \mathcal{U}_m$, i.e.,

$$\|x^m - x^*\|_A = \min \left\{ \|x - x^*\|_A : x \in x^0 + \mathcal{U}_m \right\}. \quad (10.6)$$

If the spaces are nested (cf. (10.2)), the norm $\|x^m - x^*\|_A$ decreases weakly with increasing m . This statement also holds for minimisation later on.

In the case of $\Phi(\cdot, \cdot, A) \in \mathcal{L}_{>0}$, i.e., $NA > 0$, the minimisation in (10.6) uses the norm $\|\cdot\|_{NA}$ instead of $\|\cdot\|_A$.

The classical CG method in §10.2 will lead us to (10.6) with $\mathcal{U}_m = \mathcal{K}_m(A, r^0)$.

10.1.5.2 Residual Norm

The norm $\|x^m - x^*\|_A$ coincides with $\|A(x^m - x^*)\|_2 = \|r^m\|_2$. The minimisation of the residual

$$\|r^m\|_2 = \min \left\{ \|A(x - x^*)\|_2 : x \in x^0 + \mathcal{U}_m \right\} \quad (10.7)$$

implies the orthogonality

$$r^m \perp A\mathcal{U}_m =: \mathcal{V}_m. \quad (10.8)$$

The latter statement can also be written as $A^\top r^m \perp \mathcal{U}_m$. For general matrices, the use of Krylov subspaces leads us to the GMRES method in §10.5.1.

The minimisation of $F(x) = \frac{1}{2} \|N^{1/2}r\|_2^2 + c$ for some $N > 0$ (cf. §10.1.4) generalises (10.7) to

$$\|N^{1/2}r^m\|_2 = \min \left\{ \|N^{1/2}A(x - x^*)\|_2 : x \in x^0 + \mathcal{U}_m \right\}.$$

In the case of Hermitian matrices A , the realisation with Krylov spaces is given in §10.3.

One must be aware of the fact that a small residual $\|r^m\|_2$ does not necessarily imply that the error $\|x^m - x^*\|_2$ is small (cf. Remark 2.35).

10.1.5.3 Euclidean Norm

The first idea may be to approximate the solution x of $Ax = b$ by the best approximation x^* in $x^0 + \mathcal{U}_m$:

$$\|x^m - x^*\|_2 = \min \left\{ \|x - x^*\|_2 : x \in x^0 + \mathcal{U}_m \right\}.$$

The first variation yields the orthogonality condition $e^m \perp \mathcal{U}_m$. In terms of the condition (10.3), this can be written as

$$r^m \perp A^{-H} \mathcal{U}_m. \quad (10.9)$$

However, in general, this problem is not feasible. The cost for computing x^m is at least as high as solving the system $Ax = b$. For instance, $x^1 = x^0 + \alpha u$ (u normalised vector with $\mathcal{U}_1 = \text{span}\{u\}$) is the minimiser if $\alpha = -\langle e^0, u \rangle$. However, $e^0 = x^0 - x^*$ is not available unless the exact solution x^* is known. Therefore, evaluating the scalar product $\langle e^0, u \rangle$ causes a problem.

Nevertheless, the problem becomes solvable if the subspace \mathcal{U}_m can be written as $\mathcal{U}_m = A^H \mathcal{V}_m$ and a basis $\{v^1, \dots, v^m\}$ of \mathcal{V}_m is known. Then the basis of \mathcal{U}_m can be chosen as $\{u^1, \dots, u^m\}$ with $u^j := A^H v^j$. In this case, condition (10.9) becomes $r^m \perp \mathcal{V}_m$.

For $A = A^H$ and Krylov spaces \mathcal{V}_m , this approach is realised by the method of orthogonal directions in §10.4.

10.2 Conjugate Gradient Method

Concerning books on Krylov methods we refer, e.g., to Greenbaum [167, Part I], Liesen–Strakos [265], Meurant [283], Saad [328], Stoer [355], and van der Vorst [373, §§5–12]. The history is described by Golub–O’Leary [155].

10.2.1 First Formulation

In the following, the gradient method and the conjugate directions in §§9.2–9.3 will be combined. In order not to lose optimality with respect to the previous search directions, we only permit conjugate directions. The residuals (negative gradients) are used to determine the search direction p^m in (9.24). As for the gradient method we assume

$$A > 0 \quad \text{and} \quad F(x) = \frac{1}{2} \langle Ax, x \rangle - \langle b, x \rangle.$$

After constructing (linearly independent) p^0, p^1, \dots, p^{m-1} , we can orthogonalise r^m with respect to the energy scalar product $\langle \cdot, \cdot \rangle_A$ (cf. Remark A.26a):

$$p^m := r^m - \sum_{\ell=0}^{m-1} \frac{\langle Ar^m, p^\ell \rangle}{\langle Ap^\ell, p^\ell \rangle} p^\ell, \quad (10.10a)$$

$$p^0 := r^0. \quad (10.10b)$$

Note that for $m = 0$ the empty sum in (10.10a) implies (10.10b).

Remark 10.6. (a) p^m in (10.10a) is conjugate to all p^ℓ with $0 \leq \ell \leq m-1$.

(b) The directions p^ℓ span the Krylov subspace

$$\mathcal{K}_m(A, r^0) = \text{span}\{p^0, \dots, p^{m-1}\} = \text{span}\{r^0, \dots, r^{m-1}\}. \quad (10.11a)$$

(c) Having constructed x^m and its residual r^m by the method of conjugate directions, the vectors r^m and p^m can only vanish simultaneously. This means that either $x^m = x^*$ is the exact solution or $p^m \neq 0$ holds.

(d) The residual is orthogonal to the preceding subspaces:

$$r^m \perp \mathcal{K}_\ell(A, r^0) \quad \text{for all } \ell \leq m. \quad (10.11b)$$

Proof. (a) By construction (10.10a), $\langle Ap^m, p^j \rangle = 0$ holds for $j < m$.

(b) Equation (10.11a) holds for $m = 1$. Let (10.11a) be valid for m . Definition (10.10a) implies the identity $\text{span}\{\mathcal{K}_m(A, r^0), p^m\} = \text{span}\{\mathcal{K}_m(A, r^0), r^m\}$ because of Exercise 8.8a and yields assertion (10.11a) for $m+1$.

(d) Repeat (9.25b) stated in Theorem 9.27.

(c) By (10.10a), $p^m = 0$ follows from $r^m = 0$. Assume the case of $p^m = 0$. (10.10a) shows that $r^m \in \mathcal{K}_m(A, r^0)$. On the other hand, $r^m \perp \mathcal{K}_m(A, r^0)$ holds (cf. (10.11b)). Both statements together imply that $r^m = 0$. \square

A first provisional representation of the conjugate gradient method reads as follows:

start: x^0 arbitrary; $r^0 := b - Ax^0$; (10.12a)

Loop over $m = 0, 1, \dots, n-1$: $(n := \#I)$

stop if $r^m = 0$, otherwise
compute p^m from r^m according to (10.10a,b) (10.12b)

$x^{m+1} := x^m + \lambda_{\text{opt}}(r^m, p^m, A) p^m$ with λ_{opt} in (9.24c); (10.12c)

$r^{m+1} := r^m - \lambda_{\text{opt}}(r^m, p^m, A) Ap^m$; (10.12d)

The properties of this method are summarised below.

Theorem 10.7. (a) Let m_0 be the value when the loop (10.12b–d) terminates with $r^{m_0} = 0$ and $x^{m_0} = x^*$. Assuming exact arithmetic, $m_0 = \deg_A(e^0) = \deg_A(r^0)$ holds. Since $m_0 \leq n := \#I$, the loop terminates latest after n steps.

(b) The iterates x^m ($0 \leq m \leq m_0$) can be characterised by each of the following minimisation problems:

$$F(x^m) = \min \left\{ F\left(x^0 + \sum_{\ell=0}^{m-1} \lambda_\ell p^\ell\right) : \lambda_0, \dots, \lambda_{m-1} \in \mathbb{K} \right\}, \quad (10.13a)$$

$$F(x^m) = \min \left\{ F\left(x^0 + \sum_{\ell=0}^{m-1} \mu_\ell r^\ell\right) : \mu_0, \dots, \mu_{m-1} \in \mathbb{K} \right\}, \quad (10.13b)$$

$$F(x^m) = \min \left\{ F\left(x^0 + p_{m-1}(A) r^0\right) : p_{m-1} \in \mathcal{P}_{m-1} \right\}. \quad (10.13c)$$

(c) The minima (10.13a–c) can also be expressed by the energy norm $\|\cdot\|_A$:

$$\begin{aligned} \|e^m\|_A &= \min_{\lambda_0, \dots, \lambda_{m-1} \in \mathbb{K}} \left\| e^0 + \sum_{\ell=0}^{m-1} \lambda_\ell p^\ell \right\|_A = \min_{\mu_0, \dots, \mu_{m-1} \in \mathbb{K}} \left\| e^0 + \sum_{\ell=0}^{m-1} \mu_\ell r^\ell \right\|_A \\ &= \min_{p_{m-1} \in \mathcal{P}_{m-1}} \|e^0 + p_{m-1}(A) r^0\|_A = \min_{\xi \in \mathcal{K}_m(A, r^0)} \|e^0 + \xi\|_A. \end{aligned} \quad (10.13d)$$

Proof. (b) Because of (10.11a), all minimisation problems (10.13a–c) are of the form

$$F(x^m) = \min\{F(x^0 + \xi) : \xi \in \mathcal{K}_m(A, r^0)\}.$$

This statement coincides with (10.11b) in Remark 10.6d.

(c) The equivalence of the statements in the parts (b) and (c) follows from (9.3): $F(x) = \|x - A^{-1}b\|_A^2 + \text{const.}$

(a) For $m^* = \deg_A(e^0)$, there is a polynomial $p = p_{m^*}$ of degree m^* with $p(A)e^0 = 0$. The scaling can be chosen so that $p(0) = 1$ (cf. Lemma 8.12). Define $q \in \mathcal{P}_{m^*-1}$ by $p(\xi) = 1 - q(\xi)\xi$. Since $0 = p(A)e^0 = e^0 - q(A)Ae^0 = e^0 + q(A)r^0$, the minimum in (10.13d) yields $e^{m^*} = 0$. This proves that the first $m = m_0$ with $e^{m_0} = 0$ satisfies $m_0 \leq m^*$. On the other hand, $e^{m_0} = 0$ and (10.13d) prove that there is some polynomial $p(\xi) = 1 - q(\xi)\xi$ of degree m_0 with $p(A)e^0 = 0$. Hence $m_0 \geq m^* = \deg_A(e^0)$. \square

The proposed algorithm (10.12a–d) can significantly be simplified in step (10.12b). Computing most of the scalar products $\langle Ar^m, p^\ell \rangle$ in (10.10a) can be avoided.

Lemma 10.8. $\langle Ar^m, p^\ell \rangle = 0$ holds for all $0 \leq \ell \leq m-2$, $m \leq m_0$.

Proof. We have $\langle Ar^m, p^\ell \rangle = \langle r^m, Ap^\ell \rangle$. Equation (10.11a) and inclusion (8.9) show that $Ap^\ell \in A\mathcal{K}_{\ell+1}(A, r^0) \subset \mathcal{K}_{\ell+2}(A, r^0) \subset \mathcal{K}_m(A, r^0)$. Therefore, the assertion follows from (10.11b): $r^m \perp \mathcal{K}_m(A, r^0)$. \square

Only the term for $\ell = m-1$ does remain in the sum (10.10a):

$$p^m := r^m - \frac{\langle Ar^m, p^{m-1} \rangle}{\langle Ap^{m-1}, p^{m-1} \rangle} p^{m-1} = r^m - \frac{\langle r^m, Ap^{m-1} \rangle}{\langle Ap^{m-1}, p^{m-1} \rangle} p^{m-1}. \quad (10.14)$$

The second representation in (10.14) has the advantage that only the product Ap^{m-1} is needed which already appears in the denominator, in λ_{opt} , and in (10.12d).

10.2.2 CG Method (Applied to Richardson's Iteration)

Using (10.14), we present the CG method (10.12a–d) in the following form ('CG' abbreviates 'conjugate gradient').

$\gamma_{\text{CG}}[\Phi_1^{\text{Rich}}]$	CG method (applied to Richardson's iteration)	(10.15)
start:	x^0 arbitrary; $r^0 := b - Ax^0$; $p^0 := r^0$;	(10.15a)
Loop over	$m = 0, 1, \dots, n-1$: stop if $r^m = 0$, otherwise:	
	$x^{m+1} := x^m + \lambda_{\text{opt}} p^m$ with	(10.15b)
	$\lambda_{\text{opt}} := \lambda_{\text{opt}}(r^m, p^m, A) = \langle r^m, p^m \rangle / \langle Ap^m, p^m \rangle$;	(10.15c)
	$r^{m+1} := r^m - \lambda_{\text{opt}} Ap^m$;	(10.15d)
	$p^{m+1} := r^{m+1} - \frac{\langle r^{m+1}, Ap^m \rangle}{\langle Ap^m, p^m \rangle} p^m$;	(10.15e)

Exercise 10.9. The following alternatives are equivalent to (10.15c,e):

$$\lambda_{\text{opt}}(r^m, p^m, A) = \|r^m\|_2^2 / \langle Ap^m, p^m \rangle, \quad (10.15c')$$

$$p^{m+1} = r^{m+1} + \frac{\|r^{m+1}\|_2^2}{\|r^m\|_2^2} p^m. \quad (10.15e')$$

Remark 10.10. One CG step $x^m \mapsto x^{m+1}$ requires one multiplication Ap^m and, in addition, only simple vector operations and scalar products. On the other hand, the storage requirement is higher. Besides x^m , also r^m and p^m are needed.

The CG method was first presented in 1952 by Stiefel [353] in a paper still worth reading. Independently, the method was described in the same year by Hestenes (cf. Hestenes [218], Hestenes–Stiefel [219]).

The CG method can be interpreted in two completely different ways:

- as a direct method,
- as an iterative method.

Formally, the CG algorithm is a direct method because it produces the exact solution x^* after finitely many operations (see m_0 in Theorem 10.7a). For the practical performance, this is not true. Since the later and smaller residuals r^m arise from linear combinations of larger quantities, cancellation leads to an error amplification, so that the vectors $\{p^0, \dots, p^{n-1}\}$ no longer form a conjugate system. After losing the orthogonality, two cases may appear:

- stagnation: the errors e^m fluctuate about the reached level of accuracy,
- instability: the errors start to grow again.

The first case is harmless, provided that the reached error level is sufficient. The second case will happen for many Krylov methods discussed later. This is the reason that there are many equivalent algorithms, i.e., algorithms producing identical results under exact arithmetic, but behaving differently under floating-point perturbation. In the best case, there are ‘stabilised’ versions which do not become unstable. There is a further, still more severe problem. Division by $\langle Ap^m, p^m \rangle$ already appears in (10.15c). A division by zero leads to a breakdown of the algorithm. A *lucky breakdown* happens if a vanishing divisor only appears if x^m is already the exact solution (so that the algorithm need not to be continued). In the ‘unlucky’ cases, the ‘stabilised’ versions should overcome this difficulty (cf. §10.3.3). In any case, one can state that the Krylov methods cannot be used as a practical method for the *direct* solution of large linear systems.

It was Reid [319] how emphasised the use of the CG method as an *iterative* method. Although the limit process $m \rightarrow \infty$ does not make sense,¹ the decrease of the error e^m in a range $0 \leq m \leq m_0$ with $m_0 \ll n = \#I$ is all we need for practical applications, provided that at least e^{m_0} is small enough.

The problem caused by floating-point perturbations suggests a modification towards an infinite CG iteration.² Assume that perturbations get out of control after (more than) k steps. Then, after every k steps (i.e., for $m = 0, k, 2k, \dots$), we start again with the last descent direction $p^k := r^k$, etc. This method is usually called the *restarted CG method*:

$$\begin{aligned} \text{start: } & x^0 \text{ arbitrary starting iterate,} \\ \text{iteration } m = 1, 2, \dots: & x^m: \text{as in (10.15b,d),} \\ & r^m, p^m: \text{as in (10.15c–e), if } m \text{ is not a multiple of } k; \\ & r^m := p^m := b - Ax^m, \quad \text{if } m = 0, k, 2k, \dots \end{aligned} \tag{10.16}$$

10.2.3 Convergence Analysis

The convergence analysis is based on the following observation corresponding to Remark 9.9 in the case of the gradient method. Property (10.17d) stated below coincides with the characterisation in §10.1.5.1.

Proposition 10.11. *Let x^0, \dots, x^{m_0} be the sequence of the CG iterates.*

(a) *The CG results can be regarded as the results of the semi-iterative Richardson iteration Φ_1^{Rich} . The related polynomials $p_k \in \mathcal{P}_k$ in (8.6c) with $p_k(1) = 1$ yield the error representation*

$$e^k = x^k - x^* = p_k(M_1^{\text{Rich}})e^0 = p_k(I - A)e^0 \quad (M_1^{\text{Rich}} = I - A). \tag{10.17a}$$

¹ The terms ‘convergence’ and ‘asymptotic convergence rate’ lose their meaning since no limit can be formed.

² This is still a nonlinear iteration. If $x^m = A^{-1}b$, the *lucky breakdown* stops the iteration.

(b) p_k and $q_{k-1}(\xi) := [p_k(1 - \xi) - 1]/\xi$ are the optimal polynomials solving the respective minimisation problems

$$\|e^k\|_A = \|p_k(M_1^{\text{Rich}})e^0\|_A \leq \|\tilde{p}_k(M_1^{\text{Rich}})e^0\|_A \quad (10.17b)$$

for all polynomials $\tilde{p}_k \in \mathcal{P}_k$ with $\tilde{p}_k(1) = 1$,

$$\|e^k\|_A = \|e^0 + q_{k-1}(A)r^0\|_A \leq \|e^0 + \tilde{q}_{k-1}(A)r^0\|_A \quad (10.17c)$$

for all polynomials $\tilde{q}_{k-1} \in \mathcal{P}_{k-1}$,

$$\|e^k\|_A = \min \left\{ \|x - x^*\|_A : x \in x^0 + \mathcal{K}_k(A)r^0 \right\}. \quad (10.17d)$$

Proof. (a) (10.15b) shows that $x^k = x^0 + \sum_{\nu=0}^{k-1} \beta_\nu p^\nu$ with $\beta_\nu := \lambda_{\text{opt}}(r^\nu, p^\nu, A)$, i.e.,

$$x^k - x^0 = e^k - e^0 \in \text{span}\{p^0, \dots, p^{k-1}\} = \mathcal{K}_k(A, r^0) \quad (\text{cf. (10.11a)}).$$

Hence, there is a polynomial $q_{k-1} \in \mathcal{P}_{k-1}$ with $e^k = e^0 - q_{k-1}(A)r^0$. Since $r^0 = -Ae^0$, $e^k = \hat{p}_k(A)e^0$ holds for the polynomial $\hat{p}_k(\xi) := 1 + \xi q_{k-1}(\xi) \in \mathcal{P}_k$. The related polynomial $p_k(\xi) := \hat{p}_k(1 - \xi)$ satisfies the consistency condition $p_k(1) = \hat{p}_k(0) = 1$. The identity $p_k(M_1^{\text{Rich}})e^0 = p_k(I - A)e^0 = \hat{p}_k(A)e^0 = e^k$ proves that $p_k \in \mathcal{P}_k$ is the polynomial in (10.17a).

(b) Since the CG results satisfy (10.13d), the polynomial q_{k-1} is the minimiser in (10.17c). Problem (10.17b) is equivalent to (10.17c) and (10.17d). \square

Remark 10.12. The CG iterates x^m are not the solutions of the minimisation problem posed in §8.3.1 because there the minimisation is required with respect to the Euclidean norm $\|\cdot\|_2$. However, if $\|\cdot\|_2$ is replaced by $\|\cdot\|_A$, the CG method offers the possibility of solving the modified minimisation problem $\|p_m(M)e^0\|_A = \min$ without knowledge of the initial error e^0 and the spectrum of $M_1^{\text{Rich}} = I - A$ (equivalently, of the spectrum of A).

Remark 10.13. For any polynomial $P_m \in \mathcal{P}_m$ satisfying $P_m(1) = 1$, the errors $e^m = x^m - x^*$ of the CG iterates satisfy the error estimate

$$\|e^m\|_A \leq \max \{|P_m(1 - \lambda)| : \lambda \in \sigma(A)\} \|e^0\|_A. \quad (10.18)$$

Proof. (10.17b) shows that $\|e^m\|_A \leq \|P_m(I - A)\|_A \|e^0\|_A$. The matrix norm $\|\cdot\|_A$ has the representation $\|X\|_A = \|A^{1/2}XA^{-1/2}\|_2$ (cf. (C.5d)). $A^{1/2}$ commutes with polynomials in A : $A^{1/2}P_m(I - A)A^{-1/2} = P_m(I - A)$. The assertion (10.18) follows from $\|P_m(I - A)\|_2 = \max\{|P_m(1 - \lambda)| : \lambda \in \sigma(A)\}$. \square

The following theorem shows that—as in the case of the Chebyshev method—an order improvement can be achieved.

Theorem 10.14. Let A be positive definite with $\lambda := \lambda_{\min}(A)$, $\Lambda := \lambda_{\max}(A)$ and abbreviate the spectral condition number by $\kappa = \kappa(A) = \Lambda/\lambda$. The errors e^m of the CG iterates x^m satisfy the estimate

$$\|e^m\|_A \leq \frac{2 \left(1 - \frac{1}{\kappa}\right)^m}{\left(1 + \frac{1}{\sqrt{\kappa}}\right)^{2m} + \left(1 - \frac{1}{\sqrt{\kappa}}\right)^{2m}} \|e^0\|_A = \frac{2c^m}{1 + c^{2m}} \|e^0\|_A \quad (10.19)$$

$$\text{with } c := \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} = \frac{\sqrt{\Lambda} - \sqrt{\lambda}}{\sqrt{\Lambda} + \sqrt{\lambda}}.$$

Proof. Let P_m be the transformed Chebyshev polynomial (8.27a) belonging to $\sigma_M := [a, b] \supset \sigma(M^{\text{Rich}}) = \sigma(I - A)$ with $a = 1 - \Lambda$ and $b = 1 - \lambda$. (10.18) and (8.27b) yield $\|e^m\|_A \leq \|e^0\|_A / C_m$. (8.28c) proves (10.19). \square

The error estimate (10.19) uses an upper bound that may be too pessimistic. It is based on the Chebyshev polynomial P_m which is the optimal choice for minimising $\max\{|P_m(\xi)| : \xi \in \sigma_M = [a, b]\}$, but not necessarily for minimising $\max\{|P_m(\xi)| : \xi \in \sigma(M^{\text{Rich}}) = \sigma(I - A)\} = \max\{|P_m(1 - \lambda)| : \lambda \in \sigma(A)\}$. This leads to the following statement.

Remark 10.15. Although the asymptotic convergence rate of the gradient method depends exclusively on the spectral condition number $\kappa(A)$ and therefore the extreme eigenvalues, the convergence of the CG method is influenced by the whole spectrum.

The following simple example will illustrate this fact. Assume that the inclusion $\sigma(M^{\text{Rich}}) \subset [a, b]$ with $a = 1 - \Lambda$, $b = 1 - \lambda$ can be strengthened to $\sigma(M^{\text{Rich}}) \subset \sigma_M := [a, a'] \cup [b', b]$ with $a \leq a' < b' \leq b$. Then one may find a polynomial P_m for which $\max\{|P_m(1 - \lambda)| : \lambda \in \sigma(A)\}$ is smaller than for the Chebyshev polynomial (cf. §8.3.6). Hence, P_m yields a better estimate than (10.19). Generally speaking, if the eigenvalues of A are not distributed uniformly over $[\lambda, \Lambda]$ (e.g., if they accumulate in smaller subintervals), the CG method converges better than estimated by (10.19).

Exercise 10.16. If the spectrum $\sigma(M^{\text{Rich}}) = \{\lambda, \Lambda\}$ contains only the extreme eigenvalues λ and Λ , the cg method yields $x^{m_0} = x^*$ for $m_0 \leq 2$.

Even if the eigenvalue distribution permits no better polynomial than the Chebyshev polynomial, the ratios $\|e^{m+1}\|_A / \|e^m\|_A$ improve with increasing iteration number m and become smaller than $c \approx 1 - 2/\sqrt{\kappa}$ in (10.19). The reason is as follows. In the case of the gradient method (9.11a–c), the error e^m converges to the subspace $V := \text{span}\{v_1, v_2\}$ spanned by the eigenvectors belonging to $\lambda := \lambda_{\min}(A)$ and $\Lambda := \lambda_{\max}(A)$ (see the proof of Corollary 9.11). For the CG case, this behaviour cannot occur. If the CG error e^m lies exactly in the subspace V , 2 = $\dim V$ steps of the CG methods would be sufficient to obtain $e^{m+2} = 0$. It can be proved that the CG error moves towards V^\perp . Restricting the matrix A to V^\perp , we obtain the spectrum $\sigma(A) \setminus \{\lambda, \Lambda\}$ and the condition is Λ_2 / λ_2 , where λ_2 is the second smallest and Λ_2 the second largest eigenvalue. Hence, after a certain number of steps, the error ratios behave more like $c \approx 1 - 2/\sqrt{\Lambda_2 / \lambda_2} < c$. A precise analysis of this superconvergence phenomenon is given by van der Sluis–van der Vorst [370]. See also Strakos [357].

10.2.4 CG Method Applied to Positive Definite Iterations

10.2.4.1 Standard Version

As the gradient method, the method of conjugate gradients can be applied to other positive definite iterations than the Richardson method. This yields the so-called *preconditioned CG method* (but notice that the gradients are preconditioned not the CG method). Assume $\Phi \in \mathcal{L}_{\text{pos}}$. Hence, the standard assumption $A > 0$ implies $N > 0$ for the matrix $N = N[\Phi]$ in

$$x^{m+1} = x^m - N(Ax^m - b) \quad \text{with } A, N \text{ positive definite.} \quad (10.20a)$$

As in (9.15b), we introduce $\check{A} := N^{1/2}AN^{1/2}$ and $\check{b} := N^{1/2}b$. Algorithm (10.20a) is equivalent to the Richardson iteration (10.20b) for solving $\check{A}\check{x} = \check{b}$:

$$\check{x}^{m+1} = \check{x}^m - (\check{A}\check{x}^m - \check{b}). \quad (10.20b)$$

Applying the CG algorithm (10.15a–e) to $\check{A}\check{x} = \check{b}$, we obtain:

$$\text{start: } \check{x}^0 := N^{-1/2}x^0; \quad \check{r}^0 := \check{b} - \check{A}\check{x}^0; \quad \check{p}^0 := \check{r}^0; \quad (10.21a)$$

for $m = 0, 1, 2, \dots$ (while $\check{r}^m \neq 0$):

$$\check{x}^{m+1} := \check{x}^m + \lambda_{\text{opt}} \check{p}^m \text{ with} \quad (10.21b)$$

$$\lambda_{\text{opt}} := \lambda_{\text{opt}}(\check{r}^m, \check{p}^m, \check{A}) = \langle \check{r}^m, \check{p}^m \rangle / \langle \check{A}\check{p}^m, \check{p}^m \rangle; \quad (10.21c)$$

$$\check{r}^{m+1} := \check{r}^m - \lambda_{\text{opt}} \check{A} \check{p}^m \quad (= \check{b} - \check{A}\check{x}^{m+1}); \quad (10.21d)$$

$$\check{p}^{m+1} := \check{r}^{m+1} - \langle \check{r}^{m+1}, \check{A}\check{p}^m \rangle / \langle \check{A}\check{p}^m, \check{p}^m \rangle \check{p}^m; \quad (10.21e)$$

Insert $\check{A} = N^{1/2}AN^{1/2}$ and $\check{b} = N^{1/2}b$, define x^m and p^m by

$$\check{x}^m = N^{-1/2}x^m, \quad \check{p}^m = N^{-1/2}p^m \quad (10.21f)$$

and use $N^{1/2}r^m = N^{1/2}(b - Ax^m) = \check{b} - \check{A}\check{x}^m = \check{r}^m$. (10.21a–e) becomes

$$\text{start: } x^0 \text{ arbitrary; } r^0 := b - Ax^0; \quad p^0 := Nr^0; \quad (10.22a)$$

iteration: for $m = 0, 1, 2, \dots$ (while $r^m \neq 0$):

$$x^{m+1} := x^m + \lambda_{\text{opt}} p^m \text{ with} \quad (10.22b)$$

$$\lambda_{\text{opt}} := \lambda_{\text{opt}}(r^m, p^m, A) = \langle r^m, p^m \rangle / \langle Ap^m, p^m \rangle; \quad (10.22c)$$

$$r^{m+1} := r^m - \lambda_{\text{opt}} Ap^m; \quad (10.22d)$$

$$p^{m+1} := Nr^{m+1} - \frac{\langle Nr^{m+1}, Ap^m \rangle}{\langle Ap^m, p^m \rangle} p^m; \quad (10.22e)$$

The expression (10.22c) coincides with the original definition (9.6a) of λ_{opt} . (10.22e) shows that the search directions p^m are produced from the ‘preconditioned’

gradient Nr^m by an A -orthogonalisation. Exploiting the equivalent formulations (10.15c',e'), we end up with

$$\begin{aligned}\lambda_{\text{opt}} &:= \langle Nr^m, r^m \rangle / \langle Ap^m, p^m \rangle; \\ p^{m+1} &:= Nr^{m+1} + \frac{\langle Nr^{m+1}, r^{m+1} \rangle}{\langle Nr^m, r^m \rangle} p^m.\end{aligned}$$

If one carries along the variables x^m , p^m , r^m , and $\rho_m := \langle Nr^m, r^m \rangle$ during the iteration, the CG algorithm $\Upsilon_{\text{CG}}[\Phi]$ takes the form (10.23a–f):

start:	x^0 arbitrary;	(10.23)
$r^0 := b - Ax^0$;	$p^0 := Nr^0$;	$\rho_0 := \langle p^0, r^0 \rangle$;
iteration: for $m = 0, 1, 2, \dots$ (while $m < n := \#I$ and $r^m \neq 0$):		
$a^m := Ap^m$;	$\lambda_{\text{opt}} := \rho_m / \langle a^m, p^m \rangle$;	(10.23b)
$x^{m+1} := x^m + \lambda_{\text{opt}} p^m$;		(10.23c)
$r^{m+1} := r^m - \lambda_{\text{opt}} a^m$;		(10.23d)
$q^{m+1} := Nr^{m+1}$;	$\rho_{m+1} := \langle q^{m+1}, r^{m+1} \rangle$;	(10.23e)
$p^{m+1} := q^{m+1} - \frac{\rho_{m+1}}{\rho_m} p^m$;		(10.23f)

The error estimate for $e^m = x^m - x^*$ follows as in §9.2.4, since the inequality (10.19) for $\check{e}^m = \check{x}^m - \check{x}^* = N^{-1/2}e^m$ can be transferred to e^m : $\|\check{e}^m\|_{\check{A}} = \|e^m\|_A$. Notice that $\kappa = \kappa(\check{A}) = \kappa(N^{1/2}AN^{1/2}) = \kappa(NA) = \Gamma/\gamma$ with Γ and γ in (10.24a).

Theorem 10.17 (error estimate). Assume $\Phi \in \mathcal{L}_{\text{pos}}$ and $A > 0$. The matrix $W = N^{-1}$ of the third normal form is assumed to satisfy

$$\gamma W \leq A \leq \Gamma W \quad (\gamma > 0, \text{ cf. (9.18a)}). \quad (10.24a)$$

Then the iterates x^m of the CG method $\Upsilon_{\text{CG}}[\Phi]$ in (10.23a–f) are the minimisers of $\min\{\|x - x^*\|_A : x = x^0 + \mathcal{K}_m(NA)Nr^0\}$ and fulfil the energy norm estimate

$$\|e^m\|_A \leq \frac{2c^m}{1 + c^{2m}} \|e^0\|_A \quad \text{with } c = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} = \frac{\sqrt{\Gamma} - \sqrt{\gamma}}{\sqrt{\Gamma} + \sqrt{\gamma}}, \quad \kappa = \frac{\Gamma}{\gamma}. \quad (10.24b)$$

Lemma 10.18. (a) $m_0 = \deg_{\check{A}}(\check{e}^0) = \deg_{NA}(e^0) = \deg_{AN}(r^0) \leq n = \#I$ is the first index m_0 with $r^{m_0} = 0$ and $x^{m_0} = x^*$.

(b) The search directions generated by (10.23a–f) are conjugate with respect to the original matrix A :

$$\langle p^k, p^\ell \rangle_A = 0 \quad \text{for } k \neq \ell. \quad (10.25)$$

(c) The statements in (10.11a,b) become

$$\begin{aligned}r^m \perp \mathcal{K}_m(NA, Nr^0) &= \text{span}\{p^0, \dots, p^{m-1}\} \\ &= \text{span}\{Nr^0, \dots, Nr^{m-1}\}.\end{aligned}$$

(d) The iterate x^m is the minimiser of the expressions

$$\begin{aligned} F(x^m) &= \min_{\lambda_0, \dots, \lambda_{m-1} \in \mathbb{K}} F\left(x^0 + \sum_{\ell=0}^{m-1} \lambda_\ell p^\ell\right) = \min_{\mu_0, \dots, \mu_{m-1} \in \mathbb{K}} F\left(x^0 + N \sum_{\ell=0}^{m-1} \mu_\ell r^\ell\right) \\ &= \min_{p_{m-1} \in \mathcal{P}_{m-1}} F(x^0 + p_{m-1}(NA)Nr^0) = \min_{\xi \in \mathcal{K}_m(NA, Nr^0)} F(x^0 + \xi). \end{aligned}$$

Proof. Part (a) is identical to Theorem 10.7a. Part (b) follows from (10.21f),

$$\begin{aligned} \langle \check{p}^k, \check{p}^\ell \rangle_{\check{A}} &= \langle \check{A}\check{p}^k, \check{p}^\ell \rangle = \left\langle N^{1/2} AN^{1/2} N^{-1/2} p^k, N^{-1/2} p^\ell \right\rangle = \langle Ap^k, p^\ell \rangle \\ &= \langle p^k, p^\ell \rangle_A \end{aligned}$$

and the \check{A} -orthogonality of the search directions \check{p}^k . Parts (c) and (d) are consequences of (10.13a–c) applied to the \vee -quantities in (10.21f). \square

The alternative reformulation $\bar{x}^{m+1} := \bar{x}^m - (\bar{A}\bar{x}^m - \bar{b})$ used in §9.2.4.2 will be discussed in §10.3.

10.2.4.2 Directly Positive Definite Case

Assume $\Phi \in \mathcal{L}_{>0}$, i.e., the iteration $\Phi(\cdot, \cdot, A)$ is directly positive definite: $N[A]A > 0$ (cf. Definition 5.14). Now we substitute A , x , b by $\hat{A} := NA$, $\hat{x} = x$, $\hat{b} = Nb$ in the CG algorithm (10.21a–e). Reformulation the algorithm in terms of the quantities A , x , b yields

$$\begin{aligned} \text{start: } \quad &x^0 \text{ arbitrary; } \quad r^0 := b - Ax^0; \quad p^0 := Nr^0; \quad m := 0; \\ \text{iteration: } \quad &x^{m+1} := x^m + \lambda_{\text{opt}} p^m \text{ with} \\ &\lambda_{\text{opt}} := \lambda_{\text{opt}}(Nr^m, p^m, NA) = \langle Nr^m, p^m \rangle / \langle NAp^m, p^m \rangle; \\ &r^{m+1} := r^m - \lambda_{\text{opt}} A p^m; \\ &p^{m+1} := Nr^{m+1} - \frac{\langle Nr^{m+1}, NAp^m \rangle}{\langle NAp^m, p^m \rangle} p^m; \end{aligned}$$

Proposition 10.19. (a) The final iteration number is $m_0 = \deg_{NA}(e^0)$.

(b) The directions p^m are NA -orthogonal.

(c) The transformed residuals are orthogonal:

$$Nr^m \perp \mathcal{K}_m(NA, Nr^0) = \text{span}\{p^0, \dots, p^{m-1}\} = \text{span}\{Nr^0, \dots, Nr^{m-1}\}.$$

(d) $\|e^m\|_{NA} \leq \frac{2c^m}{1+c^{2m}} \|e^0\|_{NA}$ holds with $c = \frac{\sqrt{\Gamma} - \sqrt{\gamma}}{\sqrt{\Gamma} + \sqrt{\gamma}}$, where γ and Γ are the minimal and maximal eigenvalues of NA .

10.2.5 Numerical Examples

We choose the Poisson model problem for $h = 1/32$. Applying the CG method to the Richardson iteration (i.e., algorithm $\Upsilon_{\text{CG}}[\Phi_1^{\text{Rich}}]$ in (10.15a–e)) yields the results given in Table 10.1. Due to inequality (10.19), the convergence factors $\|e^m\|_A/\|e^{m-1}\|_A$ measured with respect to the energy norm $\|\cdot\|_A$ should become smaller than $c = (\sqrt{\Lambda} - \sqrt{\lambda})/(\sqrt{\Lambda} + \sqrt{\lambda})$. Inserting the eigenvalues λ and Λ in (3.1b,c) for $h = 1/32$, we obtain $c = 0.9063471$. In fact, the convergence factor decreases from 0.9 to 0.66 when $m = 30$ increases to $m = 90$. This ‘superlinear’ convergence behaviour illustrates the improvement of the effective condition during the iteration as discussed in the last paragraph of §10.2.3.

Table 10.2 reports the CG results for $h = \frac{1}{32}$ with the SSOR and ILU iteration as basic iterations. The optimal SSOR parameter is the same as for Table 9.2. The ILU iteration is the modified five-point version ILU₅ with $\omega = -1$ and enlargement of the diagonal by 5 (cf. §7.3.10). The condition of the SSOR method determined in §9.2.5 is $\kappa \approx 7.66$. This yields the value $c \approx 0.47$ for c in (10.24b). In the SSOR case, the averaged convergence factors $(\|e^m\|_A/\|e^0\|_A)^{\frac{1}{m}}$

m	value in the middle	$\frac{\ e^m\ _A}{\ e^{m-1}\ _A}$
1	-0.00186560978	0.670874
2	-0.00460087980	0.791286
3	-0.00739241614	0.860663
4	-0.01111605755	0.865691
10	-0.04408187826	0.917138
20	-0.11796241337	0.939358
30	0.40673579950	0.918423
40	0.49137792828	0.843496
50	0.50013929834	0.832459
60	0.50010381735	0.738779
70	0.50001053720	0.761377
80	0.50000013936	0.708295
90	0.50000000342	0.661969
100	0.50000000001	

Table 10.1 Results of $\Upsilon_{\text{CG}}[\Phi_1^{\text{Rich}}]$ applied to the Poisson model problem with $h = 1/32$.

5-point ILU with $\omega = -1$		SSOR with $\omega = 1.8212691200$	
m	$u_{16,16}$	$u_{16,16}$	$\frac{\ e^m\ _A}{\ e^{m-1}\ _A}$
1	0.2262513522	0.0285107511	0.457624
2	0.5320480495	0.1146321025	0.307093
3	0.4582969109	0.2093879771	0.599140
4	0.4818928890	0.3500438579	0.530214
5	0.4827955876	0.4301535841	0.491911
10	0.4999129317	0.4992951874	0.464830
11	0.5000044282	0.4998541213	0.465082
12	0.4999850353	0.4999456258	0.394760
20	0.5000000033	0.5000000087	0.320139
21	0.5000000026	0.5000000020	0.487606
22	0.5000000008	0.5000000055	0.405755
23	0.5000000002	0.5000000041	0.408013
24	0.5000000000	0.5000000000	0.332715
25	0.5000000000	0.5000000005	0.432772
26	0.5000000000	0.5000000001	0.334264
27	0.5000000000	0.5000000001	0.366209
28	0.5000000000	0.5000000000	0.365471
29	0.5000000000	0.5000000000	0.487797
30	0.5000000000	0.5000000000	0.776690

Table 10.2 The CG method (10.21a–e) applied to the ILU and SSOR iterations.

are around 0.47 until $m = 11$. Afterwards they decrease to 0.42 for $m \approx 30$. The values $u_{16,16}$ ('value in the middle') given in Table 10.2 show that for $m \geq 27$ the rounding errors acquire the upper hand. Nevertheless, the CG algorithm is stable.

The superlinear convergence behaviour mentioned in connection with Table 10.1 should not be overrated. Its advantage can be exploited only if m becomes sufficiently large. In the case of Table 10.1, 'sufficiently large' means $m \geq 30$; in the SSOR case of Table 10.2, it is $m \geq 17$. Inspecting the values in these tables illustrates the following dilemma:

- Either the iteration is fast (as in Table 10.2). Then one would like to stop the iteration before reaching the critical value of m indicating the appearance of superconvergence.
- Or the iteration is slow (as in Table 10.1). Then one would prefer to replace the iteration Φ with a better one.

10.2.6 Amount of Work of the CG Method

One iteration step (10.23b–f) requires one evaluation of $p \mapsto Ap$ and $r \mapsto Nr$, three vector additions, three multiplications of a vector by a scalar number, and two scalar products. This adds up to

$$\text{CG-Work}(\Phi) = C(A) + C(N) + 8n$$

arithmetic operations for $\Upsilon_{\text{CG}}[\Phi]$, where

$$C(A): \text{work for } p \mapsto Ap, \quad C(N): \text{work for } r \mapsto Nr.$$

Performing the Φ -iteration step in the form $\Phi(x, b) = x - N(Ax - b)$, we need $C(A) + C(N) + 2n$ operations, so that

$$\text{CG-Work}(\Phi) = \text{Work}(\Phi) + 6n.$$

Hence, as in the semi-iterative case (cf. §8.3.9), the cost factor is equal to

$$C_{\Phi,\text{cg}} = C_\Phi + 6/C_A.$$

According to the analysis of convergence behaviour discussed above, we choose $c = (\sqrt{\Lambda} - \sqrt{\lambda})/(\sqrt{\Lambda} + \sqrt{\lambda})$ in (10.24b) as the asymptotic rate on which we base the effective amount of work:

$$\text{Eff}_{\text{cg}}(\Phi) = - \left(C_\Phi + \frac{6}{C_A} \right) \log \left(\frac{\sqrt{\Lambda} - \sqrt{\lambda}}{\sqrt{\Lambda} + \sqrt{\lambda}} \right).$$

Remark 10.20. Even if these numbers coincide exactly with those obtained in §8.3.9 for the Chebyshev method, one has to emphasise one important advantage of the CG method: The eigenvalue bounds γ and Γ may be unknown to the user. Vice versa, the efficacy of the Chebyshev method deteriorates if too pessimistic bounds γ , Γ are inserted.

10.2.7 Suitability for Secondary Iterations

Section 5.5 describes composed iterations arising from $x \mapsto x - B^{-1}(Ax - b)$ by replacing the exact solution of $B\delta = d$ with the approximation by a secondary iteration. Now we can start with $\delta^0 = 0$ and perform m steps of the CG algorithm. Positive and negative comments concerning this approach are given in the next lemma.

Lemma 10.21. *Let A and B be positive definite matrices.*

$$\Phi_A(x, b) = x - B^{-1}(Ax - b)$$

is the primary iteration. For solving $B\delta = d$, the CG method $\Upsilon_{\text{CG}}[\Phi_B]$ based on the iteration

$$\Phi_B(\delta, d) = \delta - C^{-1}(B\delta - d)$$

with a starting iterate $\delta^0 = 0$ is inserted as a secondary solver. The number k of CG steps is chosen such that $2c^k \leq \varepsilon$ holds with

$$c = (\sqrt{\Lambda} - \sqrt{\lambda})/(\sqrt{\Lambda} + \sqrt{\lambda}), \quad 0 < \delta C \leq B \leq \Delta C.$$

The composed iteration Φ_k is no longer linear, but it still can be written in the form

$$\Phi_k(x, b) = M_k(Ax - b)x + N_k(Ax - b)b \quad (10.26a)$$

with matrices $M_k(d)$, $N_k(d)$ depending on the defect $d = Ax - b$. They have the contraction number (10.26b) with respect to the energy norm:

$$\|M_k(Ax - b)\|_A \leq \|M_A\|_A + \varepsilon \|A^{1/2}B^{-1}A^{1/2}\|_2 \quad (M_A = I - B^{-1}A). \quad (10.26b)$$

Before proving the lemma, we comment on (10.26b). If, as in §5.5.1, B is a preconditioner with $\kappa(B^{-1}A) = \|A^{1/2}B^{-1}A^{1/2}\|_2 = \mathcal{O}(1)$, the right-hand side in (10.26b) is bounded by $\|M_A\|_A + C\varepsilon$. For example, one should choose ε such that

$$\|M_A\|_A + C\varepsilon \leq \frac{1}{2}(1 + \|M_A\|_A) < 1.$$

Proof of Lemma 10.21. The right-hand side d in $B\delta = d$ is the defect $d = Ax^m - b$ (cf. (5.18a)). Because of $\delta^0 = 0$, the error estimate (10.24b) yields the B -energy norm $\|\delta^k - \delta\|_B \leq \varepsilon \|\delta^0 - \delta\|_B = \varepsilon \|\delta\|_B$, $\delta := B^{-1}d$. From

$$\|\delta\|_B = \|B^{1/2}\delta\|_2 = \|B^{-1/2}A(x^m - x^*)\|_2 \leq \|B^{-1/2}AB^{-1/2}\|_2\|x^m - x^*\|_B,$$

we deduce

$$\begin{aligned} \|x^{m+1} - x^*\|_B &= \|x^m - \delta^k - x^*\|_B \leq \|x^m - \delta - x^*\|_B + \|\delta^k - \delta\|_B \\ &= \|\Phi_A(x^m, b) - x^*\|_B + \|\delta^k - \delta\|_B \\ &\leq \|M_A\|_B \|x^m - x^*\|_B + \varepsilon \|\delta\|_B \\ &\leq \left[\|M_A\|_B + \varepsilon \|B^{-1/2}AB^{-1/2}\|_2 \right] \|x^m - x^*\|_B. \end{aligned}$$

The identity $\|B^{-1/2}AB^{-1/2}\|_2 = \|A^{1/2}B^{-1/2}\|_2^2 = \|A^{1/2}B^{-1}A_2^{1/2}\|$ (cf. (B.21a)) proves the contraction number (10.26b). The definition of $M_k(Ax - b)$ and $N_k(Ax - b)$ in (10.26a) is obvious. Since the CG method is nonlinear (analogous to Remark 9.8a), Φ_k is also. \square

Remark 10.22. The composed iteration Φ_k defined in Lemma 10.21 is not well suited to be the basic iteration for the Chebyshev or CG method because the matrix $W_k(\delta) = A(I - M_k(\delta))$, $\delta = Ax - b$, of the third normal form of Φ_k depends on the value of the iterates x^m . Concerning this problem, see Golub–Overton [156] and Axelsson–Vassilevski [17].

10.2.8 Three-Term Recursion for p^m

Finally, we describe another formulation of the CG method. The three-term formulation is less important for the CG method itself, but is required as a stabilisation of, e.g., the CR algorithm in §10.3.3.

Inserting definition (10.23b,d): $r^{m+1} := r^m - \lambda Ap^m$ into (10.23f), one obtains $p^{m+1} := Nr^m - \lambda NAp^m + \text{const} \cdot p^m$. Since the scaling of the search direction is irrelevant, we may replace p^{m+1} by $-p^{m+1}/\lambda$. Because $Nr^m \in \mathcal{K}_{m+1}(NA, Nr^0)$ and $p^m \in \mathcal{K}_{m+1}(NA, Nr^0)$, the following ansatz is justified:

$$p^{m+1} := NA p^m - \sum_{\mu=0}^m \alpha_{\mu, m+1} p^{m-\mu}. \quad (10.27)$$

Condition (10.25) states that $\langle Ap^{m+1}, p^m \rangle = 0$ and determines the coefficients

$$\alpha_{0, m+1} = \langle ANAp^m, p^m \rangle / \langle Ap^m, p^m \rangle,$$

since $\langle Ap^{m-\mu}, p^m \rangle = 0$ for $\mu > 0$. Similarly we obtain

$$a_{1, m+1} = \langle ANAp^m, p^{m-1} \rangle / \langle Ap^{m-1}, p^{m-1} \rangle.$$

Lemma 10.23. Assume $(AN)^H = NA$. Then the coefficients in (10.27) satisfy $\alpha_{\mu, m+1} = 0$ for $\mu \geq 2$.

Proof. The condition $\langle Ap^{m+1}, p^{m-\mu} \rangle = 0$ yields the equation

$$\langle ANAp^m, p^{m-\mu} \rangle = \alpha_{\mu,m} \langle Ap^{m-\mu}, p^{m-\mu} \rangle.$$

The assertion follows from

$$\begin{aligned} \langle ANAp^m, p^{m-\mu} \rangle &= \langle Ap^m, NAp^{m-\mu} \rangle \\ &\stackrel{(10.27)}{=} \left\langle Ap^m, p^{m+1-\mu} + \sum_{\nu=0}^{m-\mu} \alpha_{\mu,m+1-\mu} p^{m-\mu-\nu} \right\rangle = 0. \quad \square \end{aligned}$$

Thanks to Lemma 10.23, p^{m+1} can be calculated from the three-term recursion

$$\begin{aligned} p^{m+1} &= NA p^m - \alpha_0 p^m - \alpha_1 p^{m-1} \\ \text{with } \alpha_0 &= \frac{\langle ANAp^m, p^m \rangle}{\langle Ap^m, p^m \rangle}, \quad \alpha_1 = \frac{\langle ANAp^m, p^{m-1} \rangle}{\langle Ap^{m-1}, p^{m-1} \rangle}, \end{aligned}$$

where the last term is absent for $m = 0$ (formally, we may set $\alpha_1 = 0, p^{-1} = 0$). The CG algorithm (10.23a–f) is equivalent to (10.28a–e):

$$\mathbf{start:} \quad x^0 \text{ arbitrary; } r^0 := b - Ax^0; p^{-1} := 0; p^0 := Nr^0; \quad (10.28a)$$

iteration: for $m = 0, 1, 2, \dots$ while $\langle Ap^m, p^m \rangle \neq 0$:

$$x^{m+1} := x^m + \lambda_{\text{opt}} p^m; \quad r^{m+1} := r^m - \lambda_{\text{opt}} Ap^m \quad \text{with} \quad (10.28b)$$

$$\lambda_{\text{opt}} := \langle r^m, p^m \rangle / \langle Ap^m, p^m \rangle; \quad (10.28c)$$

$$p^{m+1} := NA p^m - \alpha_0 p^m - \alpha_1 p^{m-1} \quad \text{with} \quad (10.28d)$$

$$\alpha_0 := \frac{\langle ANAp^m, p^m \rangle}{\langle Ap^m, p^m \rangle}; \quad \alpha_1 = \frac{\langle ANAp^m, p^{m-1} \rangle}{\langle Ap^{m-1}, p^{m-1} \rangle}; \quad (10.28e)$$

where again $\alpha_1 := 0$ is chosen for $m = 0$.

The next theorem is based on the very weak assumption $(AN)^H = NA$ which follows from $A > 0, N > 0$ or from $A = A^H, N = N^H$.

Theorem 10.24. Assume $(AN)^H = NA$. Let m_0 be the maximal index such that the directions generated in (10.28d) satisfy $\langle Ap^m, p^m \rangle \neq 0$ for all $0 \leq m \leq m_0$.

(a) The quantities x^m, r^m, p^m ($0 \leq m \leq m_0$) in (10.28a–e) satisfy $r^m = b - Ax^m$ and

$$\begin{aligned} \langle Ap^m, p^\ell \rangle &= \langle r^m, Nr^\ell \rangle = \langle r^m, p^\ell \rangle = 0 \quad \text{for } 0 \leq \ell < m, \\ \text{span}\{p^0, \dots, p^m\} &= \mathcal{K}_{m+1}(NA, Nr^0) \supset \text{span}\{Nr^0, \dots, Nr^m\} \end{aligned}$$

for $0 \leq m \leq m_0$. More precisely, we have

$$Nr^m \in \text{span}\{p^m, p^{m-1}\} \quad \text{for } 0 \leq m \leq m_0. \quad (10.29)$$

(b) As long as algorithm (10.22a–e) does not terminate, (10.22a–e) and (10.28a–e) produce the same iterates x^m , whereas the search directions p^m may differ by a nonvanishing factor.

(c) Assume, in addition, that $N + N^H > 0$. If the iteration (10.28a–e) terminates because of $p^m = 0$, the iterate x^m is already the exact solution.

Proof. The assertion is proved by induction. The start $m = 0$ is trivial. Let the statements hold for $0, 1, \dots, m - 1$. We abbreviate $\mathcal{K}_m(NA, Nr^0)$ by \mathcal{K}_m .

(i) For the proof of $\langle Ap^m, p^\ell \rangle = 0$, we use (10.28d):

$$Ap^m = ANAp^{m-1} - \alpha_0 Ap^{m-1} - \alpha_1 Ap^{m-2}.$$

For $\ell \in \{m-2, m-1\}$, the definitions of α_0 and α_1 prove $\langle Ap^m, p^\ell \rangle = 0$. Let $\ell \leq m-3$. The assumption $(AN)^H = NA$ yields $\langle ANAp^{m-1}, p^\ell \rangle = \langle Ap^{m-1}, NAp^\ell \rangle$. From $p^\ell \in \text{span}\{p^0, \dots, p^\ell\} = \mathcal{K}_{\ell+1}$, we conclude that

$$NAp^\ell \in \mathcal{K}_{\ell+2} \subset \mathcal{K}_{m-1} = \text{span}\{p^0, \dots, p^{m-2}\} \perp Ap^{m-1}.$$

Since Ap^{m-1} and Ap^{m-2} are also perpendicular to p^ℓ , $\langle Ap^m, p^\ell \rangle = 0$ follows.

(ii) By induction $\mathcal{K}_m = \text{span}\{p^0, \dots, p^{m-1}\}$ holds. We use again (10.28d): $p^m = NAp^{m-1} - \alpha_0 p^{m-1} - \alpha_1 p^{m-2} \in N\mathcal{K}_m + \text{span}\{p^0, \dots, p^{m-1}\} \subset \mathcal{K}_{m+1}$. This proves $\text{span}\{p^0, \dots, p^m\} \subset \mathcal{K}_{m+1}$. On the other hand, we have

$$\mathcal{K}_{m+1} \subset \mathcal{K}_m + N\mathcal{K}_m = \text{span}\{p^0, \dots, p^{m-1}\} + NA\text{span}\{p^0, \dots, p^{m-1}\} \ni p^m$$

because of (10.28d). This proves the reverse inclusion $\mathcal{K}_{m+1} \subset \text{span}\{p^0, \dots, p^m\}$.

(iii) $0 = \langle r^m, p^\ell \rangle = \langle r^{m-1}, p^\ell \rangle - \lambda_{\text{opt}} \langle Ap^m, p^\ell \rangle = 0$ holds for $\ell < m-1$ by induction and follows for $\ell = m-1$ by definition of λ_{opt} . This proves $r^m \perp \mathcal{K}_m$.

(iv) Now we prove (10.29). The definition of r^m in (10.28b) shows that $Nr^m = Nr^{m-1} - \lambda NAp^{m-1}$. By induction $Nr^{m-1} \in \text{span}\{p^{m-2}, p^{m-1}\}$ holds, while (10.28d) yields $NAp^{m-1} = p^m + \alpha_0 p^{m-1} + \alpha_1 p^{m-2} \in \text{span}\{p^{m-2}, p^{m-1}, p^m\}$. Hence ANr^m has the representation

$$ANr^m = b_0 Ap^m + b_1 Ap^{m-1} + b_2 Ap^{m-2}.$$

The scalar product with p^{m-2} yields the value $b_2 = \frac{\langle ANr^m, p^{m-2} \rangle}{\langle Ap^{m-2}, p^{m-2} \rangle}$. By assumption $(AN)^H = NA$, $\langle ANr^m, p^{m-2} \rangle = \langle r^m, NAp^{m-2} \rangle$ holds. Since $NAp^{m-2} \in N\mathcal{K}_{m-1} \subset \mathcal{K}_m$, part (iii) proves $b_2 = 0$ and $Nr^m \in \text{span}\{p^{m-1}, p^m\}$ follows.

(v) $\langle r^m, Nr^\ell \rangle = 0$ for $\ell < m$ is a consequence of (10.29) and $r^m \perp \mathcal{K}_m$.

(vi) Part (b) holds, since another scaling of p^m does not change x^m .

(vii) If $p^m = 0$, (10.29) implies $Nr^m \in \text{span}\{p^0, \dots, p^{m-1}\}$. Since $\langle r^m, p^\ell \rangle = 0$ for $\ell < m$, we conclude that $\langle r^m, Nr^m \rangle = 0$ and the assumption $N + N^H > 0$ implies that $r^m = 0$. \square

10.3 Method of Conjugate Residuals (CR)

10.3.1 Algorithm

In the case of the gradient method, a residual oriented transformation is discussed in §9.2.4.2. Under the assumption $A > 0$, the iteration $\Phi \in \mathcal{L}_{\text{pos}}$ with $N > 0$ is transformed to $\bar{x}^{m+1} := \bar{x}^m - (\bar{A}\bar{x}^m - \bar{b})$ with $\bar{A} := A^{1/2}NA^{1/2} > 0$, $\bar{b} := A^{1/2}Nb$, $\bar{x}^m := A^{1/2}x^m$, $\bar{p}^m = A^{1/2}p^m$, $\bar{r}^m = A^{1/2}Nr^m$ (cf. (9.19)). As in (10.21a–e), we can formulate the CG algorithm (10.21a–e) with A, x, b, p, r replaced by $\bar{A}, \bar{x}, \bar{b}, \bar{p}, \bar{r}$. Then we substitute these quantities by the original ones and obtain the following algorithm $\Upsilon_{\text{CR}}[\Phi]$:

$$\begin{aligned} \text{start: } & x^0 \text{ arbitrary; } r^0 := b - Ax^0; \quad p^0 := Nr^0; \quad m = 0; \\ \text{iteration: } & x^{m+1} := x^m + \lambda_{\text{opt}} p^m \text{ with} \\ & \lambda_{\text{opt}} := \lambda_{\text{opt}}(r^m, NAp^m, N) = \frac{\langle Nr^m, Ap^m \rangle}{\langle NAp^m, Ap^m \rangle}; \\ & r^{m+1} := r^m - \lambda_{\text{opt}} A p^m; \\ & p^{m+1} := Nr^{m+1} - \frac{\langle ANr^{m+1}, NAp^m \rangle}{\langle NAp^m, Ap^m \rangle} p^m; \end{aligned} \quad (10.30)$$

For $N = I$, this method is equivalent to the *method of the conjugate residuals* (CR) of Stiefel [354].

The following statements follow from the properties of the CG method applied to $\bar{A}, \bar{x}, \bar{b}, \bar{p}, \bar{r}$ after a reformulation by A, x, b, p, r .

Proposition 10.25. (a) The number $m_0 = \deg_{\bar{A}}(\bar{e}^0) = \deg_{NA}(e^0) = \deg_{AN}(r^0)$ is the same as in Lemma 10.18a.

(b) The directions p^m are ANA-orthogonal.

(c) The statements in (10.11a,b) become

$$ANr^m \perp \mathcal{K}_m(NA, Nr^0) = \text{span}\{p^0, \dots, p^{m-1}\} = \text{span}\{Nr^0, \dots, Nr^{m-1}\}.$$

(d) The convergence rate c is the same c as in Theorem 10.17. Note that the involved norms are different. Here the residuals are the minimisers of

$$\min \left\{ \|N^{1/2}A(x - x^*)\|_2 : x = x^0 + \mathcal{K}_m(NA, Nr^0) \right\}$$

and are bounded by

$$\|N^{1/2}r^m\|_2 \leq \frac{2c^m}{1 + c^{2m}} \|N^{1/2}r^0\|_2.$$

In the case of $N = I$, the CR method corresponds to the formulation in §10.1.5.2 with the Krylov spaces $\mathcal{U}_m = \mathcal{K}_m(A, r^0)$ and $\mathcal{V}_m = A\mathcal{K}_m(A, r^0)$ (note that $A = A^H$).

10.3.2 Application to Hermitian Matrices

In the following we assume

$$A = A^H \text{ regular} \quad \text{and} \quad N > 0.$$

Since $A^H NA > 0$, the denominator $\langle NAp^m, Ap^m \rangle$ in (10.30) vanishes if and only if $p^m = 0$. Hence, the algorithm (10.30) is applicable as long as $p^m \neq 0$. In the indefinite case, however, there is a severe difference to the conjugate gradient method. The CG method for $A > 0$ terminates with $r^m = 0$, i.e., $x^m = A^{-1}b$ ('lucky breakdown'), whereas for an indefinite matrix A an unlucky breakdown may occur.

Remark 10.26. Assume that $A = A^H$ has positive and negative eigenvalues. Then there are initial values $x^0 \neq A^{-1}b$ so that $\lambda_{\text{opt}}(r^0, NAp^0, N) = 0$. Then $p^1 = 0$ leads to a breakdown, while $x^1 = x^0$ is still different from the true solution.

Proof. $\lambda_{\text{opt}}(r^0, NAp^0, N) = 0$ follows from $\langle Ap^0, p^0 \rangle = \langle NANp^0, p^0 \rangle = 0$ which holds for certain $p^0 \neq 0$. Since $p^1 \perp_{NAN} p^0$ and $p^1 \in \text{span}\{p^0\}$ because of $\lambda_{\text{opt}} = 0$, $p^1 = 0$ follows. \square

Lemma 10.27. Let $A = A^H$ and $N > 0$. Assume that the algorithm (10.30) for a fixed x^0 is applicable for all $0 \leq m \leq m_0$. Then, as in Proposition 10.25b–c, the search directions p^m are ANA-orthogonal and

$$ANr^m \perp \mathcal{K}_m(NA, Nr^0) = \text{span}\{p^0, \dots, p^{m-1}\} = \text{span}\{Nr^0, \dots, Nr^{m-1}\}$$

holds. The iterate x^m in (10.30) minimises the norm

$$\|N^{1/2}r^m\|_2 = \min \left\{ \|N^{1/2}A(x - x^*)\|_2 : x \in x^0 + \mathcal{K}_m(NA, Nr^0) \right\}. \quad (10.31)$$

Proof. (i) Concerning the first two statements, the previous proof by induction can be repeated without change.

(ii) Note that $x^m - x^0 \in \mathcal{K}_m := \text{span}\{p^0, \dots, p^{m-1}\} = \mathcal{K}_m(NA, Nr^0)$. Because of $ANA > 0$, $\{\langle A(x - x^*), NA(x - x^*) \rangle : x - x^0 \in \mathcal{K}_m\}$ attains its minimum at $x = x^m$ if and only if the gradient $ANA(x^m - x^*) = -ANr^m$ is orthogonal to \mathcal{K}_m . This, however, is the second statement of the lemma. \square

The reason for the breakdown mentioned in Remark 10.26 is that the spaces $\text{span}\{Nr^0, Nr^1\} = \text{span}\{Nr^0\}$ and $\text{span}\{Nr^0, NANr^0\}$ differ. This fact suggests that the subspace $\mathcal{K}_m(NA, Nr^0) = \text{span}\{Nr^0, \dots, (NA)^{m-1}Nr^0\}$ is better suited than $\text{span}\{Nr^0, \dots, Nr^{m-1}\}$.

Even if $\langle ANr^m, Nr^m \rangle = 0$ does not occur during the calculations, it may happen that Nr^m is 'almost' contained in $\mathcal{K}_m(NA, Nr^0)$, leading to a numerical instability of the algorithm. One remedy is constructing the search directions p^m by the three-term recursion explained in §10.2.8.

10.3.3 Stabilised Method of Conjugate Residuals

Using the three-term recursion in algorithm (10.30), we obtain the following algorithm $\mathcal{T}_{\text{CR}}^{\text{stab}}[\Phi]$:

$\mathcal{T}_{\text{CR}}^{\text{stab}}[\Phi]$	stabilised method of the conjugate residuals	(10.32)
start:	x^0 arbitrary; $r^0 := b - Ax^0$; $p^{-1} := 0$; $p^0 := Nr^0$;	(10.32a)
iteration:	for $m = 0, 1, 2, \dots$ while $\langle Ap^m, NAp^m \rangle \neq 0$:	
	$x^{m+1} := x^m + \lambda p^m$; $r^{m+1} := r^m - \lambda Ap^m$ with	(10.32b)
	$\lambda := \langle r^m, NAp^m \rangle / \langle Ap^m, NAp^m \rangle$;	(10.32c)
	$p^{m+1} := NAp^m - \alpha_0 p^m - \alpha_1 p^{m-1}$ with	(10.32d)
	$\alpha_0 := \frac{\langle ANAp^m, NAp^m \rangle}{\langle Ap^m, NAp^m \rangle}$; $\alpha_1 = \frac{\langle ANAp^m, NAp^{m-1} \rangle}{\langle Ap^{m-1}, NAp^{m-1} \rangle}$;	(10.32e)

Exercise 10.28. By $ANAp^m$ appearing in (10.32e), algorithm (10.32a–e) seems to cost two multiplications by the matrix A per iteration step. Rewrite algorithm (10.32a–e) with an additional recursion for $a^m := Ap^m$ so that only one multiplication by A is needed.

Theorem 10.29. Assume $(AN)^H = NA$. Let m_0 be the maximal index such that the directions generated in (10.28d) satisfy $\langle Ap^m, NAp^m \rangle \neq 0$ for all $0 \leq m \leq m_0$.
(a) The quantities x^m , r^m , p^m ($0 \leq m \leq m_0$) in (10.28a–e) satisfy $r^m = b - Ax^m$ and

$$\begin{aligned} \langle Ap^m, NAp^\ell \rangle &= \langle r^m, NANr^\ell \rangle = \langle r^m, NAp^\ell \rangle = 0 \quad \text{for } 0 \leq \ell < m, \\ \text{span}\{p^0, \dots, p^m\} &= \mathcal{K}_{m+1}(NA, Nr^0) \supset \text{span}\{Nr^0, \dots, Nr^m\} \end{aligned}$$

for $0 \leq m \leq m_0$. More precisely, we have

$$Nr^m \in \text{span}\{p^m, p^{m-1}\} \quad \text{for } 0 \leq m \leq m_0. \quad (10.33)$$

(b) As long as algorithm (10.30) does not terminate, (10.30) and (10.32a–e) produce the same iterates x^m , whereas the search directions may differ by a nonvanishing factor.

(c) Assume in addition that $N + N^H > 0$. If the iteration (10.32a–e) terminates because of $p^m = 0$, the iterate x^m is already the exact solution.

Proof. The assertion is proved by induction. The start $m = 0$ is trivial. Let the statements hold for $0, 1, \dots, m - 1$. We abbreviate $\mathcal{K}_m(NA, Nr^0)$ by \mathcal{K}_m .

(i) For the proof of $\langle Ap^m, NAp^\ell \rangle = 0$, we use (10.32d):

$$Ap^m = ANAp^{m-1} - \alpha_0 Ap^{m-1} - \alpha_1 Ap^{m-2}.$$

For $\ell \in \{m-2, m-1\}$, the definitions of α_0 and α_1 prove $\langle Ap^m, NAp^\ell \rangle = 0$. Let $\ell \leq m-3$. The assumption $(AN)^H = NA$ yields $\langle ANAp^{m-1}, NAp^\ell \rangle = \langle Ap^{m-1}, (NA)^2 p^\ell \rangle$. From $p^\ell \in \text{span}\{p^0, \dots, p^\ell\} = \mathcal{K}_{\ell+1}$, we conclude that $NAp^\ell \in \mathcal{K}_{\ell+2} \subset \mathcal{K}_{m-1} = \text{span}\{p^0, \dots, p^{m-2}\}$ and $NA\mathcal{K}_{m-1} \perp Ap^{m-1}$. Since Ap^{m-1} and Ap^{m-2} are also perpendicular to NAp^ℓ , $\langle Ap^m, NAp^\ell \rangle = 0$ follows.

(ii) By induction, $\mathcal{K}_m = \text{span}\{p^0, \dots, p^{m-1}\}$ holds. We again use (10.32d): $p^m = NAp^{m-1} - \alpha_0 p^{m-1} - \alpha_1 p^{m-2} \in NAK_m + \text{span}\{p^{m-2}, p^{m-1}\} \subset \mathcal{K}_{m+1}$. This proves $\text{span}\{p^0, \dots, p^m\} \subset \mathcal{K}_{m+1}$. On the other hand, the inclusion

$$\mathcal{K}_{m+1} \subset \mathcal{K}_m + NAK_m = \text{span}\{p^0, \dots, p^{m-1}\} + NA \text{span}\{p^0, \dots, p^{m-1}\} \ni p^m$$

follows from (10.32d) proving the reverse inclusion $\mathcal{K}_{m+1} \subset \text{span}\{p^0, \dots, p^m\}$.

(iii) $0 = \langle r^m, NAp^\ell \rangle = \langle r^{m-1}, NAp^\ell \rangle - \lambda_{\text{opt}} \langle Ap^m, NAp^\ell \rangle = 0$ holds for $\ell < m-1$ by induction and follows for $\ell = m-1$ by definition of λ_{opt} . This proves $r^m \perp NAK_m$.

(iv) For the proof of (10.33), use the definition of r^m in (10.32b): $Nr^m = Nr^{m-1} - \lambda NAp^{m-1}$. By induction $Nr^{m-1} \in \text{span}\{p^{m-2}, p^{m-1}\}$ holds, while (10.32d) yields $NAp^{m-1} = p^m + \alpha_0 p^{m-1} + \alpha_1 p^{m-2} \in \text{span}\{p^{m-2}, p^{m-1}, p^m\}$. Hence Nr^m has the representation

$$Nr^m = b_0 p^m + b_1 p^{m-1} + b_2 p^{m-2}.$$

Using part (i), we obtain $b_2 = \frac{\langle ANr^m, NAp^{m-2} \rangle}{\langle Ap^{m-2}, NAp^{m-2} \rangle}$ by taking the scalar product of ANr^m with NAp^{m-2} . $(AN)^H = NA$ implies that $\langle ANr^m, NAp^{m-2} \rangle = \langle r^m, (NA)^2 p^{m-2} \rangle$ holds. Since $NAp^{m-2} \in NAK_{m-1} \subset \mathcal{K}_m$, part (iii) proves $b_2 = 0$, and $Nr^m \in \text{span}\{p^{m-1}, p^m\}$ follows.

(v) $\langle r^m, NANr^\ell \rangle = 0$ for $\ell < m$ is a consequence of (10.29) and $r^m \perp NAK_m$.

(vi) Statement (b) follows as in Theorem 10.24. For Part (c), use that $p^m = 0$ implies $Nr^m = cp^{m-1}$ for some $c \in \mathbb{K}$. Obviously, $c = 0$ and $r^m = 0$ follow from $0 = \langle ANr^m, NAp^{m-1} \rangle = \langle r^m, (NA)^2 p^{m-1} \rangle$. This equation holds since $p^m = 0$ implies $\mathcal{K}_m = \mathcal{K}_{m+1}$ and therefore $\langle r^m, (NA)^2 p^{m-1} \rangle = 0$ because of $r^m \perp NAK_m = NAK_{m+1} = (NA)^2 \mathcal{K}_m$. \square

10.3.4 Convergence Results for Indefinite Matrices

Lemma 10.27 carries over to algorithm (10.32) since it produces the same iterates x^m . The error estimate in Proposition 10.25d cannot be transferred directly to indefinite matrices because the spectrum of NA no longer lies in the positive part. In the general case, the resulting convergence speed is definitely slower than in the positive definite case. Note that the quantity c below is defined in terms of κ , whereas c in Proposition 10.25d is derived from $\sqrt{\kappa}$. Hence, in general, the typical acceleration by the conjugate gradient technique does not take place, but notice Theorem 10.31.

Theorem 10.30. Assume $N > 0$, $A = A^H$ regular, and $\kappa = \kappa(NA)$. Then the iterates x^m of algorithm (10.32) satisfy the error estimate

$$\|N^{1/2}A(x^m - x^*)\|_2 \leq \frac{2c^\mu}{1 + c^{2\mu}} \|N^{1/2}A(x^0 - x^*)\|_2 \quad (10.34)$$

with $c := (\kappa - 1)/(\kappa + 1)$ and $\frac{m}{2} - 1 < \mu \leq \frac{m}{2}$, where $\mu \in \mathbb{N}_0$. Hence, the asymptotic convergence rate amounts to $\sqrt{c} = 1 - 1/\kappa + \mathcal{O}(\kappa^{-2})$.

Proof. For odd m , we exploit the monotone convergence $\|N^{1/2}Ae^{m+1}\|_2 \leq \|N^{1/2}Ae^m\|_2$ following from (10.31). Therefore, consider an even $m = 2\mu$. Analogously to Remark 10.13,

$$\|N^{1/2}Ae^m\|_2 \leq \left(\max_{\lambda \in \sigma(NA)} |P_m(1 - \lambda)| \right) \|N^{1/2}Ae^0\|_2 \quad (10.35)$$

holds for any polynomial $P_m \in \mathcal{P}_m$ with $P_m(1) = 1$. Let p_μ be a polynomial of degree $\leq \mu = \frac{m}{2}$ with $p_\mu(1) = 1$. $P_m(\xi) := p_\mu(\xi(2 - \xi))$ is of degree $\leq m$ and satisfies $P_m(1) = 1$. Evidently, $P_m(1 - \lambda) = p_\mu(1 - \lambda^2)$ holds, from which

$$\|N^{1/2}Ae^m\|_2 \leq \max \{ |p_\mu(1 - \lambda^2)| : \lambda \in \sigma(NA) \} \|N^{1/2}Ae^0\|_2.$$

follows. If $\lambda \in \sigma(NA)$, we have $|\lambda| \in [\gamma, \Gamma]$ and $\lambda^2 \in [\gamma^2, \Gamma^2]$, where

$$\gamma := 1/\rho(A^{-1}N^{-1}) = \min\{|\lambda| : \lambda \in \sigma(NA)\}, \quad \Gamma := \rho(NA).$$

Since $[\gamma^2, \Gamma^2]$ lies in the positive half-axis, the Chebyshev polynomial (8.27a) yields the following estimate with $c = (\Gamma - \gamma)/(\Gamma + \gamma) = (\kappa - 1)/(\kappa + 1)$:

$$\max \{ |p_\mu(1 - \lambda^2)| : \lambda \in \sigma(NA) \} \leq \max_{\gamma^2 \leq \xi \leq \Gamma^2} |p_\mu(1 - \xi)| \leq \frac{2c^\mu}{1 + c^{2\mu}}. \quad \square$$

Estimate (10.34) may be too pessimistic. Often a milder form of indefiniteness occurs. If, for instance, the Helmholtz equation $-\Delta u - cu = f$ with $c > 0$ is discretised, A has eigenvalues λ_μ^h ($1 \leq \mu \leq n = n_h$), where

$$\lambda_\mu^h = \lambda_{\mu,0}^h - c, \quad 0 < \lambda_{\mu,0}^h : \text{eigenvalues of the Poisson model case (3.1a).}$$

For $h \rightarrow 0$, the discrete eigenvalues λ_μ^h tend to the Laplace eigenvalues λ_μ which cannot accumulate (cf. [193, §11]). Therefore the following properties are satisfied:

$$\text{The number } k \text{ of negative eigenvalues is bounded for } h \rightarrow 0. \quad (10.36a)$$

$$\begin{aligned} \text{For all } h > 0, \text{ the nonpositive eigenvalues} \\ \text{belong to } [-c_1, -c_0] \text{ with } 0 < c_0 \leq c_1. \end{aligned} \quad (10.36b)$$

$$\text{The positive eigenvalues are in } [\gamma, \Gamma] \text{ with } 0 < \gamma \leq \Gamma. \quad (10.36c)$$

Let $k = k_h$ be the number of negative eigenvalues λ_μ^h , $1 \leq \mu \leq k$. Define

$$\pi_h(1 - \xi) = \prod_{\mu=1}^k (1 - \xi/\lambda_\mu^h).$$

Let p_μ be the Chebyshev polynomial (8.27a) of degree $\mu := m - k$ for $a = 1 - \Gamma$ and $b = 1 - \gamma$. The product $P_m(\xi) := \pi_h(\xi)p_\mu(\xi)$ is of degree m with $P_m(1) = 1$. Since $P_m(1 - \lambda) = 0$ holds for the negative eigenvalues $\lambda \in \sigma(NA)$, the factor on the right-hand side in (10.35) reduces to

$$\max \{ |P_m(1 - \lambda)| : \lambda \in [\gamma, \Gamma] \} \leq \max \{ |\pi_h(1 - \lambda)| : \lambda \in [\gamma, \Gamma] \} \frac{2c^\mu}{1 + c^{2\mu}}$$

with $c := \frac{\sqrt{\Gamma} - \sqrt{\gamma}}{\sqrt{\Gamma} + \sqrt{\gamma}}$ (cf. (10.36c)). $|\pi_h(1 - \lambda)|$ can be estimated by $(1 + \Gamma/c_0)^k$ (cf. (10.36b)). The m -th root of the bound $(1 + \Gamma/c_0)^k \frac{2c^\mu}{1 + c^{2\mu}}$ tends to c . Hence, the asymptotic convergence rate is not influenced by the negative eigenvalues. This proves the next theorem.

Theorem 10.31. Assume $A = A^H$, $N > 0$, and let the eigenvalues of NA satisfy (10.36a–c). Replace the spectral condition number $\kappa(NA)$ by the possibly smaller number $\kappa := \Gamma/\gamma$ (γ, Γ in (10.36c)). Then the error estimate for the algorithm (10.32) of the conjugate residuals reads

$$\|N^{1/2}A(x^m - x^*)\|_2 \leq 2 \left(\frac{1 + \Gamma/c_0}{c} \right)^k \|N^{1/2}A(x^0 - x^*)\|_2$$

with the asymptotic convergence rate $c := \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} = \frac{\sqrt{\Gamma}-\sqrt{\gamma}}{\sqrt{\Gamma}+\sqrt{\gamma}}$ and c_0 in (10.36a).

An alternative to the method (10.32) of conjugate residuals is the application of the standard CG method to the Kaczmarz iteration (cf. §5.6.3). Then the convergence speed is as slow as in Theorem 10.30. In the situation of (10.36a–c), the convergence rate would not improve.

m	value in the middle	$\frac{\ e^m\ _A}{\ e^{m-1}\ _A}$
1	0.2222124445	0.157356
2	0.4269164370	0.537790
3	0.4510237348	0.439627
4	0.4759275765	0.438732
10	0.4998558015	0.384330
20	0.5000000047	0.338399
21	0.5000000029	0.389211
22	0.5000000012	0.407384
23	0.5000000003	0.317164
24	0.5000000002	0.442606
25	0.5000000000	0.691876
26	0.5000000000	0.768926
27	0.5000000000	0.955932
28	0.5000000000	1.02596
29	0.5000000000	1.03161
30	0.5000000000	1.03076

10.3.5 Numerical Examples

For reasons of comparison, we first test the positive definite Poisson model problem with $h = \frac{1}{32}$. We apply the CR method to the ILU iteration (five-point pattern) with the same parameters as in Table 10.2. The results given in Table 10.3 are similar to those of the standard CG method in Table 10.2.

Next, we choose the discrete Helmholtz equation $-\Delta u - 50u = f$ as an indefinite example. Here the matrix A is the Poisson model matrix minus $50I$. It has three negative eigenvalues $\lambda_1 = -30.277$,

Table 10.3 $\Upsilon_{\text{CR}}^{\text{stab}}[\Phi_{\text{ILU}}]$: CR method for the Poisson model problem applied to the 5-point-ILU iteration ($\omega = -1$, $h = 1/32$).

$\lambda_2 = \lambda_3 = -0.7866$, while $\lambda_4 = 28.7$ is the smallest positive eigenvalue. For the modified ILU decomposition, the diagonal must be enlarged by 55 (cf. Remark 7.44). The results of Table 10.4 show that the reduction factor moves toward the asymptotic convergence rate and is of a size similar to the positive definite case of Table 10.3. The stagnation for $m \geq 33$ is due to rounding. In both examples the algorithm behaves stable.

m	value in the middle	$\frac{\ e^m\ _A}{\ e^{m-1}\ _A}$	m	value in the middle	$\frac{\ e^m\ _A}{\ e^{m-1}\ _A}$
1	-1.129805206	1.36998	15	0.5017304154	0.97466
2	0.5616735534	0.41788	16	0.5019212154	0.86920
3	0.9170148791	0.77945	17	0.5018558645	0.56086
4	0.7375934000	0.78685	18	0.5017568935	0.32511
5	0.6675855715	0.88467	19	0.5008067252	0.27287
6	0.5834957931	0.95835	20	0.5003741869	0.19130
7	0.5440078825	0.99228	21	0.5003841894	0.35875
8	0.5222771713	1.00338	30	0.4999998664	0.30740
9	0.5099064832	1.00768	31	0.4999999994	0.40910
10	0.5053055088	1.00956	32	0.4999999838	0.69469
11	0.5029466483	1.01213	33	0.4999999962	0.90769
12	0.5020970259	1.01621	34	0.4999999984	0.97862
13	0.5015223028	1.01159	35	0.4999999986	0.98121
14	0.5015388760	1.00335	36	0.4999999994	0.99385

Table 10.4 CR method $\gamma_{\text{CR}}^{\text{stab}}[\Phi_{\text{ILU}}]$ for an indefinite problem based on the 5-point-ILU iteration.

10.4 Method of Orthogonal Directions

The CG method (10.15a–e) minimises the error $\|e^m\|_A = \|A^{1/2}e^m\|_2$ with respect to the energy norm over the Krylov space $\mathcal{K}_m(A, r^0)$. The method of conjugate residuals (with $N = I$) minimises the residual $\|r^m\|_2 = \|Ae^m\|_2$ over the same space. A more natural norm would be $\|e^m\|_2$. Then the search directions p^m should be orthogonal in the usual sense. This can be achieved by replacing the Krylov space $\mathcal{K}_m(A, r^0)$ by $A\mathcal{K}_m(A, r^0) = \mathcal{K}_m(A, Ar^0)$. The corresponding algorithm (10.37) is described by Fridman [140] (1963) and called the *method of orthogonal directions* (OD) since the search directions form an orthogonal system if $N=I$. The application of OD to an iteration Φ with the matrix $N[\Phi] > 0$ takes the form

$\gamma_{\text{OD}}[\Phi]$	method of orthogonal directions	(10.37)
start:	x^0 arbitrary; $r^0 := b - Ax^0$; $q^{-1} := r^0$; $q^0 := ANq^{-1}$;	(10.37a)
iteration:	for $m = 0, 1, 2, \dots$ while $q^m \neq 0$: $x^{m+1} := x^m + \lambda p^m$; $r^{m+1} := r^m - \lambda Ap^m$ with $p^m := Nq^m$; $\rho_m := \langle q^m, p^m \rangle$; $\lambda := \frac{\langle r^m, p^{m-1} \rangle}{\rho_m}$;	(10.37b)
	$q^{m+1} := Ap^m - \alpha_0 q^m - \alpha_1 q^{m-1}$ with	(10.37c)
	$\alpha_0 := \langle Ap^m, p^m \rangle / \rho_m$; $\alpha_1 := \langle Ap^m, p^{m-1} \rangle / \rho_m$;	(10.37d)

where $\alpha_1 := 0$ is set for $m = 0$. The method (10.37) is unstable as we observe from the results in Tables 10.5 and 10.6. A stabilisation is given by Stoer [355, (3.16)]. On the other hand, we can do without it if only a few iteration steps are required.

m	value in the middle	$\ e^m\ _2$	$\frac{\ e^m\ _2}{\ e^{m-1}\ _2}$	$\sqrt[m]{\frac{\ e^m\ _2}{\ e^0\ _2}}$
1	-4.57492859 ₁₀ -2	2.93956 ₁₀ -1	3.92836 ₁₀ -1	3.92836 ₁₀ -1
10	4.989708480 ₁₀ -1	5.16807 ₁₀ -4	3.74216 ₁₀ -1	4.82976 ₁₀ -1
11	5.002475524 ₁₀ -1	1.91138 ₁₀ -4	3.69844 ₁₀ -1	4.71399 ₁₀ -1
15	5.000015863 ₁₀ -1	5.80274 ₁₀ -6	4.67856 ₁₀ -1	4.56356 ₁₀ -1
16	5.000085958 ₁₀ -1	2.76800 ₁₀ -6	4.77016 ₁₀ -1	4.57621 ₁₀ -1
17	4.999867395 ₁₀ -1	4.98160 ₁₀ -6	1.79971 ₁₀ +0	4.96007 ₁₀ -1
18	4.999923552 ₁₀ -1	1.35781 ₁₀ -5	2.72567 ₁₀ +0	5.45253 ₁₀ -1
19	4.999645661 ₁₀ -1	3.77863 ₁₀ -5	2.78287 ₁₀ +0	5.94095 ₁₀ -1
20	4.998667835 ₁₀ -1	1.05920 ₁₀ -4	2.80315 ₁₀ +0	6.42016 ₁₀ -1
27	5.290373141 ₁₀ -1	7.10159 ₁₀ -2	3.20465 ₁₀ +0	9.16477 ₁₀ -1
30	2.379020942 ₁₀ +0	1.33836 ₁₀ +0	1.77625 ₁₀ +0	1.01957 ₁₀ +0

Table 10.5 OD method $\Upsilon_{OD}[\Phi_{ILU}]$ applied to the same problem as in Table 10.3.

m	value in the middle	$\ e^m\ _2$	$\frac{\ e^m\ _2}{\ e^{m-1}\ _2}$	$\sqrt[m]{\frac{\ e^m\ _2}{\ e^0\ _2}}$
1	1.288025563 ₁₀ +0	4.58268 ₁₀ -1	6.12419 ₁₀ -1	6.12419 ₁₀ -1
10	5.107964511 ₁₀ -1	2.08681 ₁₀ -1	9.93821 ₁₀ -1	8.80119 ₁₀ -1
20	5.084149051 ₁₀ -1	1.00523 ₁₀ -2	4.81485 ₁₀ -1	8.06139 ₁₀ -1
30	5.000072697 ₁₀ -1	5.10416 ₁₀ -6	4.28537 ₁₀ -1	6.72659 ₁₀ -1
35	4.999124543 ₁₀ -1	2.09700 ₁₀ -4	2.28682 ₁₀ +0	7.91591 ₁₀ -1
40	4.178915511 ₁₀ -1	5.64009 ₁₀ -2	6.53540 ₁₀ +0	9.37412 ₁₀ -1

Table 10.6 OD method $\Upsilon_{OD}[\Phi_{ILU}]$ for the indefinite problem in Table 10.4.

The proof of the following theorem is left to the reader.

Theorem 10.32. Assume that $N > 0$ and $A = A^H$. Let m_0 be the largest index with $q^m \neq 0$. $q^{m_0+1} = 0$ implies $x^{m_0+1} = x^*$. For all $0 \leq m \leq m_0$, (10.38a–c) hold:

$$\langle Nq^k, q^\ell \rangle = 0 \quad \text{for } 0 \leq k \neq \ell \leq m_0, \tag{10.38a}$$

$$\langle Nq^k, q^k \rangle \neq 0 \quad \text{for } 0 \leq k \leq m_0,$$

$$r^m \perp N\mathcal{K}_m(AN, r^0), \tag{10.38b}$$

$$\text{span}\{q^0, \dots, q^{m-1}\} = AN\mathcal{K}_m(AN, r^0) = \mathcal{K}_m(AN, ANr^0). \tag{10.38c}$$

x^m is the minimiser $\min\{\|N^{-1/2}(x - x^*)\| : x \in x^0 + N\mathcal{K}_m(AN, r^0)\}$.

This case corresponds to the choice of the spaces in §10.1.5.3. The connection with the Lanczos method is described by Paige–Saunders [307]. The method SYMMLQ defined there is a further stabilisation of the method (10.37).

A review of the algorithms discussed above and of additional variants is given by Stoer [355].

10.5 Solution of Nonsymmetric Systems

Some of the methods described above do not require the assumption (9.1) of positive definiteness of A and are also applicable to indefinite but still symmetric matrices. The nonsymmetric situation is more difficult.

10.5.1 Generalised Minimal Residual Method (GMRES)

The following method generalises the minimal residual iteration described in §9.4 and corresponds to the approach in §10.1.5.2.

10.5.1.1 General Setting and Convergence

The ‘generalised minimal residual method’ described by Saad–Schultz [329] (see also Walker [388]) determines the vector in the affine space $x^0 + \mathcal{K}_m(A, r^0)$ minimising the residual:

$$x^m = \operatorname{argmin} \left\{ \|b - Ax\|_2 : x \in x^0 + \mathcal{K}_m(A, r^0) \right\}. \quad (10.39)$$

We recall that the control of the residual might be questionable (cf. Remark 2.35).

As the CG method, GMRES (with exact arithmetic) yields the true solution after at least $\#I$ steps.

Proposition 10.33. *For regular A and $m_0 := \deg_A(e^0) = \deg_A(r^0) \leq \#I$, the iterate x^{m_0} is the exact solution x^* .*

Proof. For regular A , the statements $p_{m_0}(A)e^0 = 0$ and $p_{m_0}(A)r^0 = -Ap_{m_0}(A)e^0 = 0$ are equivalent. Let $p_{m_0} \in \mathcal{P}_{m_0}$ be the polynomial with $p_{m_0}(A)e^0 = 0$. Note that $p_{m_0}(0) \neq 0$ by Lemma 8.12. After a suitable scaling, $p_{m_0}(0) = 1$ holds so that $p_{m_0}(\xi) = 1 - \xi q_{m_0-1}(\xi)$. The correction $q_{m_0-1}(A)r^0 \in \mathcal{K}_{m_0}(A, r^0)$ yields

$$x^{m_0} - A^{-1}b = e^0 + q_{m_0-1}(A)r^0 = (I - Aq_{m_0-1}(A))e^0 = p_{m_0}(A)e^0 = 0,$$

i.e., x^{m_0} is the exact solution. □

In the case of a general matrix A , one cannot expect other convergence statements than $x^{m_0} = A^{-1}b$, as the following example shows.

Example 10.34. Define $A \in \mathbb{R}^{n \times n}$ by the entries $A_{ij} = \begin{cases} 1 & j - i = 1 \bmod n \\ 0 & \text{otherwise} \end{cases}$ (e.g., $A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$ for $n = 3$). Then there are initial values x^0 so that the equality $\|r^m\|_2 = \|r^0\|_2$ holds for all residuals $r^m = b - Ax^m$ with $m < n$.

Proof. Choose x^0 such that $r^0 = [1 \ 0 \ \dots \ 0]^\top$ is the first unit vector. The solution x^m of (10.39) is of the form $x^m = x^0 + q_{m-1}(A)r^0$. The corresponding residual is $r^m = r^0 - q_{m-1}(A)Ar^0 = p_m(A)r^0$ with the polynomial $p_m(\xi) := 1 - \xi q_{m-1}(\xi)$, i.e., $p_m(\xi) = \sum_{\nu=0}^m a_\nu \xi^\nu$ with $a_0 = 1$. Note that the optimal polynomial p_m minimises $\|r^m\|_2$. One checks that the product $A^\nu r^0$ is the μ -th unit vector with $\mu = n+1-\nu \bmod n$. Hence $r^m = p_m(A)r^0 = [a_0 \ a_{n-1} \ \dots \ a_2 \ a_1]^\top$ yields the squared norm $\|r^m\|_2^2 = \sum_{\nu=0}^m |a_\nu|^2$. The minimum is achieved for $a_1 = a_2 = \dots = a_{n-1} = 0$ resulting in $\|r^m\|_2 = 1 = \|r^0\|_2$. \square

Better results can be obtained if A is Hermitian: $A = A^\text{H}$. However, in this case, the cheaper method of conjugate residuals can be applied, which yields the same iterates (set $N = I$ in Lemma 10.27).

In the case of $A + A^\text{H} > 0$, the convergence can be derived from the convergence of the minimal residual iteration (cf. §9.4).

Proposition 10.35. *Assume $A + A^\text{H} > 0$. Then the residuals of GMRES satisfy $\|r^m\|_2 \leq c^m \|r^0\|_2$ with c in (9.26).*

Proof. By construction, $r^m = p_m(A)r^0$ holds with a polynomial $p_m \in \mathcal{P}_m$ with $p_m(0) = 1$. The minimal residual iteration yields the sequence (\hat{x}^k) with residuals \hat{r}^k . Assume $\hat{r}^0 = r^0$. There are polynomials $q_k \in \mathcal{P}_1$ with $q_k(0) = 1$ and $\hat{r}^k = q_k(A)\hat{r}^{k-1}$. The product $\hat{p}_m(\xi) := \prod_{k=1}^m q_k(\xi)$ satisfies $\hat{r}^m = \hat{p}_m(A)r^0$, $\hat{p}_m \in \mathcal{P}_m$, and $\hat{p}_m(0) = 1$. The optimality of the GMRES algorithm yields $\|r^m\|_2 = \min\{\|\rho_m(A)r^0\|_2 : \rho_m \in \mathcal{P}_m, \rho_m(0) = 1\} \leq \|\hat{r}^m\|_2 \leq c^m \|r^0\|_2$. \square

10.5.1.2 Arnoldi Basis

Let $\{v^1, \dots, v^m\}$ be any basis of $\mathcal{K}_m(A, r^0)$ (this is possible if and only if $m \leq \deg_A(r^0)$). According to (10.8), the minimiser x^m and its residual r^m are characterised by $r^m \perp A\mathcal{K}_m(A, r^0)$. The ansatz $x^m = x^0 + \sum_{\nu=1}^m \alpha_\nu v^\nu$ yields

$$r^m = b - Ax^m = r^0 - \sum_{\nu=1}^m \alpha_\nu Av^\nu,$$

and $r^m \perp A\mathcal{K}_m(A, r^0)$ produces the m equations

$$0 = \langle r^m, Av^\mu \rangle = \langle r^0, Av^\mu \rangle - \sum_{\nu=1}^m \alpha_\nu \langle Av^\nu, Av^\mu \rangle \quad (1 \leq \mu \leq m) \quad (10.40)$$

for the m unknown factors α_ν .

Lemma 10.36. *For regular $A \in \mathbb{K}^{I \times I}$, the matrix $G_m := (\langle Av^\nu, Av^\mu \rangle)_{1 \leq \nu, \mu \leq m}$ is regular for all $m \leq \deg_A(r^0)$ so that the system (10.40) is uniquely solvable.*

Proof. Since A is regular, $\{Av^1, \dots, Av^m\}$ is also a basis of $A\mathcal{K}_m(A, r^0)$. Hence, the Gram matrix G_m is regular. \square

For the actual computation, the basis should be suitably chosen. One strategy is to arrange the vectors v^k such that $\mathcal{K}_m(A, r^0) = \text{span}\{v^1, \dots, v^m\}$ for all $m \leq \deg_A(r^0)$; i.e., $\mathcal{K}_{m+1}(A, r^0) = \text{span}\{\mathcal{K}_m(A, r^0), v^m\}$. For the purpose of stability, the basis should be orthonormal. Finally, the basis should be such that the involved computational work is as small as possible.

Instead of the orthonormalisation procedure in Remark A.26a, we use the *Arnoldi algorithm*:

```
w0 := r0; h0,-1 := \|r0\|2; m := 0;
while hm,m-1 ≠ 0 do
begin vm := wm/hm,m-1;
  for i := 1 to m do him := ⟨Avm, vi⟩;
  wm+1 := Avm - ∑i=1m himvi; hm+1,m := \|wm\|2;
  m := m + 1
end;
```

One easily checks that $\langle v^m, v^i \rangle = \delta_{mi}$; i.e., $(v^i)_{1 \leq i \leq m}$ is an orthonormal basis of $\mathcal{K}_m(A, r^0)$. The construction implies the property

$$Av^m = \sum_{i=1}^{m+1} h_{im}v^i. \quad (10.41)$$

Therefore, $\langle Av^k, v^i \rangle = h_{ik}$ holds, where we define $h_{ik} := 0$ for $i > k + 1$. We form the matrices

$$V_m = [v^1 \ v^2 \ \dots \ v^m] \in \mathbb{K}^{I \times m}, \quad H_m = (h_{ik})_{1 \leq i,k \leq m} \in \mathbb{K}^{m \times m},$$

$$\hat{H}_{m+1} = (h_{ik})_{1 \leq i \leq m+1, 1 \leq k \leq m} \in \mathbb{K}^{(m+1) \times m}.$$

Note that H_m and \hat{H}_{m+1} are Hessenberg matrices, i.e., $h_{ik} = 0$ for $i > k + 1$. From (10.41), we derive

$$V_m^\mathsf{H} AV_m = H_m, \quad V_{m+1}^\mathsf{H} AV_m = \hat{H}_{m+1}.$$

The ansatz $x^m \in x^0 + \mathcal{K}_m(A, r^0)$ becomes $x^m = x^0 + V_m z^m$ for a vector $z^m \in \mathbb{K}^m$ to be determined. The residual is $r^m = r^0 - AV_m z^m$. Note that $r^0 = \|r^0\|_2 v^1 = \|r^0\|_2 V_{m+1} \mathbf{e}^1$ (\mathbf{e}^1 is the first unit vector). Since V_{m+1} is an orthogonal matrix, $V_{m+1} V_{m+1}^\mathsf{H}$ is the orthogonal projection onto $\mathcal{K}_{m+1}(A, r^0)$. Therefore, $\text{range}(AV_m) \subset \mathcal{K}_{m+1}(A, r^0)$ implies that

$$AV_m = (V_{m+1} V_{m+1}^\mathsf{H})(AV_m) = V_{m+1} \hat{H}_{m+1}.$$

Together we obtain

$$r^m = V_{m+1} \left[\|r^0\|_2 \mathbf{e}^1 + \hat{H}_{m+1} z^m \right].$$

Exploiting again the orthogonality of V_{m+1} , we conclude that

$$\|r^m\|_2 = \left\| \left[\|r^0\|_2 \mathbf{e}^1 + \hat{H}_{m+1} z^m \right] \right\|_2$$

has to be minimised over all $z^m \in \mathbb{K}^m$ (cf. Exercise B.22). This is a least-squares problem as considered in Remark B.23: apply the QR decomposition: $\hat{H}_{m+1} = QR$ and solve $Rz^m = -\|r^0\|_2 Q^H \mathbf{e}^1$. Because of the Hessenberg form of \hat{H}_{m+1} , the QR decomposition is rather cheap (m Givens rotations have to be applied).

Remark 10.37 (cost). The cost of the m -th GMRES step is $\mathcal{O}(m \#I)$, so that m steps yield a total amount of $\mathcal{O}(m^2 \#I)$ operations. The storage cost is $\mathcal{O}(m \#I)$.

The reason is that the involved matrix \hat{H}_{m+1} has Hessenberg structure instead of a tridiagonal one. The existence of short recursions as in the classical CG method is connected with the B -normality of A as discussed in Liesen–Saylor [264].

In the case of a Hermitian matrix $A = A^H$, the Hessenberg structure becomes a tridiagonal one and short recursions can be applied. The resulting method is called MINRES (cf. van der Vorst [373, §6.4]).

10.5.1.3 GMRES(m)

The increasing cost mentioned above is the reason for introducing a restart after a fixed number of m steps. After reaching the GMRES iterate x^m , this value is used as the new starting value for the next m GMRES steps. The size of m may be determined by the maximal available storage S_{\max} : $\mathcal{O}(m \#I) \leq S_{\max}$.

Since already for GMRES no convergence statement for $m < \deg_A(e^0)$ could be given in the general case, the situation is even worse for GMRES(m). In this case, not even $x^m = A^{-1}b$ for $m = n$ can be expected. An alternative approach is to restrict the orthogonalisation to the last m directions.

10.5.2 Full Orthogonalisation Method (FOM)

The full orthogonalisation method or Arnoldi method tries to determine $x^m \in x^0 + \mathcal{K}_m(A, r^0)$ such that

$$r^m \perp \mathcal{K}_m(A, r^0)$$

(we recall that $r^m \perp A\mathcal{K}_m(A, r^0)$ holds for GMRES).

In the general case, the method can break down without obtaining the exact solution. For instance, $r^0 \neq 0$ with $\langle Ar^0, r^0 \rangle = 0$ yields a breakdown since $x^1 = x^0 + \alpha r^0$ leads to a zero division in $\alpha = \|r^0\|_2^2 / \langle Ar^0, r^0 \rangle$.

If the method can be performed successfully, $x^{m_0} = A^{-1}b$ holds for the index $m_0 = \deg_A(r^0)$. For a proof, use that $r^{m_0} \in \mathcal{K}_{m_0+1}(A, r^0) = \mathcal{K}_{m_0}(A, r^0)$ can be perpendicular to $\mathcal{K}_{m_0}(A, r^0)$ only if $r^{m_0} = 0$.

10.5.3 Biconjugate Gradient Method and Variants

The biconjugate gradient method (abbreviated as BCG or BiCG) uses two different Krylov subspaces $\mathcal{K}_m(A, r^0)$ and $\mathcal{K}_m(A^\mathsf{H}, r_*^0)$. Here r_*^0 is any vector with $\langle r^0, r_*^0 \rangle \neq 0$. As the original conjugate gradient method, it uses a short recursion for the search directions $p^m \in \mathcal{K}_{m+1}(A, r^0)$ and $p_*^m \in \mathcal{K}_{m+1}(A^\mathsf{H}, r_*^0)$. As a result the residuals are *biconjugate*: $\langle r^i, r_*^j \rangle = 0$ for $i \neq j$, while $\langle Ap^i, p_*^j \rangle = 0$ for $i \neq j$. The formulation of the method goes back to Lanczos [255] and Fletcher [136]. This method does not aim at the minimisation of the error in some norm.

The use of A^H in the algorithm may lead to problems since sometimes only a subroutine for $x \mapsto Ax$ is available. On the other hand, all vectors $v \in \mathcal{K}_m(A^\mathsf{H}, r_*^0)$ have the representation $p_m(A^\mathsf{H})r_*^0$ with some polynomial $p_m \in \mathcal{P}_m$. The arising scalar products $\langle v, x \rangle$ with $x = q_m(A)r^0 \in \mathcal{K}_m(A, r^0)$ can be rewritten as $\langle p_m(A^\mathsf{H})r_*^0, x \rangle = \langle r_*^0, p_m(A)q_m(A)r^0 \rangle$. Fortunately, the products $p_m q_m$ are of the form $p_m^2(\xi)$ or $\xi p_m^2(\xi)$. This gives rise to the conjugate gradient squared method CGS by Sonneveld [344] (see also Sonneveld–Wesseling–de Zeeuw [345]).

A stabilised version of CGS called Bi-CGSTAB is developed by van der Vorst [372]. For details, see the original papers or van der Vorst [373, §7], Kanzow [233, §7], Saad [328, §§7.3–7.4], Gutknecht [173, 174, 175], and Bank–Chan [26].

10.5.4 Further Remarks

Since matrices that are not positive definite require more or less involved CG variants, another remedy is worth being considered. As in §5.5, an indefinite or non-symmetric problem can be preconditioned by a positive definite matrix B , so that for solving $B\delta = d$ the standard CG method can be applied as a secondary iteration.

Concus–Golub [98] and Widlund [396] describe an interesting method for general matrices A that are split into their symmetric and skew-symmetric parts: $A = A_0 + A_1$, $A_0 = \frac{1}{2}(A + A^\mathsf{H})$. For many applications, A_0 proves to be positive definite. A two-sided transformation by $A^{-1/2}$ yields the matrix $A' := I - S$ with the skew-symmetric term $S := A^{-1/2}A_1A^{-1/2}$. The eigenvalues of A' lie in a complex interval instead of a real one (cf. Hageman–Young [212, p. 336]. For the respective CG version, one finds an error estimate with respect to the A_0 -energy norm, depending on $\Lambda := \|A_0^{-1}A_1\|_2$ and leading to the asymptotic convergence rate $1 - \mathcal{O}(1/\Lambda)$. In the cases of systems arising from partial differential equations, Λ is usually h -independent, leading to a convergence rate independent of the discretisation parameter h . For each step of the algorithm, one system $A_0\delta = d$ must be solved. This fact limits practicability. Under similar assumptions, the multigrid iteration of the second kind even achieves a convergence rate $\mathcal{O}(h^\tau)$ with positive (!) exponent τ (cf. §11.9.1).

Young calls *NA symmetrisable* if there is a similarity transformation such that $WNAW^{-1} > 0$. Then there exist a matrix Z with $ZNA > 0$. The methods called ORTHODIR, ORTHOMIN, and ORTHORES are based on this assumption (cf. Hageman–Young [212, pp. 340–346]).

Part III

Special Iterations

Beside classical iterations there are modern iterative techniques, which are the subject of the third part.

The multigrid methods of Chapter 11 are the first iterations achieving linear complexity for a large class of problems. In particular, they apply to discretisations of elliptic boundary value problems. The multigrid method is a recursive algorithm using a product iteration built from a smoothing iteration and a coarse-grid correction. Due to the close connection to the discretisation of elliptic problems, the convergence analysis uses similar tools as the error analysis of finite element methods.

The multigrid method also applies to discretisations of integral equations. This variant is called the multigrid method of the second kind and is historically the first example of a multigrid approach. In this case, the fully populated matrix should be treated by the hierarchical matrix technique mentioned in Appendix D.

Since the present computer architecture is characterised by parallel processors, there is strong interest in distributing the computational effort. This leads to the concept of domain decomposition methods. Various versions of these iterations are discussed in Chapter 12. Although this technique starts from the geometric decomposition of the underlying domain, there are generalisations to an algebraic decomposition leading to the class of subspace iteration methods (cf. §12.5). As described in §12.9, the concept of subspace iterations also applies to multigrid iterations.

An important aspect of the quality of iterative methods is robustness. It is a considerable disadvantage if an iterative technique has to be adapted to any new application case. Then the computer time is minimised at the cost of human working time. A method is called robust if the adaptation to the actual problem is minimal. The hierarchical LU iteration in Chapter 13 is such a robust method. Using the technique of hierarchical matrices, we determine a rather accurate LU decomposition of the underlying matrix and use the forward and backward substitution by LU as preconditioner. This guarantees fast convergence, while the hierarchical matrix computation is of almost linear complexity. The algebraic version in §13.4.2 underlines the robustness of this approach.

In the case of partial differential equations in many variable or in the presence of many parameters, the solution x of the linear systems $Ax = b$ may require a storage far beyond the computer capacity. Since such problems are often tensor-structured, tensor-based methods may be applied for approximating the solution. In Chapter 14 we introduce into the techniques suited for such problems.

Chapter 11

Multigrid Iterations

Abstract Multigrid methods belong to the class of fastest linear iterations, since their convergence rate is bounded independently of the step size h . Furthermore, their applicability does not require symmetry or positive definiteness. Books devoted to multigrid are Hackbusch [183], Wesseling [395], Trottenberg–Oosterlee–Schuller [367], Shaidurov [338], and Vassilevski [378]; see also [205, pp. 1–312].

The ‘smoothing step’ and the ‘coarse-grid correction’ together with the involved restrictions and prolongations are typical ingredients of the multigrid iteration. They are introduced in Section 11.1 for the Poisson model problem. The two-grid iteration explained in Section 11.2 is the first step towards the multigrid method. The iteration matrix is provided in §11.2.3. First numerical examples are presented in §11.2.4.

Before a more general proof of convergence is presented, Section 11.3 investigates the one-dimensional model problem. The proof demonstrates the complementary roles of the smoothing part and the coarse-grid correction. Moreover, the dependence of the convergence rate on the number of smoothing steps is determined.

The multigrid iteration is defined in Section 11.4. Its computational work is discussed and numerical examples are presented. The iteration matrix is described in §11.4.4.

The nested iteration presented in Section 11.5 is a typical technique combined with the multigrid iteration. In principle, it can be combined with any iteration, provided that a hierarchy of discretisations is given. Besides a reduction of the computational work, the nested iteration technique allows us to adjust the iteration error to the discretisation error.

A general convergence analysis of the W-cycle is presented in Section 11.6. Stronger statements are possible in the positive definite case which is studied in Section 11.7. Here, also the V-cycle convergence is proved. As long as lower order terms are responsible for the nonsymmetric structure, the symmetric convergence results can be transferred as shown in §11.7.6. This includes the case of the V-cycle.

Possible combinations with semi-iterative methods are discussed in Section 11.8. Concluding comments are given in Section 11.9.

11.1 Introduction

Multigrid iterations consist of two complementary parts: the *smoothing step* and the *coarse-grid correction*. Below we explain both steps in the case of the Poisson model problem.

11.1.1 Smoothing

Let A be the matrix of the Poisson model problem with step size h . As the simplest example we choose Richardson's iteration:

$$x^{m+1} = \Phi_{\Theta}^{\text{Rich}}(x^m, b) = x^m - \Theta(Ax^m - b) \quad \text{with} \quad (11.1a)$$

$$\Theta = \frac{1}{8}h^2 \approx 1/\lambda_{\max}(A) = 1/\rho(A) \quad (\text{cf. (3.1c)}). \quad (11.1b)$$

$\rho(A) = \lambda_{\max}(A)$ is the eigenvalue corresponding to the eigenfunction

$$e^{\alpha\beta}(x, y) = 2h \sin(\alpha\pi x) \sin(\beta\pi y) \quad (1 \leq \alpha, \beta \leq N-1, (x, y) \in \Omega_h) \quad (11.2a)$$

of the highest frequency $\alpha = \beta = N-1$ (cf. (3.2)). The convergence rate $\rho(M_{\Theta}^{\text{Rich}}) = 1 - \Theta\lambda_{\min} \approx 1 - \lambda_{\min}/\lambda_{\max} \leq 1 - \mathcal{O}(h^2)$ is attained by the lowest frequency $\alpha = \beta = 1$, i.e., when the error $e^m = x^m - x$ is a multiple of the eigenfunction $e^{1,1}$.

All $x \in X := \mathbb{R}^I$ can be represented by the orthonormal eigenvector basis (11.2a):

$$x = \sum_{\alpha, \beta=1}^{N-1} \xi_{\alpha\beta} e^{\alpha\beta} \quad \text{with } \xi_{\alpha\beta} := \langle x, e^{\alpha\beta} \rangle. \quad (11.2b)$$

Since high frequencies α, β correspond to *strong oscillations* of the sine functions (11.2a), we define

$$X_{\text{osc}} := \text{span} \{ e^{\alpha\beta} : 1 \leq \alpha, \beta \leq N-1, \max\{\alpha, \beta\} > \frac{N}{2} \} \quad (11.2c)$$

as a subspace of the *oscillatory components*. Note that at least one of the indices α, β lies in the *high-frequency part* $(N/2, N)$ of the frequency interval $[1, N-1]$. If we are able to generate an approximation x^0 , whose error lies in the subspace X_{osc} ,

$$e^0 := x^0 - x \in X_{\text{osc}} \quad (e^0 \text{ is the error, not an eigenvector!}), \quad (11.2d)$$

the simple Richardson iteration yields fast convergence.

Lemma 11.1. *Assume the Poisson model case with (11.2d). Then all succeeding errors e^m also belong to X_{osc} and satisfy the error estimate*

$$\|e^m\|_2 \leq \frac{3}{4} \|e^{m+1}\|_2, \quad (11.3)$$

i.e., restricted to X_{osc} , the convergence rate is h -independent.

Proof. Since the vectors (11.2a) are orthonormal (cf. Lemma 3.2), we have

$$\|x\|_2^2 = \sum_{\alpha,\beta=1}^{N-1} |\xi_{\alpha\beta}|^2 \quad \text{for } x \text{ in (11.2b).}$$

Because of $M e^{\alpha\beta} = (1 - \Theta \lambda_{\alpha\beta}) e^{\alpha\beta}$, applying the iteration matrix $M = I - \Theta A$ to the error e^m with coefficients $\xi_{\alpha\beta}$ yields e^{m+1} satisfying

$$\begin{aligned} \|e^{m+1}\|_2^2 &\leq \sum |1 - \Theta \lambda_{\alpha\beta}|^2 |\xi_{\alpha\beta}|^2 \leq \max |1 - \Theta \lambda_{\alpha\beta}|^2 \sum |\xi_{\alpha\beta}|^2 \\ &= \max |1 - \Theta \lambda_{\alpha\beta}|^2 \|e^m\|_2^2, \end{aligned}$$

with $\lambda_{\alpha\beta} = 4h^{-2} [\sin^2(\alpha\pi h/2) + \sin^2(\beta\pi h/2)]$ (cf. (3.1a)). The maximum has to be taken over all α, β appearing in (11.2c). By symmetry, we may restrict the frequencies to $0 < \alpha < N$ and $N/2 < \beta < N$. For these α, β ,

$$2h^{-2} = 4h^{-2} \sin^2(\pi/4) < \lambda_{\alpha\beta} \leq \lambda_{N-1, N-1} < 8h^{-2}$$

holds; hence, $|1 - \Theta \lambda_{\alpha\beta}| = |1 - \frac{h^2}{8} \lambda_{\alpha\beta}| < 1 - \frac{h^2}{8} 2h^{-2} = \frac{3}{4}$ proves the desired inequality (11.3). \square

The statement of the lemma is not of direct practical use because the assumption (11.2d) cannot be established in practice (at least not with less work than for solving $Ax = b$ exactly). However, we can conclude the following estimate involving the smooth subspace $X_{\text{sm}} := X_{\text{osc}}^\perp = \text{span}\{e^{\alpha\beta} : 1 \leq \alpha, \beta \leq N/2\}$.

Conclusion 11.2. Split the starting error e^0 into

$$e^0 = e_{\text{osc}}^0 + e_{\text{sm}}^0, \quad e_{\text{osc}}^0 \in X_{\text{osc}}, \quad e_{\text{sm}}^0 \in X_{\text{sm}} := X_{\text{osc}}^\perp.$$

Then, after m steps of Richardson's iteration (11.1a,b), we have

$$\begin{aligned} e^m &= e_{\text{osc}}^m + e_{\text{sm}}^m \quad \text{with} \\ e_{\text{osc}}^m &= M^m e_{\text{osc}}^0 \in X_{\text{osc}}, \quad e_{\text{sm}}^m = M^m e_{\text{sm}}^0 \in X_{\text{sm}}, \\ \|e_{\text{osc}}^m\|_2 &\leq \left(\frac{3}{4}\right)^m \|e_{\text{osc}}^0\|_2, \end{aligned}$$

while e_{sm}^m converges only very slowly to 0. Since e_{osc}^m decreases faster than e_{sm}^m , the smooth part of e^m has increased, and one may regard e^m as ‘smoother’ than e^0 . The smoothness of e^m can be measured by the ratio $\|e_{\text{sm}}^m\|_2 / \|e_{\text{osc}}^m\|_2$.

For illustration purposes, we present the numerical results for the system

$$Ax = b \quad \text{with} \quad A = h^{-2} \text{tridiag}\{-1, 2, -1\} \quad (11.4a)$$

of $n = N - 1 = \frac{1}{h} - 1$ equations corresponding to the one-dimensional Poisson boundary value problem

$$-u''(x) = f(x) \quad \text{for } 0 < x < 1, \quad u(0) = u_0, \quad u(1) = u_1. \quad (11.4b)$$

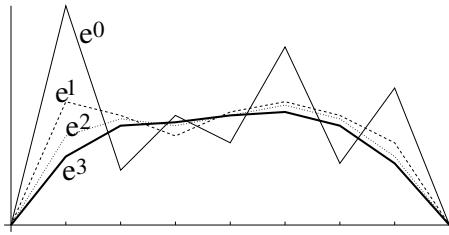


Fig. 11.1 Errors $e^m \in X_{\text{osc}}$ of example (11.4a).

Figure 11.1 shows the (piecewise linearly connected) initial values e_i^0 ($0 \leq i \leq N = 8$) and the errors e^m of the first three Richardson iterates. The errors e^m are insignificantly smaller than e^0 but clearly smoother.

We call iterative methods such as the Richardson iteration (11.1a,b) *smoothing iterations* and use the symbol \mathcal{S} instead of Φ .

Exercise 11.3. The choice of Θ in (11.1b) is not the optimal value. Determine Θ such that the bound in (11.3) becomes minimal.

In the following, an iteration Ψ with a complementary property is desired: Ψ should effectively reduce the smooth components in

$$X_{\text{sm}} = X_{\text{osc}}^\perp = \text{span} \{e^{\alpha\beta} : 1 \leq \alpha, \beta \leq N/2\}.$$

Ψ is not required to have good convergence with respect to the subspace X_{osc} . Then the product method $\Psi \circ \mathcal{S}$ would have the property that for both subspaces X_{osc} and X_{sm} one of the factors in $\Psi \circ \mathcal{S}$ yields fast convergence.

Unfortunately, none of the methods mentioned so far has this property. To construct such an iteration Ψ , we adhere to the concept that smooth grid functions can be well approximated by using a coarser grid. After introducing coarser grids in §§11.1.2–11.1.4, we shall return to the construction of the *coarse-grid correction* Ψ in §11.1.5.

11.1.2 Hierarchy of Systems of Equations

For the following considerations, we have to embed the problem $Ax = b$ into a family of systems. In the model case, for all step sizes $h = 1/N$, we obtain a system $Ax = b$ depending on N or h , respectively. Let

$$h_0 > h_1 > \dots > h_{\ell-1} > h_\ell > \dots \quad \text{with } \lim_{\ell \rightarrow \infty} h_\ell = 0$$

be a sequence of step sizes, which may be generated, e.g., by

$$h_\ell := h_0/2^\ell \quad (\ell \geq 0). \tag{11.5a}$$

The index ℓ is the level-number. $\ell = 0$ corresponds to the coarsest grid. In the model case, for which the grid $\Omega_\ell := \Omega_{h_\ell}$ is contained in the unit square, the step size

$$h_0 = 1/2 \tag{11.5b}$$

is the coarsest one. Then $\Omega_0 = \Omega_{h_0}$ contains only one interior grid point.

Each step size h_ℓ (i.e., each level ℓ) corresponds to a system

$$A_\ell x_\ell = b_\ell \quad (\ell = 0, 1, 2, \dots) \quad (11.6a)$$

of order n_ℓ , which in the model case amounts to

$$n_\ell = (N_\ell - 1)^2 = (1/h_\ell - 1)^2. \quad (11.6b)$$

The family of systems (11.6a) for $\ell = 0, 1, 2, \dots$ represents the hierarchy of systems of equations. The actual problem $Ax = b$ to be solved corresponds to a particular level $\ell = \ell_{\max}$. For solving $A_\ell x_\ell = b_\ell$ at $\ell = \ell_{\max}$, we shall use all lower levels $\ell < \ell_{\max}$.

Remark 11.4. Concerning the construction of the family $\{A_\ell, b_\ell\}_{\ell=0,1,\dots}$ of discretisations, we mention two quite different approaches:

- (A) A maximal level ℓ_{\max} and the corresponding system $A_{\ell_{\max}} x_{\ell_{\max}} = b_{\ell_{\max}}$ are given. Then auxiliary problems $A_\ell x_\ell = b_\ell$ for $\ell < \ell_{\max}$ are created in some way.
- (B) The discretisation starts with $A_0 x_0 = b_0$. Local (or global) grid refinement is used to construct $A_\ell x_\ell = b_\ell$ for $\ell = 1, 2, \dots$ until the discretisation error is sufficiently small.

11.1.3 Prolongation

The vectors x_ℓ and b_ℓ in (11.6a) are elements of the vector space

$$X_\ell = \mathbb{R}^{n_\ell}. \quad (11.7)$$

To connect different levels $\ell = 0, 1, 2, \dots, \ell_{\max}$, we introduce the prolongation

$$p : X_{\ell-1} \rightarrow X_\ell \quad (\ell \geq 1), \quad (11.8)$$

which is assumed to be a linear and injective mapping (more precisely; a family of mappings¹ for all $\ell \geq 1$) from the coarse grid into the fine one.

In the one-dimensional case (11.4a), the vector x_ℓ can be regarded as a grid function defined on $\Omega_\ell = \{\mu h_\ell : 0 < \mu < N_\ell = 1/h_\ell\}$. The vector x_ℓ is rewritten as u_ℓ if it is understood as a grid function on Ω_ℓ with values

$$u_\ell(\mu h_\ell) = x_{\ell,\mu} \quad (1 \leq \mu \leq N_\ell - 1),$$

i.e., for all step sizes the arguments of u_ℓ belong to the interval $\Omega = (0, 1)$ and are restricted to the nodal points in Ω_ℓ . For ease of notation, we include the boundary values

$$u_\ell(0) = u_\ell(1) = 0. \quad (11.9)$$

¹ A more precise notation would be $p_{\ell,\ell-1}$ indicating the involved levels. However, the context will uniquely determine the levels.

An obvious proposal for the prolongation p is piecewise linear interpolation between the grid points of $\Omega_{\ell-1}$:

$$(pu_{\ell-1})(\xi) := u_{\ell-1}(\xi) \quad \text{for } \xi \in \Omega_{\ell-1} \subset \Omega_\ell, \quad (11.10a)$$

$$(pu_{\ell-1})(\xi) := \frac{1}{2} [u_{\ell-1}(\xi + h_\ell) + u_{\ell-1}(\xi - h_\ell)] \quad \text{for } \xi \in \Omega_\ell \setminus \Omega_{\ell-1}, \quad (11.10b)$$

where definition (11.9) is used at $\xi = h_\ell$ and $\xi = 1 - h_\ell$. A shorter characterisation of the prolongation p is the symbol (11.10c):

$$p = \begin{bmatrix} \frac{1}{2} & 1 & \frac{1}{2} \end{bmatrix}. \quad (11.10c)$$

The stencil in (11.10c) indicates that the unit vector $x_{\ell-1} = (\dots, 0, 1, 0, \dots)^T$ is mapped into $x_\ell = px_{\ell-1} = (\dots, 0, \frac{1}{2}, 1, \frac{1}{2}, 0, \dots)^T$.

For the two-dimensional Poisson equation, the vector x_ℓ is represented by the grid function u_ℓ :

$$u_\ell(\xi, \eta) = x_{\ell,ij} \quad \text{for } 1 \leq i, j \leq N_\ell - 1, \quad (\xi, \eta) = (ih_\ell, jh_\ell) \in \Omega_\ell,$$

where the boundary values are defined by

$$u_\ell(\xi, \eta) := 0 \quad \text{for } \xi = 0 \text{ or } \xi = 1 \text{ or } \eta = 0 \text{ or } \eta = 1.$$

The two-dimensional generalisation of the piecewise linear interpolation (11.10a,b) (bilinear interpolation) reads as follows:

$$(pu_{\ell-1})(\xi, \eta) := u_{\ell-1}(\xi, \eta) \quad \text{for } (\xi, \eta) \in \Omega_{\ell-1} \subset \Omega_\ell,$$

$$(pu_{\ell-1})(\xi, \eta) := \frac{1}{2} [u_{\ell-1}(\xi + h_\ell, \eta) + u_{\ell-1}(\xi - h_\ell, \eta)] \quad \text{for } \xi/h_\ell \text{ odd, } \eta/h_\ell \text{ even,}$$

$$(pu_{\ell-1})(\xi, \eta) := \frac{1}{2} [u_{\ell-1}(\xi, \eta + h_\ell) + u_{\ell-1}(\xi, \eta - h_\ell)] \quad \text{for } \xi/h_\ell \text{ even, } \eta/h_\ell \text{ odd,}$$

$$(pu_{\ell-1})(\xi, \eta) := \frac{1}{4} \left[\begin{array}{l} u_{\ell-1}(x + h_\ell, h + h_\ell) + u_{\ell-1}(x - h_\ell, h - h_\ell) \\ + u_{\ell-1}(x - h_\ell, h + h_\ell) + u_{\ell-1}(x + h_\ell, h - h_\ell) \end{array} \right] \quad \text{for } \xi/h_\ell \text{ and } \eta/h_\ell \text{ odd.}$$

The abbreviation of p defined above is the star

$$p = \begin{bmatrix} 1/4 & 1/2 & 1/4 \\ 1/2 & 1 & 1/2 \\ 1/4 & 1/2 & 1/4 \end{bmatrix} \quad (\text{nine-point prolongation}), \quad (11.11)$$

since the application of p to a unit vector yields the values indicated in (11.11) extended by zero is the remaining grid.

In general, the stencil

$$p = \begin{bmatrix} \pi_{-1,1} & \pi_{0,1} & \pi_{1,1} \\ \pi_{-1,0} & \pi_{0,0} & \pi_{1,0} \\ \pi_{-1,-1} & \pi_{0,-1} & \pi_{1,-1} \end{bmatrix} \quad (11.12)$$

describes the following mapping, where the summation is taken over all i, j with $(\xi - ih_\ell, \eta - jh_\ell) \in \Omega_{\ell-1}$:

$$(pu_{\ell-1})(\xi, \eta) := \sum_{i,j} \pi_{ij} u_{\ell-1}(\xi - ih_\ell, \eta - jh_\ell) \quad \text{for } (\xi, \eta) \in \Omega_\ell.$$

Other linear interpolations as well as prolongations of higher order are discussed by Hackbusch [183, §3.4]. A so-called *matrix-dependent prolongation* is defined by (11.12) with the coefficients

$$\pi_{00} := 1, \quad \pi_{\pm 1,0} := -\frac{\sum_j \alpha_{\mp 1,j}}{\sum_j \alpha_{0,j}}, \quad \pi_{0,\pm 1} := -\frac{\sum_i \alpha_{i,\mp 1}}{\sum_i \alpha_{i,0}}, \quad (11.13a)$$

$$(A_\ell p u_{\ell-1})(\xi, \eta) = 0 \quad \text{for } \xi/h_\ell \text{ and } \eta/h_\ell \text{ odd,} \quad (11.13b)$$

where $\alpha_{i,j}$ are the coefficients of A_ℓ according to (1.13a,b). Condition (11.13b) determines $\pi_{\pm 1, \pm 1}$ (cf. Hackbusch [183, §10.3] and de Zeeuw [104]).

11.1.4 Restriction

The restriction r is a linear and surjective mapping

$$r : X_\ell \rightarrow X_{\ell-1} \quad (\ell \geq 1),$$

which maps fine-grid functions into coarse-grid functions. If $\Omega_{\ell-1} \subset \Omega_\ell$ holds as in the model case, the simplest choice is the trivial restriction

$$(r_{\text{triv}} u_\ell)(\xi, \eta) = u_\ell(\xi, \eta) \quad \text{for } (\xi, \eta) \in \Omega_{\ell-1}.$$

However, because of certain disadvantages, we advise against its use (cf. Hackbusch [183, §3.5]). Instead, we define $(ru_\ell)(\xi, \eta)$ as the weighted mean of the neighbouring values. The stencil

$$r = \begin{bmatrix} \rho_{-1,1} & \rho_{0,1} & \rho_{1,1} \\ \rho_{-1,0} & \rho_{0,0} & \rho_{1,0} \\ \rho_{-1,-1} & \rho_{0,-1} & \rho_{1,-1} \end{bmatrix} \quad (11.14)$$

characterises the restriction

$$(ru_\ell)(\xi, \eta) = \sum_{i,j=-1}^1 \rho_{ij} u_\ell(\xi + ih_\ell, \eta + jh_\ell) \quad \text{for } (\xi, \eta) \in \Omega_{\ell-1}.$$

The nine-point prolongation (11.11) corresponds to the nine-point restriction

$$r = \frac{1}{4} \begin{bmatrix} 1/4 & 1/2 & 1/4 \\ 1/2 & 1 & 1/2 \\ 1/4 & 1/2 & 1/4 \end{bmatrix}, \quad (11.15)$$

which can be considered as the adjoint to (11.11), where the definition of adjoint mappings is based on the scalar products

$$\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_\ell \quad \text{with} \quad \langle u_\ell, v_\ell \rangle_\ell = h_\ell^d \sum_{\alpha \in I} u_{\ell,\alpha} \overline{v_{\ell,\alpha}} \quad (11.16)$$

for X_ℓ . d is the dimension of the grid $\Omega_\ell \subset \mathbb{R}^d$. The adjoint mapping is denoted by p^* . Since p can also be considered as a matrix, the transposed matrix p^\top is defined. Because of the different weighting factors h_ℓ^d in (11.16), p^* and p^\top differ by a factor as stated in the next exercise.

Exercise 11.5. Assume the two-dimensional case $d = 2$ and prove that the mapping adjoint to p defined in (11.12) is r in (11.14) with $\rho_{ij} = \pi_{ij}/4$. Prove for general d that $p^* = 2^{-d} p^\top$.

Having fixed the prolongation, we can always choose the adjoint mapping

$$r := p^* \quad (11.17)$$

as a restriction. For example, we can define a matrix-dependent restriction by (11.13a,b) and (11.17).

11.1.5 Coarse-Grid Correction

Let \bar{x}_ℓ be the result of a few steps of the smoothing iteration (11.1a,b). The corresponding error $\bar{e}_\ell := \bar{x}_\ell - x_\ell$ is the exact correction; i.e., the solution can be obtained by

$$x_\ell = \bar{x}_\ell - \bar{e}_\ell.$$

Since $A_\ell \bar{e}_\ell = A_\ell(\bar{x}_\ell - x_\ell) = A_\ell \bar{x}_\ell - A_\ell x_\ell = A_\ell \bar{x}_\ell - b_\ell$, the correction \bar{e}_ℓ satisfies the equation

$$A_\ell \bar{e}_\ell = d_\ell \quad \text{with the defect } d_\ell := A_\ell \bar{x}_\ell - b_\ell. \quad (11.18a)$$

According to considerations in §11.1.1, \bar{e}_ℓ is smooth. Therefore, it should be possible to approximate \bar{e}_ℓ by using the coarse grid: $\bar{e}_\ell \approx p e_{\ell-1}$. As ansatz for $e_{\ell-1}$, we take the coarse-grid equation corresponding to (11.18a):

$$A_{\ell-1} e_{\ell-1} = d_{\ell-1} \quad \text{with } d_{\ell-1} := r d_\ell. \quad (11.18b)$$

Assume that we are able to solve the coarse-grid equation (11.18b) exactly:

$$e_{\ell-1} = A_{\ell-1}^{-1} d_{\ell-1}. \quad (11.18c)$$

Its image $p e_{\ell-1}$ under the prolongation p should approximate the solution \bar{e}_ℓ of (11.18a), so that the coarse-grid correction is completed by

$$x_\ell^{\text{new}} := \bar{x}_\ell - p e_{\ell-1}. \quad (11.18d)$$

In compact form, the coarse-grid correction (11.18a–d) reads as follows:

$$\bar{x}_\ell \longmapsto x_\ell^{\text{new}} := \bar{x}_\ell - p A_{\ell-1}^{-1} r(A_\ell \bar{x}_\ell - b_\ell).$$

Renaming \bar{x}_ℓ and x_ℓ^{new} by x_ℓ^m and x_ℓ^{m+1} , the mapping above defines an iterative method which we call the *coarse-grid correction*:

$$\Phi_\ell^{\text{CGC}}(x_\ell, b_\ell) := x_\ell - p A_{\ell-1}^{-1} r(A_\ell x_\ell - b_\ell). \quad (11.19)$$

Remark 11.6. The iteration matrix M_ℓ^{CGC} and the matrix N_ℓ^{CGC} of the second normal form of the coarse-grid correction are

$$M_\ell^{\text{CGC}} = I - p A_{\ell-1}^{-1} r A_\ell, \quad N_\ell^{\text{CGC}} = p A_{\ell-1}^{-1} r.$$

Φ_ℓ^{CGC} (as such without smoothing) is not an interesting iteration as stated next.

Remark 11.7. The coarse-grid correction Φ_ℓ^{CGC} is consistent, but not convergent.

Proof. The consistency is a consequence of the second normal form. $n_\ell > n_{\ell-1}$ implies $\dim X_\ell > \dim X_{\ell-1}$; hence, the kernel of the restriction r is nontrivial. Let $0 \neq x \in \ker(r)$. Since $M_\ell^{\text{CGC}} \eta = \eta$ for $\eta := A_\ell^{-1} x$, the matrix M_ℓ^{CGC} has an eigenvalue $\lambda = 1$, so that $\rho(M_\ell^{\text{CGC}}) \geq 1$ indicates divergence. \square

For systems obtained from Galerkin discretisation (cf. Proposition E.16 and Hackbusch [183, Note 3.6.6]), the so-called *Galerkin product* representation of $A_{\ell-1}$ is valid:

$$A_{\ell-1} = r A_\ell p. \quad (11.20)$$

Remark 11.8. Given $A = A_{\ell_{\max}}$, one can use (11.20) as a recursive definition of the coarse-grid matrices A_ℓ for $\ell = \ell_{\max} - 1, \dots, 0$, provided that suitable mappings r and p are available (see case (A) in Remark 11.4). If one uses the definition (11.17) of r , only the prolongations p have to be defined.

Lemma 11.9. Assume (11.20). Then $\Phi_\ell^{\text{CGC}}(\hat{x}_\ell, b_\ell) = \hat{x}_\ell$ holds for all vectors \hat{x}_ℓ with $\hat{e}_\ell = \hat{x}_\ell - x_\ell \in \text{range}(p)$; i.e., Φ_ℓ^{CGC} is a projection (cf. Definition 5.12).

Proof. Use $M_\ell^{\text{CGC}} p = p - p A_{\ell-1}^{-1} r A_\ell p = p - p A_{\ell-1}^{-1} A_{\ell-1} = p - p = 0$. \square

The next exercise shows that the coarse-grid equation (11.18b) is a reasonable ansatz for $e_{\ell-1}$.

Exercise 11.10. Let A_ℓ be positive definite. The best approximation of $\bar{e}_\ell \in X_\ell$ with respect to the A_ℓ norm $\|x_\ell\|_A := \langle A_\ell x_\ell, x_\ell \rangle_\ell^{1/2}$ is $p e_{\ell-1}$, where $p = r^*$ according to (11.17) and $e_{\ell-1}$ is the solution of (11.18b) with the Galerkin matrix (11.20).

11.2 Two-Grid Method

11.2.1 Algorithm

The smoothing iteration \mathcal{S}_ℓ is defined in §11.1.1 and the coarse-grid correction Φ_ℓ^{CGC} is constructed in §11.1.5. The two-grid iteration is the product iteration

$$\Phi_\ell^{\text{TGM}} := \Phi_\ell^{\text{CGC}} \circ \mathcal{S}_\ell^\nu \quad (\ell \geq 1, \nu \geq 1)$$

(cf. §5.4), where ν is the number of smoothing steps. In algorithmic notation, the iteration Φ_ℓ^{TGM} takes the form

function	$\Phi_\ell^{\text{TGM}}(x_\ell, b_\ell);$	(11.21a)
begin	for $i := 1$ to ν do $x_\ell := \mathcal{S}_\ell(x_\ell, b_\ell);$	(11.21a)
	$d_{\ell-1} := r(A_\ell x_\ell - b_\ell);$	(11.21b)
	$e_{\ell-1} := A_{\ell-1}^{-1} d_{\ell-1};$	(11.21c)
	$x_\ell := x_\ell - p e_{\ell-1};$	(11.21d)
	$\Phi_\ell^{\text{TGM}} := x_\ell$	(11.21e)
end;		

11.2.2 Modifications

As stated in Proposition 5.25b, $\Phi_\ell^{\text{CGC}} \circ \mathcal{S}_\ell^\nu$ has the same convergence behaviour as

$$\Phi_\ell^{\text{TGM}(\nu_1, \nu_2)} := \mathcal{S}_\ell^{\nu_2} \circ \Phi_\ell^{\text{CGC}} \circ \mathcal{S}_\ell^{\nu_1} \quad \text{with } \nu = \nu_1 + \nu_2. \quad (11.22a)$$

In this case, ν_1 pre- and ν_2 post-smoothing steps are applied. Algorithm (11.21) is the special case of iteration (11.22a) with $\nu_1 = \nu$ and $\nu_2 = 0$. In the sequel, we use the more general version (11.22a).

One may also use different iterations \mathcal{S}_ℓ and $\hat{\mathcal{S}}_\ell$ as pre- and post-smoothers:

$$\hat{\mathcal{S}}_\ell^{\nu_2} \circ \Phi_\ell^{\text{CGC}} \circ \mathcal{S}_\ell^{\nu_1}. \quad (11.22b)$$

A semi-iterative smoothing instead of (11.21a) will be discussed in §11.8.1.

11.2.3 Iteration Matrix

Lemma 11.11. *Let \mathcal{S}_ℓ be a consistent iteration with iteration matrix S_ℓ . Then $\Phi_\ell^{\text{TGM}(\nu_1, \nu_2)}$ is a consistent iteration with the iteration matrix*

$$M_\ell^{\text{TGM}(\nu_1, \nu_2)} = S_\ell^{\nu_2} (I - p A_{\ell-1}^{-1} r A_\ell) S_\ell^{\nu_1}. \quad (11.23)$$

Proof. According to Proposition 5.25b, M_ℓ^{TGM} is the product of the iteration matrices of $\hat{\mathcal{S}}_\ell^{\nu_2}$, Φ_ℓ^{CGC} , $\mathcal{S}_\ell^{\nu_1}$. Equation (11.23) follows from Remark 11.6. \square

Since \mathcal{S}_ℓ is consistent, the matrix $N_\ell^{\text{TGM}(\nu_1, \nu_2)}$ of the second normal form is implicitly determined by

$$M_\ell^{\text{TGM}(\nu_1, \nu_2)} = I - N_\ell^{\text{TGM}(\nu_1, \nu_2)} A_\ell.$$

As known from (5.12b), the second normal form matrices do not have a simple representation for product iterations. The same statement holds for $W_\ell^{\text{TGM}(\nu_1, \nu_2)}$.

11.2.4 Numerical Examples

As an example we choose the two-dimensional Poisson model problem with the step sizes h_ℓ in (11.5a,b). $A_{\ell_{\max}}$ and the auxiliary matrices A_ℓ ($\ell < \ell_{\max}$) are defined by the five-point discretisation (1.4a). The two-grid parameters are $\nu_1 = 2$, $\nu_2 = 0$. The smoothing iteration is the chequer-board variant of the Gauss-Seidel iteration (cf. (1.20)). The error norms $\|e_\ell^m\|_2 = \|x_\ell^m - x_\ell\|_2$ at level $\ell = 5$ with $h_5 = 1/64$ are shown in Table 11.1. Table 11.2 contains the reduction factors $\|x_\ell^m - x_\ell\|_2 / \|x_\ell^{m-1} - x_\ell\|_2$. The last row in Table 11.2 shows the averaged convergence factors $\rho_\ell := (\|e_\ell^8\|_2 / \|e_\ell^0\|_2)^{1/8}$. In contrast to the foregoing iterative methods, the convergence factors hardly depend on the step size. Furthermore, the convergence rate of about 0.06 is very favourable.

Since the two-grid method depends on the parameters ν_1 and ν_2 , their influence on the convergence is be investigated. As mentioned in §11.2.2, the convergence rate depends only on $\nu = \nu_1 + \nu_2$. Therefore, we may choose $\nu_1 = \nu$ and $\nu_2 = 0$ without loss of generality. The convergence factors ρ_ℓ for $h_3 = 1/16$ determined as above are shown in Table 11.3. As expected, convergence improves with increasing ν . In the last row, $\rho_3(\nu)$ is compared with the function $C/(C + \nu)$ for $C = 0.135$. It suggests the asymptotic behaviour $\rho_\ell(\nu) \approx \mathcal{O}(1/\nu)$.

m	$\ x_\ell^m - x_\ell\ _2$
0	2.935 ₁₀₋₀₂
1	1.210 ₁₀₋₀₃
2	6.206 ₁₀₋₀₅
3	3.378 ₁₀₋₀₆
4	1.939 ₁₀₋₀₇
5	1.152 ₁₀₋₀₈
6	7.058 ₁₀₋₁₀
7	4.432 ₁₀₋₁₁
8	7.188 ₁₀₋₁₂

Table 11.1 Iteration errors for $h_5 = \frac{1}{64}$.

m	$h_\ell = \frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$
1	0.10391	0.10420	0.07778	0.05465	0.03807	0.02661
2	0.06210	0.05549	0.04730	0.04336	0.04121	0.04009
3	0.06248	0.05738	0.05409	0.05238	0.05132	0.05077
4	0.06250	0.05851	0.05804	0.05565	0.05445	0.05375
5	0.06250	0.05963	0.06191	0.05866	0.05741	0.05665
6	0.06250	0.06061	0.06512	0.06089	0.05937	0.05835
7	0.06250	0.06143	0.06768	0.06292	0.06124	0.05996
8	0.06250	0.06208	0.06954	0.06463	0.06285	0.06132
ρ_ℓ	0.06654	0.06360	0.06203	0.05626	0.05248	0.04939

Table 11.2 Error ratios $\|x_\ell^m - x_\ell\|_2 / \|x_\ell^{m-1} - x_\ell\|_2$ and averaged convergence factors ρ_ℓ for the two-grid method with $\nu_1 = 2$ and $\nu_2 = 0$.

ν	1	2	3	4	5	6	10
$\rho_3(\nu)$	0.222	0.062	0.04	0.03	0.023	0.0196	0.0133
$\frac{0.135}{0.135+\nu}$	0.119	0.063	0.043	0.033	0.026	0.022	0.0133

Table 11.3 Convergence factors for different smoothing numbers ν .

11.3 Analysis for a One-Dimensional Example

In principle, one can analyse the two-grid convergence for the two-dimensional Poisson model problem (cf. Hackbusch [183, §8.1.1]); however, it is not sufficiently transparent for an introductory consideration. Therefore, we consider the tridiagonal equation (11.4a):

$$Ax = b \quad \text{with } A = h^{-2} \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix} \quad (11.24)$$

discretising the one-dimensional Poisson equation (11.4b). It should be emphasised that tridiagonal matrices are easy to solve directly. Analysis of iterative methods for these tridiagonal equations is of interest only because of the fact that the convergence properties also carry over to the general case of two or more spatial dimensions. Furthermore, this chapter serves as a demonstration of how model problems can be investigated by the help of Fourier analysis.

11.3.1 Fourier Analysis

We abbreviate the quantities at levels ℓ and $\ell - 1$ by

$$N = N_\ell, \quad N' = N_{\ell-1}, \quad h = h_\ell = 1/N, \quad h' = h_{\ell-1} = 2h.$$

The vector $x = (x_k)_{1 \leq k \leq N-1}$ is formally extended by the components

$$x_0 = x_N = 0. \quad (11.25a)$$

The vectors (grid functions) e^α with the components

$$e_k^\alpha = \sqrt{2h} \sin(\alpha k \pi h) \quad (0 \leq k \leq N) \quad (11.25b)$$

satisfy condition (11.25a) for all frequencies $\alpha \in \mathbb{Z}$. According to Exercise 3.3, $\{e^\alpha : 1 \leq \alpha \leq N-1\}$ forms an orthonormal basis. Therefore, the matrix Q built by e^α as columns is unitary: $Q^H Q = I$ (cf. Definition A.27):

$$Q := \left[e^1, e^{N-1}, e^2, e^{N-2}, \dots, e^\alpha, e^{N-\alpha}, \dots, e^{\frac{N}{2}-1}, e^{\frac{N}{2}+1}, e^{\frac{N}{2}} \right]. \quad (11.25c)$$

The reason for the special ordering of the columns will be seen next. $M := M_\ell^{\text{TGM}}$ denotes the iteration matrix of the two-grid method. Since multiplying by Q or

$Q^H = Q^{-1}$ does not change the spectral norm and spectral radius (cf. Lemma B.18), we conclude that

$$\|\hat{M}\|_2 = \|M\|_2, \quad \rho(\hat{M}) = \rho(M) \quad \text{for } \hat{M} := Q^{-1}MQ. \quad (11.25d)$$

\hat{M} is the Fourier-transformed iteration matrix. We shall show in §11.3.2 that \hat{M} has a block-diagonal structure:

$$\begin{aligned} \hat{M} &= \text{blockdiag} \{M_1, M_2, \dots, M_{N'-1}, M_{N'}\} \quad \text{with} \\ M_\alpha &: 2 \times 2 \text{ matrices for } 1 \leq \alpha \leq N' - 1, \quad M_{N'} : 1 \times 1 \text{ matrix.} \end{aligned} \quad (11.25e)$$

Applying (A.10) to \hat{M} and $\hat{M}^H \hat{M}$, we obtain the next statement.

Lemma 11.12. *Matrices of the form (11.25e) satisfy*

$$\|\hat{M}\|_2 = \max_{1 \leq \alpha \leq N'} \|M_\alpha\|_2, \quad \rho(\hat{M}) = \max_{1 \leq \alpha \leq N'} \rho(M_\alpha).$$

We choose the Richardson iteration with $\Theta = \frac{h^2}{4} \approx \frac{1}{\rho(A_\ell)}$ as the smoothing iteration. For proving the block structure (11.25e), we transform the iteration matrix

$$M = (I - pA_{\ell-1}^{-1}rA_\ell)S_\ell^\nu \quad \text{with } S_\ell = I - \Theta A_\ell, \quad \Theta = h^2/4.$$

The Fourier transform applied to the matrices A_ℓ , S_ℓ yields

$$\hat{A}_\ell := Q^{-1}A_\ell Q, \quad \hat{S}_\ell := Q^{-1}S_\ell Q. \quad (11.26a)$$

Next, we need a Fourier map $Q' : X_{\ell-1} \rightarrow X_{\ell-1}$ defined in the coarse grid space. It is defined by

$$Q' = [e'^1, e'^2, \dots, e'^{N'-1}]$$

with the orthonormal columns

$$e'_k{}^\alpha = \sqrt{4h} \sin(2\alpha k \pi h) \quad (0 \leq k \leq N'). \quad (11.26b)$$

The vectors $e'^\alpha \in X_{\ell-1}$ are obtained from e^α in (11.25b) by replacing $h = h_\ell$ with $h' = h_{\ell-1}$. Now p , $A_{\ell-1}$, and r can be transformed into

$$\hat{p} = Q^{-1}pQ', \quad \hat{A}_{\ell-1} := Q'^{-1}A_{\ell-1}Q', \quad \hat{r} := Q'^{-1}rQ.$$

One verifies that \hat{M} in (11.25d) takes the form

$$\hat{M} = (I - \hat{p}\hat{A}_{\ell-1}^{-1}\hat{r}\hat{A}_\ell)\hat{S}_\ell^\nu$$

(check that $\hat{p}\hat{A}_{\ell-1}^{-1}\hat{r}\hat{A}_\ell = Q^{-1}pA_{\ell-1}^{-1}rA_\ell Q$).

11.3.2 Transformed Quantities

Now we prove that all factors \hat{p} , $\hat{A}_{\ell-1}^{-1}$, \hat{r} , \hat{A}_ℓ , \hat{S}_ℓ^ν are block-diagonal as stated in (11.25e). According to §3.1, $A_\ell e^\alpha = \lambda_\alpha e^\alpha$ holds with $\lambda_\alpha = 4h^{-2} \sin^2(\alpha\pi h/2)$. We introduce

$$s_\alpha^2 = \sin^2(\alpha\pi h/2), \quad c_\alpha^2 = \cos^2(\alpha\pi h/2).$$

Noting that $\lambda_{N-\alpha} = s_{N-\alpha}^2 = c_\alpha^2$, we obtain

$$\hat{A}_\ell := Q^{-1} A_\ell Q = \text{blockdiag}\{A_1, \dots, A_{N'}\} \quad \text{with the blocks} \quad (11.27a)$$

$$A_\alpha = 4h^{-2} \begin{bmatrix} s_\alpha^2 & 0 \\ 0 & c_\alpha^2 \end{bmatrix} \quad \text{for } 1 \leq \alpha \leq N'-1, \quad A_{N'} = 2h^{-2}. \quad (11.27b)$$

Since $S_\ell = I - \frac{1}{4}h^2 A_\ell$ and $s_\alpha^2 + c_\alpha^2 = 1$, equations (11.27a,b) yield the result

$$\hat{S}_\ell = Q^{-1} S_\ell Q = \text{blockdiag}\{S_1, \dots, S_{N'}\} \quad \text{with the blocks} \quad (11.27c)$$

$$S_\alpha = \begin{bmatrix} c_\alpha^2 & 0 \\ 0 & s_\alpha^2 \end{bmatrix} \quad \text{for } 1 \leq \alpha \leq N'-1, \quad S_{N'} = \frac{1}{2}. \quad (11.27d)$$

Because of $A_{\ell-1} e'^\alpha = \lambda'_\alpha e'^\alpha$ with $\lambda'_\alpha = 4h'^{-2} \sin^2(\alpha\pi h'/2) = h^{-2} \sin^2(\alpha\pi h)$ and using $\sin^2(\alpha\pi h) = 4s_\alpha^2 c_\alpha^2$, we obtain the diagonal matrix

$$\hat{A}_{\ell-1} := Q'^{-1} A_{\ell-1} Q' = \text{diag}\{A'_1, \dots, A'_{N'}\} \quad \text{with } A'_\alpha = \frac{4}{h^2} s_\alpha^2 c_\alpha^2. \quad (11.27e)$$

Next, we transform p and r . Let p be defined by (11.10a–c). For r , we choose the adjoint mapping $r = p^*$:

$$r = \frac{1}{2} \begin{bmatrix} \frac{1}{2} & 1 & \frac{1}{2} \end{bmatrix}, \quad \text{i.e., } (ru_\ell)(\xi) = \frac{1}{4}u_\ell(\xi-h) + \frac{1}{2}u_\ell(\xi) + \frac{1}{4}u_\ell(\xi+h). \quad (11.27f)$$

r and \hat{r} are matrices of the format $(N'-1) \times (N-1) = (N'-1) \times (2N'-1)$. The representation

$$\hat{r} := [\text{blockdiag}\{r_1, \dots, r_{N'-1}\}, 0] \quad \text{with } r_\alpha = \sqrt{\frac{1}{2}} \begin{bmatrix} c_\alpha^2 & -s_\alpha^2 \end{bmatrix} \quad (11.27g)$$

means that the last column of $\hat{r} := Q'^{-1} r Q$ vanishes (this follows from $r e^{N'} = 0$) and that the remaining part of the format $(N'-1) \times (2N'-1)$ consists of $N'-1$ blocks r_α of size 1×2 . For the proof of (11.27g), it must be shown that

$$r e^\alpha = c_\alpha^2 e'^\alpha / \sqrt{2}, \quad r e^{N-\alpha} = -s_\alpha^2 e'^\alpha / \sqrt{2} \quad \text{for } 1 \leq \alpha \leq N'-1.$$

The restriction (11.27f) yields

$$\begin{aligned} r \sin(\alpha x \pi) &= [\sin(\alpha(x-h)\pi) + 2\sin(\alpha x \pi) + \sin(\alpha(x+h)\pi)]/4 \\ &= [1 + \cos(\alpha h \pi)] \sin(\alpha x \pi)/2 = \cos(\alpha h \pi/2)^2 \sin(\alpha x \pi) = c_\alpha^2 \sin(\alpha x \pi) \end{aligned}$$

for all frequencies α . The different scaling of the vectors e^α, e'^α explains the additional factor in $re^\alpha = c_\alpha^2 e'^\alpha / \sqrt{2}$. Since this identity holds for all $\alpha \in \mathbb{Z}$, we may replace α by $N - \alpha$: $re^{N-\alpha} = c_{N-\alpha}^2 e'^{N-\alpha} / \sqrt{2}$. For $0 \leq k \leq N'$, the equality $\sin(2\alpha k \pi h) = -\sin(2(N - \alpha)k \pi h)$ leads to $e'^{N-\alpha} = -e'^\alpha$ (cf. definition (11.26b)). Finally, $c_{N-\alpha}^2 = s_\alpha^2$ proves $re^{N-\alpha} = -s_\alpha^2 e'^\alpha / \sqrt{2}$.

p in (11.10a–c) and r in (11.27f) are connected by $r = p^*$. From $p^* = \frac{1}{2} p^\top$, we derive the representation

$$\hat{p} = Q^{-1} p Q' = Q^{-1} (2r)^\top Q' = Q^\top (2r)^\top Q' = 2 [Q'^\top r Q]^\top = 2 \hat{r}^\top.$$

Therefore, the result (11.27g) for \hat{r} proves

$$\hat{p} = Q^{-1} p Q' = \begin{bmatrix} \text{blockdiag } \{p_1, \dots, p_{N'-1}\} \\ 0 \end{bmatrix} \quad (11.27h)$$

with $p_\alpha = \sqrt{2} \begin{bmatrix} c_\alpha^2 \\ -s_\alpha^2 \end{bmatrix}$.

11.3.3 Convergence Results

Since all factors in (11.26a) have a block-diagonal structure, this carries over to \hat{M} and proves the structure (11.25e). For the 2×2 blocks M_α ($1 \leq \alpha \leq N' - 1$) and the 1×1 block $M_{N'}$, the statements (11.27b, d, e, g, h) yield

$$M_\alpha = (I - p_\alpha A_\alpha'^{-1} r_\alpha A_\alpha) S_\alpha^\nu \quad (1 \leq \alpha \leq N' - 1), \quad M_{N'} = 2^{-\nu}.$$

Inserting the representations of $p_\alpha, A_\alpha', r_\alpha, A_\alpha, S_\alpha^\nu$, we obtain

$$\begin{aligned} M_\alpha &= \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} c_\alpha^2 \\ -s_\alpha^2 \end{bmatrix} \frac{h^2}{4s_\alpha^2 c_\alpha^2} \begin{bmatrix} c_\alpha^2 & -s_\alpha^2 \end{bmatrix} 4h^{-2} \begin{bmatrix} s_\alpha^2 & 0 \\ 0 & c_\alpha^2 \end{bmatrix} \right) \begin{bmatrix} c_\alpha^2 & 0 \\ 0 & s_\alpha^2 \end{bmatrix}^\nu \quad (11.28) \\ &= \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} c_\alpha^2 & -c_\alpha^2 \\ -s_\alpha^2 & s_\alpha^2 \end{bmatrix} \right) \begin{bmatrix} c_\alpha^2 & 0 \\ 0 & s_\alpha^2 \end{bmatrix}^\nu = \begin{bmatrix} s_\alpha^2 & c_\alpha^2 \\ s_\alpha^2 & c_\alpha^2 \end{bmatrix} \begin{bmatrix} c_\alpha^2 & 0 \\ 0 & s_\alpha^2 \end{bmatrix}^\nu. \end{aligned}$$

The block M_α describes the application of M to the two functions $e^\alpha, e^{N-\alpha}$ (the respective columns of the matrix Q , cf. (11.25c)). Since $\alpha < N' < N - \alpha$, e^α corresponds to a smooth grid function and $e^{N-\alpha}$ to an oscillatory one. Obviously, the inequalities $0 < \alpha < N' < N - \alpha < N$ lead to

$$0 < s_\alpha^2 < \frac{1}{2} < c_\alpha^2 < 1. \quad (11.29)$$

The two 2×2 matrices in (11.28) characterise the coarse-grid correction and the smoothing iteration, respectively. Let the error have a representation $\sum_{\alpha=1}^{N-1} \xi_\alpha e^\alpha$ as in (11.2b). The entries $c_\alpha^2 > s_\alpha^2$ express the fact that the smooth e^α -components

converge more slowly than the nonsmooth $e^{N-\alpha}$ -components. The first matrix reflects the complementary behaviour of the coarse-grid correction: The smooth components (s_α^2 in the first column) are better reduced than the oscillatory ones (c_α^2 in the second column).

Exercise 11.13. Prove that $\rho(M_\alpha) = \rho_\nu(s_\alpha^2)$ and $\|M_\alpha\|_2 = \zeta_\nu(s_\alpha^2)$ with

$$\rho_\nu(\xi) := \xi(1-\xi)^\nu + (1-\xi)\xi^\nu, \quad (11.30a)$$

$$\zeta_\nu(\xi) := \sqrt{2 \left[\xi^2 (1-\xi)^{2\nu} + (1-\xi)^2 \xi^{2\nu} \right]}. \quad (11.30b)$$

Combining (11.25d), Lemma 11.12, and Exercise 11.13 yields

$$\rho(M) = \max\{\rho_\nu(s_\alpha^2) : 1 \leq \alpha \leq N'\}, \quad \|M_\alpha\|_2 = \max\{\zeta_\nu(s_\alpha^2) : 1 \leq \alpha \leq N'\}.$$

Since the values of s_α^2 for $1 \leq \alpha \leq N'$ are between 0 and $\frac{1}{2}$ (cf. (11.29)), the following estimates are valid:

$$\rho(M) \leq \rho_\nu := \max\{\rho_\nu(\xi) : 0 \leq \xi \leq 1/2\}, \quad (11.31a)$$

$$\|M\|_2 \leq \zeta_\nu := \max\{\zeta_\nu(\xi) : 0 \leq \xi \leq 1/2\}. \quad (11.31b)$$

The bounds ρ_ν and ζ_ν of the convergence rate and contraction numbers depend on the smoothing number ν ; however, they do not depend on the step size h . Since ρ_ν and ζ_ν decrease monotonically with increasing ν and $\rho_1 = \zeta_1 = \frac{1}{2} < 1$, the convergence of the two-grid method for the one-dimensional model problem (11.24) is proved. A more detailed discussion of the functions $\rho_\nu(\xi)$ and $\zeta_\nu(\xi)$ and their maxima in $[0, \frac{1}{2}]$ yields the following.

ν	ρ_ν	ζ_ν
1	1/2	1/2
2	1/4	1/4
3	1/8	0.150
4	0.0832	0.1159
5	0.0671	0.0947
10	0.0350	0.0496

Table 11.4 ρ_ν and ζ_ν .

Theorem 11.14. Let the two-grid method for solving the system (11.24) be characterised by Richardson's iteration with $\Theta = h^2/4$ (identical to the Jacobi iteration damped by $\frac{1}{2}$) as smoother, by the piecewise linear prolongation p , and the adjoint restriction (11.27f). Then the two-grid method with $\nu \geq 1$ smoothing steps converges with the rate ρ_ν in (11.31a), which is h -independent. The contraction number (with respect to the Euclidean norm) is bounded by ζ_ν in (11.31b). For increasing ν , these bounds have the asymptotic behaviour

$$\rho_\nu = \frac{1}{e\nu} + \mathcal{O}(\nu^{-2}), \quad \zeta_\nu = \frac{\sqrt{2}}{e\nu} + \mathcal{O}(\nu^{-2}).$$

Some values of ρ_ν , ζ_ν are listed in Table 11.4. Obviously, the quantities $\rho(M)$, $\|M\|_2$ converge with a decreasing step size parameter to their bounds ρ_ν and ζ_ν ; hence, the given estimates are strict. In §11.6 we will derive a convergence rate for general problems that also behaves like $\mathcal{O}(1/\nu)$. Theorem 11.14 demonstrates that such results about the asymptotic behaviour for large ν are not too pessimistic.

11.4 Multigrid Iteration

11.4.1 Algorithm

The two-grid method is not yet suited for practical applications because one still has to solve one system per iteration at level $\ell - 1$. The problem to be solved in (11.21c) has the form

$$A_{\ell-1} e_{\ell-1} = d_{\ell-1}; \quad (11.32)$$

hence it is of the same structure as the original problem $A_\ell x_\ell = b_\ell$. Instead of solving the system (11.32) exactly, one may approximate the solution iteratively. The iteration of choice is again the two-grid method, now applied to levels $\ell - 1$ and $\ell - 2$ instead of ℓ and $\ell - 1$. Then new auxiliary problems $A_{\ell-2} e_{\ell-2} = d_{\ell-2}$ arise, for which again the two-grid method (now at level $\ell - 2$) can be applied until equations $A_0 e_0 = d_0$ arise at the coarsest grid. The corresponding recursive method is the multigrid iteration $\Phi_\ell^{\text{MGM}(\nu_1, \nu_2)}$, which has the following algorithmic form:

procedure $\Phi_\ell^{\text{MGM}(\nu_1, \nu_2)}(x_\ell, b_\ell);$ if $\ell = 0$ then $\Phi_\ell^{\text{MGM}(\nu_1, \nu_2)} := A_0^{-1} b_0$ else begin for $i := 1$ to ν_1 do $x_\ell := \mathcal{S}_\ell(x_\ell, b_\ell);$ $d_{\ell-1} := r(A_\ell x_\ell - b_\ell);$ $e_{\ell-1}^{(0)} := 0;$ for $i := 1$ to γ do $e_{\ell-1}^{(i)} := \Phi_{\ell-1}^{\text{MGM}(\nu_1, \nu_2)}(e_{\ell-1}^{(i-1)}, d_{\ell-1});$ $x_\ell := x_\ell - p e_{\ell-1}^{(\gamma)};$ for $i := 1$ to ν_2 do $x_\ell := \hat{\mathcal{S}}_\ell(x_\ell, b_\ell);$ $\Phi_\ell^{\text{MGM}(\nu_1, \nu_2)} := x_\ell$ end;	(11.33) (11.33a) (11.33b) (11.33c) (11.33d₁) (11.33d₂) (11.33e) (11.33f) (11.33g)
--	---

The pre- and post-smoothing steps are the same as in (11.22b). Obviously, the recursive calls terminate after ℓ steps when level $\ell = 0$ is reached. Hence, the algorithm is well-defined.

ν_1 and ν_2 denote again the number of pre- and post-smoothing steps. A natural assumption is $\nu := \nu_1 + \nu_2 > 0$. For the iterative solution of the coarse-grid equation (11.32), γ steps of the iteration $\Phi_{\ell-1}^{\text{MGM}(\nu_1, \nu_2)}$ are applied to the starting value (11.33d₁). We shall see that $\gamma = 2$ is sufficient. Therefore, only the cases $\gamma = 1$ and $\gamma = 2$ are of practical interest. The multigrid iteration with $\gamma = 1$ has the name ‘V-cycle’, whereas the iteration with $\gamma = 2$ is called the ‘W-cycle’ (concerning the reason for these names, see Hackbusch [183, §2.5]).

The exact solution of linear equations is not completely avoided in the multigrid algorithm (11.33). In (11.33a), the system $A_0 x_0 = b_0$ corresponding to the coarsest grid has to be solved. Since the coarsest grid has the smallest number of grid points, the solution should not lead to practical difficulties. In the model case, according to (11.5b), $h_0 = \frac{1}{2}$ would be a possible choice of the coarsest grid size. In this case, $A_0 x_0 = b_0$ represents a single scalar equation.

Formally, the multigrid method is the product of the smoothing iteration and coarse-grid correction, where the latter almost corresponds to a composed method with a secondary iteration as described in §5.5. But different from §5.5, the auxiliary problem, which has to be approximated by the secondary iteration, does not belong to the same space X_ℓ but to the lower-dimensional space $X_{\ell-1}$.

11.4.2 Numerical Examples

The model problem with the step size $h = h_5 = 1/64$ is taken as an example. Table 11.5 shows the Euclidean norm $\|e^m\|_2$ of the errors and the reduction factors $\rho_{m+1,m} = \|e^m\|_2/\|e^{m-1}\|_2$. The parameters are $\nu_1 = 2$, $\nu_2 = 0$, $h_0 = \frac{1}{2}$. All matrices A_ℓ are defined by the five-point discretisation, p is the nine-point prolongation, and r the nine-point restriction. We choose the chequer-board Gauss–Seidel method as the smoothing iteration. The comparison of the results for $\gamma = 1$ (V-cycle) and $\gamma = 2$ (W-cycle) in Table 11.5 with the two-grid results (corresponding formally to $\gamma = \infty$; the values are copied from Table 11.2) show that $\gamma = 2$ yields almost the same fast convergence as the two-grid method, whereas the V-cycle results are less favourable.

m	$\gamma = 1$ (V-cycle)		$\gamma = 2$ (W-cycle)		$\gamma = \infty$ (two-grid algorithm) $\rho_{m+1,m}$
	$\ e^m\ _2$	$\rho_{m+1,m}$	$\ e^m\ _2$	$\rho_{m+1,m}$	
1	1.3274 ₁₀₋₁	0.1727	2.9984 ₁₀₋₀₂	0.03902	0.03807
2	2.2223 ₁₀₋₂	0.1674	1.3219 ₁₀₋₀₃	0.04408	0.04121
3	3.7656 ₁₀₋₃	0.1694	6.9050 ₁₀₋₀₅	0.05223	0.05132
4	6.4110 ₁₀₋₄	0.1702	3.7824 ₁₀₋₀₆	0.05477	0.05445
5	1.0941 ₁₀₋₄	0.1706	2.1584 ₁₀₋₀₇	0.05706	0.05741
6	1.8701 ₁₀₋₅	0.1709	1.2689 ₁₀₋₀₈	0.05879	0.05937
7	3.1996 ₁₀₋₆	0.1710	7.6788 ₁₀₋₁₀	0.06051	0.06124

Table 11.5 Multigrid iteration for the Poisson model problem with step size $h = 1/64$.

m	$\rho_{m+1,m}$	pointwise Gauss–Seidel		row Gauss–Seidel $\gamma = 1$
		$\gamma = 1$	$\gamma = 2$	
1	0.03025	1	0.1584	0.0465
2	0.04722	2	0.2602	0.0999
3	0.05308	3	0.3351	0.0952
4	0.05510	4	0.3479	0.1319
5	0.05694	5	0.3360	0.1267
6	0.05835	6	0.3142	0.1471
7	0.05970	7	0.2920	0.1304
8	0.06092	8	0.2720	0.1487
9	0.06206	9	0.2553	0.1328
10	0.06312			

Table 11.6 Multigrid convergence rates $\rho_{m+1,m}$ for Eq. (11.34) with $c = 4$ (left) and $c = 100$ (right), $h = 1/64$ and Gauss–Seidel smoothing.

In order to demonstrate that the multigrid iteration not only works well for positive definite problems, the next example is the nonsymmetric differential equation (convection-diffusion equation):

$$-\Delta u + cu_x = f$$

in $\Omega = (0, 1) \times (0, 1)$ with Dirichlet boundary values (1.1b), discretised by

$$A_\ell = h_\ell^{-2} \begin{bmatrix} -1 & & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix} + \frac{1}{2} ch_\ell^{-1} \begin{bmatrix} 0 & & \\ -1 & 0 & 1 \\ & 0 & \end{bmatrix}. \quad (11.34)$$

First, we choose $c = 4$ (for this value, all A_ℓ are M-matrices). $f = u = 0$ are taken as the right-hand side and exact solution, while $x(1 - x + y)$ serves as the starting value. The other parameters are the same as in Table 11.5. The W-cycle ($\gamma = 2$) shows a convergence rate of ≈ 0.06 (cf. Tab. 11.6) and hardly differs from the corresponding rate of the Poisson model case.

As soon as the coefficient c becomes substantially larger, e.g., $c=100$, a stability problem arises. Discretisation (11.34) yields an M-matrix for $h_5 = 1/64$, but not for larger h . A remedy is the matrix-dependent prolongation (11.13a,b) and the corresponding restriction, together with the Galerkin product (11.20) for $\ell < 5$ (cf. Hackbusch [183, §10.4]). Table 11.6 shows the convergence rates $\rho_{m+1,m}$ for $\gamma=1$ and $\gamma=2$. Different from the model case, the results for $\gamma = 2$ are hardly better than those for $\gamma = 1$. Furthermore, the rate ≈ 0.3 is not so favourable. The rate can be improved to ≈ 0.14 by row-wise Gauss–Seidel smoothing instead of the chequer-board Gauss–Seidel iteration (Table 11.6, right column).

In §§10.3.5–10.4 (cf. Tables 10.4, 10.6) the indefinite problem with the matrix

$$A_\ell := h_\ell^{-2} \begin{bmatrix} -1 & & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix} - \begin{bmatrix} 0 & & \\ 0 & 50 & 0 \\ & 0 & \end{bmatrix} \quad (11.35)$$

is solved. As we shall see in §11.6.2, for indefinite problems the choice of the coarsest step size is restricted. Here $h_0 = \frac{1}{2}$ is too coarse, but $h = \frac{1}{4}$ is possible. However, better results can be obtained with $h = \frac{1}{8}$ as the coarsest step size. Table 11.7 shows the results for $h_5 = \frac{1}{64}$, $h_0 = \frac{1}{8}$, $\gamma = 2$, nine-point prolongation, and nine-point restriction. $\nu_1 = 2$ chequer-board Gauss–Seidel steps are applied ($\nu_2 = 0$) as smoothing. The convergence rate (here 0.442) improves with decreasing grid size. Vice versa, the worse rate 0.613 results for $h=1/16$.

It is not necessary to choose h_0 sufficiently small if one uses the Kaczmarz iteration for smoothing (cf. §5.6.3), which is also convergent for the indefinite matrix (11.35). However, for the parameters $h_0 = \frac{1}{2}$, $\nu_1 = 2$, $\nu_2 = 0$, $\gamma = 2$, one obtains the rather unfavourable convergence rate 0.833 (for $h = \frac{1}{16}$ even 0.917).

m	$\ e^m\ _2$	$\rho_{m+1,m}$
1	$1.301_{10}-1$	0.169309
2	$5.607_{10}-2$	0.430985
3	$2.480_{10}-2$	0.442381
4	$1.097_{10}-2$	0.442503
5	$4.857_{10}-3$	0.442505

Table 11.7 Results for the indefinite problem (11.35).

11.4.3 Computational Work

To judge the convergence rates in §11.4.2, we have to take into account the amount of work per iteration (cf. §2.3). Because of the recursive structure, the amount of work is not quite obvious. The operations appearing in (11.33) are \mathcal{S}_ℓ in (11.33b,f), $r(A_\ell x_\ell - b_\ell)$ in (11.33c), and $x_\ell - p e_{\ell-1}$ in (11.33e). We denote the corresponding work by

$$C_S n_\ell \quad \text{operations for} \quad x_\ell \mapsto \mathcal{S}_\ell(x_\ell, b_\ell) \text{ or } \hat{\mathcal{S}}_\ell(x_\ell, b_\ell), \quad (11.36a)$$

$$C_D n_\ell \quad \text{operations for} \quad x_\ell \mapsto r(A_\ell x_\ell - b_\ell), \quad (11.36b)$$

$$C_C n_\ell \quad \text{operations for} \quad x_\ell \mapsto x_\ell - p e_{\ell-1}. \quad (11.36c)$$

Proportionality to dimension n_ℓ is a consequence of the sparsity of the matrix A_ℓ (cf. (2.28)). For standard approaches to fully populated matrices, n_ℓ would have to be replaced by n_ℓ^2 in (11.36a,b) (but see §D or Hackbusch [198, §10]).

The dimensions n_ℓ should increase with increasing level-number ℓ at least by a fixed factor C_h :

$$n_{\ell-1} \leq n_\ell / C_h \quad \text{for } \ell \geq 1. \quad (11.37)$$

Otherwise, the difficulty would arise that the auxiliary problems $A_{\ell-1} e_{\ell-1} = d_{\ell-1}$ are of a similar dimension as $A_\ell x_\ell = b_\ell$.

Remark 11.15. For the standard choice $h_\ell = h_{\ell-1}/2$ and the spatial dimension $d: \Omega \subset \mathbb{R}^d$, inequality (11.37) holds with $C_h = 2^d$. In the model case, $d=2$ is valid.

Theorem 11.16. Assume (11.36a–c) and (11.37). Let γ in (11.33d₂) satisfy

$$\gamma < C_h. \quad (11.38)$$

Then the work of the multigrid iteration is proportional to n_ℓ :

$$\begin{aligned} \text{Work}(\Phi_\ell^{\text{MGM}(\nu_1, \nu_2)}) &\leq C(\nu_1 + \nu_2) \cdot n_\ell \quad \text{with} \\ C(\nu) &= \frac{\nu C_S + C_D + C_C}{1 - \gamma/C_h} + \mathcal{O}((\gamma/C_h)^\ell). \end{aligned} \quad (11.39)$$

Proof. Let $C_\ell n_\ell$ be the work for one $\Phi_\ell^{\text{MGM}(\nu_1, \nu_2)}$ step. From the representation (11.33), we conclude that $C_\ell n_\ell \leq (\nu C_S + C_D + C_C)n_\ell + \gamma C_{\ell-1} n_{\ell-1}$. Inequality (11.37) yields $C_\ell \leq (\nu C_S + C_D + C_C) + \vartheta C_{\ell-1}$ with $\vartheta := \gamma/C_h$ and results in the geometrical sum

$$C_\ell \leq (\nu C_S + C_D + C_C)(1 + \vartheta + \dots + \vartheta^{\ell-1}) + \gamma C_0 / n_\ell,$$

where C_0 denotes the work for (11.33a) (independent of h_ℓ). Since $\gamma^\ell / n_\ell \leq \vartheta^\ell / n_1$, (11.39) follows. \square

Remark 11.17. In the two-dimensional case $d = 2$, (11.38) is satisfied for the interesting values $\gamma = 1, 2$ because of $C_h = 4$ (cf. Remark 11.15). The following constants are obtained for (11.39):

$$C_V(\nu) = \frac{4}{3}(\nu C_S + C_D + C_C) + \mathcal{O}\left((\gamma/C_h)^\ell\right) \quad \text{for } \gamma = 1, \quad (11.40a)$$

$$C_W(\nu) = 2(\nu C_S + C_D + C_C) + \mathcal{O}\left((\gamma/C_h)^\ell\right) \quad \text{for } \gamma = 2. \quad (11.40b)$$

Since $\gamma/C_h < 1$ (cf. (11.38)), formulae (11.39) and (11.40a,b) show that for increasing ℓ the work for solving $A_0x_0 = b_0$ in (11.33a) requires a vanishing portion of the total work.

Exercise 11.18. Although the one-dimensional case (11.24) is not of practical interest, one may apply the multigrid algorithm. Then (11.38) is not satisfied because $C_h = 2$ holds for the W-cycle ($\gamma = 2$). Prove that the work is equal to $\mathcal{O}(\ell n_\ell) = \mathcal{O}(n_\ell \log n_\ell)$.

For the standard multigrid parameters as used before, the work amounts to

$$C_S = 2(C_A - 1) \quad \text{for the Gauss-Seidel iteration, cf. (3.20b),} \quad (11.41a)$$

$$C_D = 2C_A + 11/4 \quad \text{for } r = \text{nine-point restriction (11.15),} \quad (11.41b)$$

$$C_C = 3/2 \quad \text{for } p = \text{nine-point prolongation (11.11).} \quad (11.41c)$$

The constants C_S and C_D improve for the Poisson model case ($C_A = 5$, since multiplications by coefficients 1 can be omitted):

$$C_S = 5 \quad \text{for the Gauss-Seidel iteration, cf. (3.21),} \quad (11.41a')$$

$$C_D = 5 + 10/4 \quad \text{for } r = \text{nine-point restriction (11.15),} \quad (11.41b')$$

If the chequer-board Gauss-Seidel method is used, some operations can be saved when applying r and p (cf. Hackbusch [183, Note 4.3.4]). Inserting formulae (11.41a–11.41b'), the numbers (11.40a,b) become

$$C_V(\nu) = \frac{8}{3}(\nu + 1)C_A + \frac{17 - 8\nu}{3} + \mathcal{O}(1/4^\ell) \quad \text{for } \gamma = 1, \quad (11.41d)$$

$$C_V(\nu) = 12 + \frac{20}{3}\nu + \mathcal{O}(1/4^\ell) \quad (\text{Poisson model case, } \gamma = 1), \quad (11.41d')$$

$$C_W(\nu) = 4(\nu + 1)C_A + \frac{17 - 8\nu}{3} + \mathcal{O}(1/2^\ell) \quad \text{for } \gamma = 2, \quad (11.41e)$$

$$C_W(\nu) = 18 + 10\nu + \mathcal{O}(1/2^\ell) \quad (\text{Poisson model case, } \gamma = 2). \quad (11.41e')$$

The corresponding effective work of the V- and W-cycle for the Poisson model problem with $\nu = 2$ is $C_{V[W]}(2)/|C_A \log(\rho)|$. Using the convergence rates ρ in Table 11.5, we obtain

$$\text{Eff}(\Phi_\ell^{\text{MGM}(2,0)}) = -C_V(2)/[5 \log(0.171)] \approx 2.89 \quad \text{for } \gamma = 1,$$

$$\text{Eff}(\Phi_\ell^{\text{MGM}(2,0)}) = -C_W(2)/[5 \log(0.06)] \approx 2.7 \quad \text{for } \gamma = 2.$$

Together with the convergence rate, the effective work is also h -independent. One should compare the numbers $\text{Eff}(\Phi_\ell^{\text{MGM}(2,0)})$ with the competing values in

Remark 8.49 (for $h = 1/32$). The numbers (11.41d',e') can also be interpreted as follows. One V-cycle step costs as much as ≈ 5 Gauss–Seidel iteration steps, one W-cycle step corresponds to 7.6 Gauss–Seidel steps.

Finally, we address the question how many smoothing steps should be performed. The numerical results in §11.4.2 have shown good agreement with the two-grid results. According to Table 11.3, these rates behave as $\approx C_\rho/(1+\nu)$, where $\nu = \nu_1 + \nu_2$. For simplification, assume that $C_C + C_D = C_S$. Then the multigrid work behaves like $\text{Work}(\Phi_\ell^{\text{MGM}(\nu_1, \nu_2)}) \approx (1+\nu)C$. With an increasing number ν of smoothing steps, the convergence improves, however, the work also increases. We have to minimise the effective work

$$-\frac{(1+\nu)C/C_A}{\log(C_\rho/(1+\nu))} = C' \frac{1+\nu}{\log(1+\nu) - \log(C_\rho)}.$$

The minimum is taken for $\nu^* = C_\rho e - 1$ and has the value $eC_\rho C/C_A$. This shows at least asymptotically that the faster the multigrid iteration (i.e., the smaller C_ρ), the smaller the number of smoothing steps should be. If, vice versa, it turns out that many smoothing steps are necessary, the multigrid method is not favourable and one should look for a better suited smoothing iteration.

11.4.4 Iteration Matrix

Since the iteration is defined recursively, the multigrid iteration matrix is also determined recursively.

Theorem 11.19. *Let S_ℓ , \hat{S}_ℓ be the iteration matrices of the respective consistent pre- and post-smoothing iterations \mathcal{S}_ℓ and $\hat{\mathcal{S}}_\ell$. Then the multigrid iteration $\Phi_\ell^{\text{MGM}(\nu_1, \nu_2)}$ is also consistent. Its iteration matrix $M_\ell^{\text{MGM}} = M_\ell^{\text{MGM}(\nu_1, \nu_2)}$ is defined by*

$$M_0^{\text{MGM}} = 0, \quad M_1^{\text{MGM}} = M_\ell^{\text{TGM}(\nu_1, \nu_2)}, \quad (11.42a)$$

$$M_\ell^{\text{MGM}} = M_\ell^{\text{TGM}(\nu_1, \nu_2)} + \hat{S}_\ell^{\nu_2} p (M_{\ell-1}^{\text{MGM}})^\gamma A_{\ell-1}^{-1} r A_\ell S_\ell^{\nu_1} \quad \text{for } \ell \geq 1. \quad (11.42b)$$

Proof. Obviously, the coarse-grid correction (11.33c–e) is consistent. Hence, Proposition 5.25a shows that Φ_0^{MGM} is consistent. For $\ell = 0$, Φ_0^{MGM} describes the exact solution, i.e., $M_0^{\text{MGM}} = 0$. For $\ell = 1$, the multi- and two-grid methods coincide. This proves (11.42a). The iteration matrix of the coarse-grid correction (11.33c–e) is

$$M_\ell^{\text{CGC}} = I - p \left[I - (M_{\ell-1}^{\text{MGM}})^\gamma \right] A_{\ell-1}^{-1} r A_\ell,$$

because we have $e_{\ell-1}^{(\gamma)} = I - p \left[I - (M_{\ell-1}^{\text{MGM}})^\gamma \right] A_{\ell-1}^{-1} r A_\ell e_\ell$ in (11.33e), as can be shown similarly as in the proof of (5.21b). (5.12a) and (11.23) prove (11.42b). \square

11.5 Nested Iteration

Contrary to the name ‘nested iteration’, the following scheme is not an iteration, but a finite technique which can be applied to any iterative method, provided that there is a hierarchy of problems

$$A_\ell x_\ell = b_\ell \quad (\ell = 0, 1, 2, \dots).$$

The latter requirement is the same as for constructing the multigrid method. Therefore, it is natural to combine the multigrid iteration with the nested iteration. This will be done in §11.5.4. First, we discuss the nested iteration independently of the multigrid method. Concerning the construction of discretisations we refer to Remark 11.4. The concept of the nested iteration is of particular help for non-linear problems, for which the choice of sufficiently close starting values is essential (cf. §11.9.5).

11.5.1 Discretisation Error and Relative Discretisation Error

We recall Remark 2.34. As long as x_ℓ is only considered as an approximation to a continuous solution of a differential equation, it makes no sense to compute x_ℓ more precisely than indicated by the discretisation error. The nested iteration provides a convenient way to obtain this goal.

For standard discretisations, the consistency order κ is known; i.e., the dependence on the step size is given by

$$\text{discretisation error: } \delta_\ell \leq C_{\text{de}} h_\ell^\kappa,$$

but the constant C_{de} is usually unknown (in principle, it may be described by the derivatives of the solution, but these are unknown). One may use error estimators, to bound δ_ℓ (cf. Verfürth [380]) and to stop the iteration as soon as the iteration error is below δ_ℓ . Since δ_ℓ is the difference² between the exact solution and x_ℓ , the triangle inequality yields an estimate $\mathcal{O}(h_\ell^\kappa + h_{\ell-1}^\kappa)$ of the difference of x_ℓ and $x_{\ell-1}$ by $\delta_\ell + \delta_{\ell+1}$. Provided that $h_\ell/h_{\ell-1}$ is uniformly bounded, the previous estimate yields the following bound of the *relative discretisation error*:

$$\|x_\ell - \tilde{p}x_{\ell-1}\| \leq C_1 h_\ell^\kappa \quad (\kappa > 0, x_\ell, x_{\ell-1} \text{ solutions to (11.6a)}). \quad (11.43)$$

Here, $\tilde{p} : X_{\ell-1} \rightarrow X_\ell$ is a suitable prolongation, which may not necessarily coincide with p in §11.1.3 and (11.33e). In the following, only the exponent κ in (11.43) must be known, not the constant C_1 .

² The precise notation of the difference might be $x - Px_\ell$ or $Rx - x_\ell$ (x and x_ℓ belong to different spaces; P and R are prolongations and restrictions between these spaces).

11.5.2 Algorithm

Obviously, the result x_ℓ^m of an iteration is more desirable, the better the starting iterate x_ℓ^0 is. So far, we did not study the choice of a starting iterate.³ Inequality (11.43) suggests using approximations to $x_{\ell-1}$ as the starting iterate of the iteration at level ℓ . The algorithm as proposed by Kronsjø-Dahlquist [248] reads as follows:

$\tilde{x}_0 :=$ suitable approximation of the solution of $A_0 x_0 = b_0$; for $\ell := 1$ to ℓ_{\max} do begin $\tilde{x}_\ell := \tilde{p} \tilde{x}_{\ell-1}$; for $i := 1$ to m_ℓ do $\tilde{x}_\ell := \Phi_\ell(\tilde{x}_\ell, b_\ell)$ end;	(11.44)
--	---------

Here, Φ_ℓ is any convergent and consistent linear iteration.⁴ The number m_ℓ of iterations is still to be determined. Theorem 11.20 will propose an appropriate choice.

Note that (11.44) is not an iteration in the proper sense, but a *finite* process. Furthermore, it produces approximate solutions \tilde{x}_ℓ for all levels $1 \leq \ell \leq \ell_{\max}$.

11.5.3 Error Analysis

First, we analyse the case of non-optimal linear iterations; i.e., the contraction number behave as

$$\|M_\ell^\Phi\| \leq 1 - c_\ell^\Phi h_\ell^\tau \quad \text{with } \tau > 0 \text{ for all } \ell \geq 1 \quad (11.45a)$$

(cf. (2.32c)). Here, M_ℓ^Φ is the iteration matrix of Φ_ℓ . An inequality opposite to condition (11.37), $n_{\ell-1} \leq n_\ell/C_h$, is

$$n_\ell \leq \bar{C}_h n_{\ell-1}.$$

By $n_\ell/n_{\ell-1} \approx (h_{\ell-1}/h_\ell)^\kappa$, the latter inequality also gives an estimate of $h_{\ell-1}/h_\ell$ appearing above. Together with the norm of \tilde{p} , we obtain an estimate of the form

$$\|\tilde{p}\|(h_{\ell-1}/h_\ell)^\kappa \leq C_2 \quad (\tilde{p} : X_{\ell-1} \rightarrow X_\ell) \quad (11.45b)$$

with κ as in (11.43). The inequalities (11.43) and (11.45a,b) must use the same family of norms in X_ℓ .

³ If one has to solve $A_\ell^{(\nu)} x_\ell^{(\nu)} = b_\ell^{(\nu)}$ ($\nu = 1, 2$) for similar data $(A_\ell^{(\nu)}, b_\ell^{(\nu)})$, one may take the solution of $A_\ell^{(1)} x_\ell^{(1)} = b_\ell^{(1)}$ as starting value for solving $A_\ell^{(2)} x_\ell^{(2)} = b_\ell^{(2)}$.

⁴ The m_ℓ -fold application of Φ_ℓ may be replaced by a semi-iteration, or acceleration methods may be used (cf. §10).

Theorem 11.20. Assume (11.43) and (11.45a,b). Fix some constant $K > 0$ and choose

$$m_\ell \geq \frac{1 + \log(C_2 + 1/K)}{c_\ell^\Phi} h_\ell^{-\tau}.$$

Then the nested iteration (11.44) with (11.49) produces results \tilde{x}_ℓ for all levels $0 \leq \ell \leq \ell_{\max}$ satisfying the error estimates

$$\|\tilde{x}_\ell - x_\ell\| \leq KC_1 h_\ell^\kappa \quad (11.46)$$

provided that the starting iterate \tilde{x}_0 satisfies inequality (11.46) for $\ell = 0$.

Proof. By assumption, (11.46) holds for $\ell = 0$. Assume (11.46) for all level numbers $\leq \ell - 1$. The starting iterate $x_\ell^0 := \tilde{p} \tilde{x}_{\ell-1}$ has an error that can be bounded by

$$\begin{aligned} \|x_\ell^0 - x_\ell\| &\leq \|\tilde{p} x_{\ell-1} - x_\ell\| + \|\tilde{p}(\tilde{x}_{\ell-1} - x_{\ell-1})\| \\ &\leq \|\tilde{p} x_{\ell-1} - x_\ell\| + \|\tilde{p}\| \|\tilde{x}_{\ell-1} - x_{\ell-1}\| \\ &\leq C_1 h_\ell^\kappa + \|\tilde{p}\| K C_1 h_{\ell-1}^\kappa \\ &\leq C_1 h_\ell^\kappa [1 + \|\tilde{p}\| (h_{\ell-1}/h_\ell)^\kappa K] \\ &\leq C_1 h_\ell^\kappa [1 + C_2 K]. \end{aligned}$$

m_ℓ iteration steps reduce the error to $\|x_\ell^{m_\ell} - x_\ell\| \leq (1 - c_\ell^\Phi h_\ell^\tau)^{m_\ell} \|x_\ell^0 - x_\ell\|$. The general inequality $1 + \xi \leq \exp(\xi)$ for all $\xi \in \mathbb{R}$ yields

$$\begin{aligned} (1 - c_\ell^\Phi h_\ell^\tau)^{m_\ell} &\leq \exp(1 - m_\ell c_\ell^\Phi h_\ell^\tau) \leq \exp\left(1 - \left(1 + \log(C_2 + \frac{1}{K})\right)\right) \\ &= 1/\left(C_2 + \frac{1}{K}\right) = \frac{K}{1 + C_2 K}. \end{aligned}$$

By the previous estimate of $\|x_\ell^0 - x_\ell\|$, (11.46) holds for ℓ . \square

Choose K somewhat smaller than one. Since $C_1 h_\ell^\kappa$ is the relative discretisation error, we obtain approximations \tilde{x}_ℓ with an iteration error similar in size:

$$\|\tilde{x}_\ell - x_\ell\| \leq K \times \text{relative discretisation error.} \quad (11.47)$$

Note that this statement holds, although the size of C_1 involved in the relative discretisation error $C_1 h_\ell^\kappa$ does not enter the algorithm.

The cost of the nested iteration is dominated by the work $\mathcal{O}(n_\ell h_\ell^{-\tau})$ at the maximal level $\ell = \ell_{\max}$. However, the standard approach using the starting value $x_{\ell_{\max}}^0 = 0$ requires $\mathcal{O}(n_{\ell_{\max}} h_{\ell_{\max}}^{-\tau} \kappa \log(1/h_{\ell_{\max}}))$ operations (cf. (2.31b)).

The analysis of the *cascade algorithm* in Bornemann–Deuflhard [56] demonstrates that the choice of the norm $\|\cdot\|$ is essential.

11.5.4 Application to Optimal Iterations

Now we assume that the iteration (as, e.g., the multigrid method; cf. Kronsjø [247]) has an h -independent contraction number:

$$\|M_\ell^\Phi\| \leq \zeta < 1 \quad \text{for all } \ell \geq 1, \quad M_\ell^\Phi: \text{iteration matrix of } \Phi_\ell. \quad (11.48)$$

Here the numbers m_ℓ in (11.44) can be chosen independently of ℓ :

$$m_\ell = m \quad (\ell \geq 1). \quad (11.49)$$

In Remark 11.22 we shall see that even the smallest possible number $m = 1$ is of practical interest.

Theorem 11.21. *Assume (11.43), (11.48), and (11.45b). The iteration number $m_\ell = m$ (cf. (11.49)) should be sufficiently large so that*

$$C_2 \zeta^m < 1. \quad (11.50)$$

Then the nested iteration (11.44) with (11.49) produces results \tilde{x}_ℓ for all levels $0 \leq \ell \leq \ell_{\max}$ satisfying the error estimates

$$\|\tilde{x}_\ell - x_\ell\| \leq C_3(\zeta^m) C_1 h_\ell^\kappa \quad \text{with} \quad C_3(\zeta^m) := \zeta^m / (1 - C_2 \zeta^m), \quad (11.51)$$

provided that the starting iterate \tilde{x}^0 satisfies inequality (11.51) for $\ell = 0$.

Proof. We repeat the induction proof of Theorem 11.20. Assume (11.51) for levels $\leq \ell - 1$. The starting iterate $x_\ell^0 := \tilde{p} \tilde{x}_{\ell-1}$ has an error that can be bounded by

$$\begin{aligned} \|x_\ell^0 - x_\ell\| &\leq \|\tilde{p} x_{\ell-1} - x_\ell\| + \|\tilde{p}(\tilde{x}_{\ell-1} - x_{\ell-1})\| \\ &\leq \|\tilde{p} x_{\ell-1} - x_\ell\| + \|\tilde{p}\| \|\tilde{x}_{\ell-1} - x_{\ell-1}\| \\ &\leq C_1 h_\ell^\kappa + \|\tilde{p}\| C_3(\zeta^m) C_1 h_{\ell-1}^\kappa \\ &\leq C_1 h_\ell^\kappa [1 + \|\tilde{p}\| (h_{\ell-1}/h_\ell)^\kappa C_3(\zeta^m)] \\ &\leq C_1 h_\ell^\kappa [1 + C_2 C_3(\zeta^m)]. \end{aligned}$$

After m iteration steps, the error is reduced to $\|x_\ell^m - x_\ell\| \leq \zeta^m \|x_\ell^0 - x_\ell\| \leq C_1 h_\ell^\kappa \{\zeta^m [1 + C_2 C_3(\zeta^m)]\}$ because of (11.48). Definition of $C_3(\cdot)$ in (11.51) shows that $\{\dots\} = C_3(\zeta^m)$ and proves (11.51) for ℓ . \square

Again the iteration error $\|\tilde{x}_\ell - x_\ell\|$ coincides up to a factor $C_3(\zeta^m)$ with the relative discretisation error $C_1 h_\ell^\kappa$, i.e., (11.47) holds with $K := C_3(\zeta^m)$.

Remark 11.22. The standard choice $h_\ell = h_{\ell-1}/2$ and the inequality $\|\tilde{p}\| \leq 1$, which is valid for standard interpolations, yield the constant $C_2 = 2^\kappa$ in (11.45b). The consistency order of the model case is $\kappa = 2$, from which $C_2 = 4$. Therefore, the factor $C_3(\zeta^m)$ is equal to

$$C_3(\zeta^m) = \zeta^m / (1 - 4\zeta^m).$$

For multigrid methods with convergence rates $\leq \zeta = 0.2$ (see the results in §11.4.2), condition (11.50) is satisfied for only one iteration step (i.e., $m = 1$) and produces the value $C_3(0.2) = 1$.

11.5.5 Amount of Computational Work

Let $C_{n\ell}$ be the work required by one step of the iteration Φ_ℓ at level ℓ and assume (11.37): $n_{\ell-1} \leq n_\ell/C_h$. The work for $\tilde{x}_{\ell-1} \mapsto \tilde{p}\tilde{x}_{\ell-1}$ is considered negligible. The total work amounting to $C_{\text{nested}}n_\ell \leq mC(n_1 + n_2 + \dots + n_\ell)$ can be estimated, using the geometrical sum $n_1 + \dots + n_\ell \leq n_\ell \sum_k C_h^{-k} \leq C_h n_\ell / (C_h - 1)$, by

$$C_{\text{nested}} \leq m C C_h / (C_h - 1).$$

For the standard case $C_h = 2^d = 4$ (cf. Remark 11.15), we obtain the result

$$\text{Work}_{(11.44)} \leq \frac{4m}{3} \text{Work}(\Phi_{\ell_{\max}}). \quad (11.52)$$

If we try to achieve an accuracy of $\varepsilon = Ch^\kappa$ at level $\ell = \ell_{\max}$ with the starting iterate $\tilde{x}_\ell := 0$ by iterating with Φ_ℓ , the work would be proportional to $\mathcal{O}(|\log \varepsilon|) = \mathcal{O}(|\log h_\ell|)$ (cf. (2.31b)). According to Remark 11.22, $m = 1$ is a realistic choice. Inequality (11.52) shows that sufficient accuracy for all levels $0 \leq \ell \leq \ell_{\max}$ can be attained with the 4/3-fold work of a single $\Phi_{\ell_{\max}}$ step.

Together with the numbers in (11.41d', e') and Table 11.5 (with $\nu_1 = 2$, $\nu_2 = 0$, $m = 1$), we obtain the following results:

$$\begin{aligned} \text{the V-cycle } (\gamma = 1) &\text{ requires } 34n_{\ell_{\max}} \text{ operations to produce} \\ \|\tilde{x}_\ell - x_\ell\| &\leq 0.53 C_1 h_\ell^\kappa \quad \text{for } 0 \leq \ell \leq \ell_{\max}, \end{aligned} \quad (11.53a)$$

$$\begin{aligned} \text{the W-cycle } (\gamma = 2) &\text{ requires } 51n_{\ell_{\max}} \text{ operations to produce} \\ \|\tilde{x}_\ell - x_\ell\| &\leq 0.08 C_1 h_\ell^\kappa \quad \text{for } 0 \leq \ell \leq \ell_{\max}. \end{aligned} \quad (11.53b)$$

The work given in (11.53b) corresponds to about 10 steps of the Gauss–Seidel iteration at level ℓ_{\max} .

Since the nested iteration (11.44) is a finite process and not an iteration, considerations in §2.3 are not applicable. How many operations are necessary, depends on the desired accuracy.

11.5.6 Numerical Examples

First, the nested iteration is applied to the differential equation

$$-\Delta u = f := -\Delta(e^{x+y^2}) \quad (11.54a)$$

with boundary values $\varphi = e^{x+y^2}$. The negative Laplacian $-\Delta$ is discretised at all levels by the standard five-point star. \tilde{p} is cubic interpolation. Let x_ℓ^* be the restriction of the exact solution e^{x+y^2} of (11.54a) to the grid Ω_ℓ . Note that x_ℓ^* does not coincide with the discrete solution x_ℓ of the system $A_\ell x_\ell = b_\ell$

corresponding to (11.54a). In Table 11.8, the results \tilde{x}_ℓ of the nested iteration are compared with x_ℓ^* because this is the error most interesting in practice. The maximum norm $\|\tilde{x}_\ell - x_\ell^*\|_\infty$ of these errors is given for cases $m = 1$ and $m = 2$ (m in (11.49)). For comparison, the last column shows the discretisation error $\|x_\ell - x_\ell^*\|_\infty$, which formally corresponds to $m = \infty$. The multigrid iteration used for solving (11.44) has the same parameters as the W-cycle ($\gamma = 2$) in Table 11.5. The data in Table 11.8 demonstrate that the choice $m = 1$ is sufficient. $m = 2$ doubles the work but cannot improve the total error $\|\tilde{x}_\ell - x_\ell^*\|_\infty$ substantially.

Analogous data are given in Table 11.9 for the differential equation

$$-\Delta u = f := -\Delta(y \sin(10x)) \quad (11.54b)$$

with a solution $y \sin(10x)$ which is oscillatory in the x direction. By the nonsmooth behaviour of the solution, the discretisation error (last column) for problem (11.54b) is nearly one digit worse than for (11.54a). Therefore, the additional error $\mathcal{O}(h_\ell^2)$ of linear interpolation \tilde{p} , which is used instead of the cubic one, is of minor consequence. Also for this example, it does not pay to perform $m = 2$ iterations per level.

ℓ	h_ℓ	$m = 1$	$m = 2$	$m = \infty$
0	1/2	7.9944658 ₁₀₋₂	7.9944658 ₁₀₋₂	7.9944658 ₁₀₋₂
1	1/4	3.9908756 ₁₀₋₂	2.9215605 ₁₀₋₂	2.8969488 ₁₀₋₂
2	1/8	1.5788721 ₁₀₋₂	8.1023136 ₁₀₋₃	8.0307789 ₁₀₋₃
3	1/16	3.2919346 ₁₀₋₃	2.0768391 ₁₀₋₃	2.0729855 ₁₀₋₃
4	1/32	5.75915491 ₁₀₋₄	5.2253758 ₁₀₋₄	5.2247399 ₁₀₋₄
5	1/64	1.3291689 ₁₀₋₄	1.3093946 ₁₀₋₄	1.3093956 ₁₀₋₄

Table 11.8 Errors $\|\tilde{x}_\ell - x_\ell^*\|_\infty$ of the nested iteration for (11.54a).

ℓ	h_ℓ	$m = 1$	$m = 2$	$m = \infty$
0	1/2	2.8249099 ₁₀₋₀	2.8249099 ₁₀₋₀	2.8249099 ₁₀₋₀
1	1/4	5.0876212 ₁₀₋₁	4.6124302 ₁₀₋₁	4.7880033 ₁₀₋₁
2	1/8	9.5881341 ₁₀₋₂	1.0330948 ₁₀₋₁	1.0308770 ₁₀₋₁
3	1/16	2.7648979 ₁₀₋₂	2.6636710 ₁₀₋₂	2.6689213 ₁₀₋₂
4	1/32	6.8798570 ₁₀₋₃	6.6486368 ₁₀₋₃	6.6506993 ₁₀₋₃
5	1/64	1.6998365 ₁₀₋₃	1.6716069 ₁₀₋₃	1.6714014 ₁₀₋₃

Table 11.9 Errors $\|\tilde{x}_\ell - x_\ell^*\|_\infty$ of the nested iteration for (11.54b).

11.5.7 Comments

Additional variants for the nested iteration (e.g., combinations with extrapolation techniques) are discussed in Hackbusch [183, §5.4, §9.3.4, §16.4] and [191, §5.6.5].

Although nonlinear systems are not the subject of this book, we remark that the nested iteration is of even greater importance for nonlinear systems of equations. In the linear case, it helps to save computer time. For nonlinear iterations, however, the availability of sufficiently good starting iterates often decides on convergence (to the desired solution) or divergence. The nested iteration with its starting value $\tilde{x}_\ell := \tilde{p} \tilde{x}_{\ell-1}$ is a suitable technique for generating such starting iterates.

A description and analysis of the nonlinear multigrid method and the corresponding nested iteration can be found in §11.9.5.

11.6 Convergence Analysis

11.6.1 Summary

The convergence proof of multigrid methods differs from the convergence proofs of other iterations because here the relationship between the equations $A_\ell x_\ell = b_\ell$ and $A_{\ell-1}x_{\ell-1} = b_{\ell-1}$ plays an important role.

As sufficient criteria, we introduce and discuss two conditions in §§11.6.2–11.6.3: the smoothing and approximation property. The smoothing property is of algebraic nature, whereas the proof of the approximation property involves the continuous problem, whose discretisation is described by $A_\ell x_\ell = b_\ell$. Together, the smoothing and approximation properties yield the convergence statement for the two-grid iteration (§11.6.4). For $\gamma \geq 2$, multigrid convergence can be concluded directly from the two-grid convergence (§11.6.5).

For positive definite matrices A_ℓ , the multigrid method can be designed as a positive definite iteration. In this case, we shall achieve even better convergence results, including the V-cycle ($\gamma = 1$; see §11.7). These results are generalised in Theorem 11.61 to the nonsymmetric case.

The analysis represented below is strongly simplified compared with that of Hackbusch [194], since presently we base our considerations mostly on the Euclidean and spectral norm. Other norms are mentioned in §11.6.6 and §11.7.2.

In contrast to what has been said above, there are multigrid methods for which convergence proofs can be performed by purely algebraic considerations. These variants will be discussed in §12.9.

11.6.2 Smoothing Property

In §11.1.1 we called a grid function $x_\ell = \sum \xi_{\alpha\beta} e^{\alpha\beta}$ (cf. (11.2b)) *smooth* if the coefficients $\xi_{\alpha\beta}$ of high frequencies α, β (corresponding to the large eigenvalues $\lambda_{\alpha\beta}$ in (3.1a)) are small. Quantitatively, one may measure the smoothness by $\|A_\ell x_\ell\|_2 = (\sum |\lambda_{\alpha\beta} \xi_{\alpha\beta}|^2)^{1/2}$. If the smoothing step (11.21a) really leads to a smoothing of the errors $e_\ell = x_\ell^0 - x_\ell$, the error $S_\ell^\nu e_\ell$ produced by the smoothing step must have a better smoothing measure $\|A_\ell S_\ell^\nu e_\ell\|_2$ than e_ℓ . Therefore, the smoothing ability is characterised by the spectral norm $\|A_\ell S_\ell^\nu\|_2$. Before defining the smoothing property, we analyse $\|A_\ell S_\ell^\nu\|_2$ for Richardson's iteration with positive definite A_ℓ :

$$S_\ell(x_\ell, b_\ell) := x_\ell - \Theta(A_\ell x_\ell - b_\ell) \quad (11.55a)$$

$$\text{with } \Theta = \Theta_\ell = 1/\rho(A_\ell) = 1/\|A_\ell\|_2. \quad (11.55b)$$

We have $\|A_\ell S_\ell^\nu\|_2 = \|A_\ell(I - \Theta A_\ell)^\nu\|_2 = \|X(I - X)^\nu\|_2/\Theta$ with $X := \Theta A_\ell$. The following lemma applies to the matrix polynomial $X(I - X)^\nu$.

Lemma 11.23. (a) For all matrices X with $0 \leq X \leq I$, the inequality

$$\|X(I - X)^\nu\|_2 \leq \eta_0(\nu) \quad (\nu \geq 0)$$

holds, where the function $\eta_0(\nu)$ is defined by

$$\eta_0(\nu) := \nu^\nu / (\nu + 1)^{\nu+1}. \quad (11.56)$$

(b) The asymptotic behaviour of $\eta_0(\nu)$ for $\nu \rightarrow \infty$ is

$$\eta_0(\nu) = \frac{1}{e\nu} + \mathcal{O}(\nu^{-2}).$$

Proof. Set $f(\xi) := \xi(1 - \xi)^\nu$. According to Lemma A.11a, we have

$$\|X(I - X)^\nu\|_2 = \rho(X(I - X)^\nu) = \max\{|f(\xi)| : \xi \in \sigma(X)\}.$$

By $f(\xi) \leq f(1/(\nu + 1)) = \eta_0(\nu)$ for all $\xi \in [0, 1] \supset \sigma(X)$, part (a) is proved. The discussion of the function $\eta_0(\nu)$ yields statement (b). \square

Remark 11.24. For $A_\ell > 0$, Richardson's method (11.55a,b) leads to

$$\|A_\ell S_\ell^\nu\|_2 \leq \eta_0(\nu) \|A_\ell\|_2 \quad \text{for all } \nu \geq 0, \ell \geq 0. \quad (11.57)$$

Note that the factor $\eta_0(\nu)$ is independent of h_ℓ and ℓ . The smoothing property, which we are going to define, is an estimate with a form similar to (11.57). Instead of $\eta_0(\nu)$, we may take an arbitrary zero sequence $\eta(\nu) \rightarrow 0$. Furthermore, it is neither necessary nor desirable to require an inequality as (11.57) for all $\nu \geq 0$.

Definition 11.25 (smoothing property). An iteration S_ℓ ($\ell \geq 0$) with iteration matrix S_ℓ satisfies the smoothing property if there are functions $\eta(\nu)$ and $\bar{\nu}(h)$ independent of ℓ with

$$\|A_\ell S_\ell^\nu\|_2 \leq \eta(\nu) \|A_\ell\|_2 \quad \text{for all } 0 \leq \nu < \bar{\nu}(h_\ell), \ell \geq 1, \quad (11.58a)$$

$$\lim_{\nu \rightarrow \infty} \eta(\nu) = 0, \quad (11.58b)$$

$$\lim_{h \rightarrow 0} \bar{\nu}(h) = \infty \quad \text{or} \quad \bar{\nu}(h) = \infty. \quad (11.58c)$$

The equality $\bar{\nu}(h) = \infty$ in (11.58c) expresses the fact that (11.58a) holds for all ν . This happens only for convergent iterations S_ℓ , as shown below.

Remark 11.26. The conditions (11.58a,b) together with $\bar{\nu}(h) = \infty$ imply convergence of S_ℓ .

Proof. $\rho(S_\ell^\nu) \leq \|S_\ell^\nu\|_2 \leq \|A_\ell^{-1}\|_2 \|A_\ell S_\ell^\nu\|_2 \leq \eta(\nu) \operatorname{cond}_2(A_\ell) < 1$ for sufficiently large ν follows from $\eta(\nu) \rightarrow 0$ and implies $\rho(S_\ell) < 1$. \square

From Remark 11.24, we conclude the next theorem.

Theorem 11.27. For $A_\ell > 0$, the Richardson iteration (11.55a,b) satisfies the smoothing property (11.58a-c) with $\eta(\nu) := \eta_0(\nu)$ and $\bar{\nu}(h) = \infty$.

The reason for the more general condition (11.58c) instead of $\bar{\nu}(h) = \infty$ is that the smoothing property can also be formulated for non-convergent iterations. Examples of divergent iterations are the Gauss–Seidel iteration for the indefinite problem (11.35), as well as Richardson’s iteration in the next remark.

Remark 11.28. Assume that the indefinite matrix $A_\ell = A_\ell^H$ has the spectrum $\sigma(A_\ell) \subset [-\alpha_\ell, \beta_\ell]$ with $0 \leq \alpha_\ell \leq \beta_\ell$ and $\lim_{\ell \rightarrow \infty} \alpha_\ell/\beta_\ell = 0$. Although the Richardson iteration with $\Theta = 1/\beta_\ell$ is divergent, it satisfies the smoothing property.

Proof. The damping factor is $\Theta = 1/\beta_\ell$. As in the proof for Lemma 11.23, we have $\|A_\ell(I - \Theta A_\ell)^\nu\|_2 \leq \max\{\eta_0(\nu), (\alpha_\ell/\beta_\ell)(1 + \alpha_\ell/\beta_\ell)^\nu\} \|A_\ell\|_2$. Define $\bar{\nu}(h_\ell)$ by $\bar{\nu}(h_\ell) := \beta_\ell/\alpha_\ell \rightarrow \infty$. For $\nu < \bar{\nu} := \bar{\nu}(h_\ell)$, the inequalities

$$(\alpha_\ell/\beta_\ell)(1 + \alpha_\ell/\beta_\ell)^\nu \leq (\alpha_\ell/\beta_\ell) \exp\{\nu \alpha_\ell/\beta_\ell\} = \frac{1}{\nu} \left(\frac{\nu}{\bar{\nu}} \exp \frac{\nu}{\bar{\nu}} \right) \leq \frac{e}{\nu}$$

follow. Hence, (11.58a–c) is satisfied by $\eta(\nu) := \max\{\eta_0(\nu), e/\nu\} = e/\nu$. \square

The assumptions of Remark 11.28 are fulfilled for discretisation of the Helmholtz equation $-\Delta u - cu = f$ ($c > 0$), because $\mathcal{O}(\alpha_\ell/\beta_\ell) = \mathcal{O}(h_\ell^2)$.

The following theorem can be considered as a perturbation lemma. It shows that the smoothing property remains valid under the perturbation of the matrix A'_ℓ into $A_\ell = A'_\ell + A''_\ell$, where A_ℓ may be indefinite and nonsymmetric.

Theorem 11.29. Let $A_\ell = A'_\ell + A''_\ell$ and $\mathcal{S}_\ell = \mathcal{S}_\ell(\cdot, \cdot, A_\ell)$ and $\mathcal{S}'_\ell = \mathcal{S}'_\ell(\cdot, \cdot, A'_\ell)$ be the smoothing iterations corresponding to A_ℓ and A'_ℓ , respectively. Their iteration matrices are denoted by S_ℓ and S'_ℓ with $S''_\ell := S_\ell - S'_\ell$. Assume that

A'_ℓ and S'_ℓ satisfy the smoothing property with $\eta'(\nu), \bar{\nu}'(h)$, (11.59a)

$$\|S'_\ell\|_2 \leq C'_S \quad \text{for all } \ell \geq 1, \quad \text{span style="float: right;">(11.59b)}$$

$$\lim_{\ell \rightarrow \infty} \|S''_\ell\|_2 = 0, \quad \text{span style="float: right;">(11.59c)}$$

$$\lim_{\ell \rightarrow \infty} \|A''_\ell\|_2 / \|A'_\ell\|_2 = 0. \quad \text{span style="float: right;">(11.59d)}$$

Then the iteration $\mathcal{S}_\ell = \mathcal{S}_\ell(\cdot, \cdot, A_\ell)$ for A_ℓ also satisfies the smoothing property. The corresponding bound $\eta(\nu)$ can be chosen, e.g., as $\eta(\nu) := 2\eta'(\nu)$.

Proof. $C_S := C'_S + \max\{\|S''_\ell\|_2 : \ell \geq 1\}$ satisfies $\|S_\ell\|_2 \leq C_S$ for all $\ell \geq 1$. Without loss of generality, we may suppose that $C_S \geq 1$. S_ℓ^ν can be split into $S_\ell^{\nu\nu} + S_\ell^{\nu(\nu)}$ with

$$\begin{aligned} \|S_\ell^{\nu(\nu)}\|_2 &= \|S_\ell^\nu - S_\ell^{\nu\nu}\|_2 \\ &= \left\| \sum_{\mu=0}^{\nu-1} S_\ell^\mu (S_\ell - S'_\ell) S'^{\nu-1-\mu} \right\|_2 = \left\| \sum_{\mu=0}^{\nu-1} S_\ell^\mu S'_\ell S''_\ell S'^{\nu-1-\mu} \right\|_2 \\ &\leq \left(\sum_{\mu=0}^{\nu-1} C_S^\mu C'_S^{\nu-1-\mu} \right) \|S''_\ell\|_2 \leq \nu C_S^{\nu-1} \|S''_\ell\|_2 \xrightarrow{(11.59c)} 0 \end{aligned} \quad (11.59e)$$

for $\ell \rightarrow \infty$. For $1 \leq \nu \leq \bar{\nu}'(h_\ell)$, we have

$$\begin{aligned} \|A_\ell S_\ell^\nu\|_2 &\leq \|A'_\ell S_\ell^{\nu\nu}\|_2 + \|A''_\ell\|_2 \|S_\ell^\nu\|_2 + \|A'_\ell\|_2 \|S_\ell^{\nu(\nu)}\|_2 \\ &\leq \eta'(\nu) \|A'_\ell\|_2 + C_S^\nu \|A''_\ell\|_2 + \nu C_S^{\nu-1} \|S_\ell''\|_2 \|A'_\ell\|_2 \\ &= \eta'(\nu) \|A_\ell\|_2 \left\{ \frac{\|A'_\ell\|_2}{\|A_\ell\|_2} + C_S^\nu \frac{\|A''_\ell\|_2}{\|A_\ell\|_2} + \nu C_S^{\nu-1} \frac{\|A'_\ell\|_2}{\|A_\ell\|_2} \|S_\ell''\|_2 \right\}. \end{aligned} \quad (11.59f)$$

By $\|A''_\ell\|_2/\|A'_\ell\|_2 \rightarrow 0$, $\|S_\ell''\|_2 \rightarrow 0$, $\|A'_\ell\|_2/\|A_\ell\|_2 \rightarrow 1$, $\|A''_\ell\|_2/\|A_\ell\|_2 \rightarrow 0$, the expression $\{\dots\}$ converges to 1 for $\ell \rightarrow \infty$ (i.e., for $h = h_\ell \rightarrow 0$) while ν is fixed. This proves that $\bar{\nu}''(h) \rightarrow \infty$ ($h \rightarrow 0$) for

$$\bar{\nu}''(h) := \sup \left\{ \nu > 0 : \frac{\|A'_\ell\|_2}{\|A_\ell\|_2} (1 + \nu C_S^{\nu-1} \|S_\ell''\|_2) + C_S^\nu \frac{\|A''_\ell\|_2}{\|A_\ell\|_2} \leq 2 \text{ for } h_\ell \leq h \right\}.$$

We define $\eta(\nu) := 2\eta'(\nu)$ and $\bar{\nu}(\eta) := \min\{\bar{\nu}'(h), \bar{\nu}''(h)\}$. For $\nu \leq \bar{\nu}(h)$, inequality (11.59f) proves the smoothing property $\|A_\ell S_\ell^\nu\|_2 \leq \eta(\nu) \|A_\ell\|_2$. \square

Usually, discretisations of elliptic differential equations satisfy the following conditions:

There is an h -independent constant c_0 such that

$$A'_\ell := \frac{1}{2}(A_\ell + A_\ell^H) + c_0 I \text{ is positive definite,} \quad (11.60a)$$

$$Ch_\ell^{-2m} \leq \|A'_\ell\|_2 \leq \bar{C}h_\ell^{-2m} \quad (2m: \text{order of the differential eq.}), \quad (11.60b)$$

$$\|A''_\ell\|_2 \leq Ch_\ell^{1-2m} \quad \text{for } A''_\ell := A_\ell - A'_\ell = \frac{1}{2}(A_\ell - A_\ell^H) - c_0 I \quad (11.60c)$$

(cf. Hackbusch [183, 201]). To apply Theorem 11.29, one proves the smoothing property for the positive definite matrix A'_ℓ and transfers this property to A_ℓ by Theorem 11.29. Condition (11.59d) follows from (11.60b,c) by $\|A''_\ell\|_2/\|A'_\ell\|_2 \leq \mathcal{O}(h_\ell) \rightarrow 0$. Since $S_\ell'' = -\Theta A''_\ell = -A''_\ell/\|A'_\ell\|_2$ in the case of the Richardson iteration, (11.59d) also implies (11.59c). (11.59b) is always satisfied by $C_S = 2$, because $S'_\ell = I - A'_\ell/\|A'_\ell\|_2$ (even $C_S = 1$ if $A'_\ell \geq 0$).

The smoothing property can be proved not only for the Richardson method but also for the damped (block-)Jacobi iteration, the 2-cyclic Gauss–Seidel iteration (in particular, the chequer-board Gauss–Seidel method for five-point formulae), and the Kaczmarz iteration. Furthermore, symmetric iterations like the symmetric Gauss–Seidel method, SSOR, and the ILU iteration (cf. deZeeuw[105]) belong to this class. The symmetric case will be considered in §11.7.3. The smoothing property does not hold, e.g., for the undamped Jacobi method or the SOR method with $\omega \geq \omega_{\text{opt}}$. For the smoothing analysis of the iterations mentioned above, see Hackbusch [183, §6.2].

The proof of Lemma 11.23 is based on the properties of the spectral norm for normal matrices. Correspondingly, statements for general matrices are proved via perturbation arguments. Nevertheless, it is possible to obtain the smoothing property

for general matrices directly. Even other norms than the spectral norm are possible. The following result by Reusken appeared in a report of 1991 and later in [323, 322].

Theorem 11.30. *Let $\|\cdot\|$ be a matrix norm corresponding to a vector norm. Let $S_\ell = I - W_\ell^{-1}A_\ell$ be the iteration matrix of the smoothing iteration and assume that*

$$\|I - 2W_\ell^{-1}A_\ell\| \leq 1, \quad (11.61a)$$

$$\|W_\ell\| \leq C \|A_\ell\| \quad (11.61b)$$

with a constant C independent of ℓ . Then the smoothing property (11.61c) holds:

$$\|A_\ell S_\ell^\nu\| \leq C \sqrt{2/(\pi\nu)} \|A_\ell\| \quad \text{for all } \nu \geq 1. \quad (11.61c)$$

The matrix in (11.61a) is the iteration matrix of $\mathcal{S}_{\vartheta=2,\ell}$, the extrapolated version of \mathcal{S}_ℓ with $\vartheta = 2$. Inequality (11.61a) does not imply convergence of $\mathcal{S}_{\vartheta=2,\ell}$, but characterises the weak contractivity or nonexpansivity (cf. Definition 7.3). The proof of the theorem is based on the following lemma.

Lemma 11.31. *Let the matrix B satisfy $\|B\| \leq 1$ with respect to a matrix norm corresponding to a vector norm. Then⁵*

$$\|(I - B)(I + B)^\nu\| \leq 2 \binom{\nu}{\lfloor \nu/2 \rfloor} \leq 2^{\nu+1} \sqrt{2/(\pi\nu)}.$$

Proof. Note that

$$\begin{aligned} (I - B)(I + B)^\nu &= (I - B) \sum_{\mu=0}^{\nu} \binom{\nu}{\mu} B^\mu = I + \sum_{\mu=1}^{\nu} \binom{\nu}{\mu} B^\mu - \sum_{\mu=0}^{\nu-1} \binom{\nu}{\mu} B^{\mu+1} - B^{\nu+1} \\ &= (I - B^{\nu+1}) + \sum_{\mu=1}^{\nu} \left[\binom{\nu}{\mu} - \binom{\nu}{\mu-1} \right] B^\mu. \end{aligned}$$

By $\|B^\mu\| \leq 1$, $\binom{\nu}{\mu-\alpha} = \binom{\nu}{\alpha}$ and $\binom{\nu}{\mu} \geq \binom{\nu}{\mu-1}$ for $\mu \leq \lfloor \nu/2 \rfloor$ we obtain

$$\begin{aligned} \|(I - B)(I + B)^\nu\| &\leq 2 + 2 \sum_{\mu=1}^{\lfloor \nu/2 \rfloor} \left| \binom{\nu}{\mu} - \binom{\nu}{\mu-1} \right| \\ &= 2 + 2 \sum_{\mu=1}^{\lfloor \nu/2 \rfloor} \left\{ \binom{\nu}{\mu} - \binom{\nu}{\mu-1} \right\} = 2 + 2 \binom{\nu}{\lfloor \nu/2 \rfloor} - 2 \binom{\nu}{0} = 2 \binom{\nu}{\lfloor \nu/2 \rfloor}. \end{aligned}$$

The sequence $a_k := \binom{2k}{k} \sqrt{k} / 2^{2k}$ is monotonically increasing and tends to $\lim a_k = \frac{1}{\sqrt{\pi}}$. The identity $\binom{\nu}{\lfloor \nu/2 \rfloor} = a_{\nu/2} 2^\nu / \sqrt{\nu/2}$ for even powers ν leads to the desired estimate $a_k \leq 1/\sqrt{\pi}$. For odd ν use $\binom{\nu}{\lfloor \nu/2 \rfloor} = \frac{1}{2} \binom{\nu+1}{(\nu+1)/2}$. \square

Proof of Theorem 11.30. $(I - B)(I + B)^\nu = 2^{\nu+1} W_\ell^{-1} A_\ell S_\ell^\nu$ holds with $B := I - 2W_\ell^{-1}A_\ell$; hence,

$$\|A_\ell S_\ell^\nu\| = 2^{-\nu-1} \|W_\ell(I - B)(I + B)^\nu\| \leq 2^{-\nu-1} \|W_\ell\| \|(I - B)(I + B)^\nu\|.$$

Assumption (11.61b) and Lemma 11.31 yield the statement. \square

⁵ $\lfloor x \rfloor = \max\{n \in \mathbb{Z} : n \leq x\}$ is the rounding down to the next integer.

Example 11.32. (a) Let $C_i > 0$ ($1 \leq i \leq 4$) be positive constants independent of ℓ with

$$\begin{aligned} C_1 I &\leq \frac{1}{2}(A_\ell + A_\ell^H) \leq C_2 h_\ell^{-2} I, \\ \|\frac{1}{2}(A_\ell - A_\ell^H)\|_2 &\leq C_3 h_\ell^{-1} I, \\ \|A_\ell\|_2 &\geq C_4 h_\ell^{-2} I. \end{aligned}$$

Set $\Theta = \Theta_\ell := h_\ell^2 C_1 / (C_1 C_2 + C_3^2)$ and $C := (C_1 C_2 + v C_3^2) / (C_1 C_4)$. Then the Richardson iteration damped by Θ_ℓ satisfies the smoothing property (11.61c) with the constant C above.

(b) Let \mathcal{S}_ℓ be the Jacobi or Gauss–Seidel iteration damped by $\vartheta = \frac{1}{2}$. Furthermore, A_ℓ is assumed to be weakly diagonally dominant. Then the smoothing property (11.61c) holds with $C = 2$ with respect to the row-sum norm $\|\cdot\|_\infty$.

Proof. (i) Theorem 3.30 proves (11.61a). (11.61b) follows with $C = 1/\Theta$.

(ii) Since $\vartheta = 1/2$, inequality (11.61a) is the estimation of the nondamped Jacobi or Gauss–Seidel iterations. Weak diagonal dominance implies (11.61a). From $\|D_\ell\|_\infty \leq \|D_\ell - E_\ell\|_\infty \leq \|A_\ell\|_\infty$ for $A = D - E - F$ (cf. (3.11a–d)), we conclude (11.61b) with $C = 1/\vartheta$. \square

11.6.3 Approximation Property

11.6.3.1 Formulation

For the coarse-grid correction, the fine-grid solution e_ℓ of $A_\ell e_\ell = d_\ell$ is replaced by $p e_{\ell-1}$ obtained from $A_{\ell-1} e_{\ell-1} = d_{\ell-1} := r d_\ell$. Therefore, $p e_{\ell-1} \approx e_\ell$, i.e., $p A_{\ell-1}^{-1} r d_\ell \approx A_\ell^{-1} d_\ell$ should be valid. We quantify this requirement by

$$\|p A_{\ell-1}^{-1} r d_\ell - A_\ell^{-1} d_\ell\|_2 \leq C_A \|d_\ell\|_2 / \|A_\ell\|_2 \quad \text{for } \ell \geq 1, d_\ell \in X_\ell.$$

This inequality can be rewritten by the matrix norm (spectral norm) as the *approximation property*

$$\|A_\ell^{-1} - p A_{\ell-1}^{-1} r\|_2 \leq C_A / \|A_\ell\|_2 \quad \text{for all } \ell \geq 1. \quad (11.63)$$

In general, proofs of the approximation property (11.63) are not of algebraic nature but use (at least indirectly) properties of the underlying boundary value problem. One possible route to the proof is as follows. Assume that $A_{\ell-1} = r A_\ell p$ holds according to (11.20). For an arbitrary restriction $r' : X_\ell \rightarrow X_{\ell-1}$, the following factorisation holds:

$$A_\ell^{-1} - p A_{\ell-1}^{-1} r = (I - p A_{\ell-1}^{-1} r A_\ell) A_\ell^{-1} = (I - p A_{\ell-1}^{-1} r A_\ell)(I - p r') A_\ell^{-1}.$$

Under suitable conditions,⁶ the solution $v_\ell := A_\ell^{-1} f_\ell$ is sufficiently smooth, so that the interpolation error

$$d_\ell = (I - pr')v_\ell = v_\ell - pr'v_\ell$$

can be estimated by $\|d_\ell\|_2 \leq C\|f_\ell\|_2/\|A_\ell\|_2$. The same tools can be used to show that $\|I - pA_{\ell-1}^{-1}rA_\ell\| \leq \text{const}$. Together, one obtains the approximation property (11.63). In case $A_{\ell-1}$ is not the Galerkin product, see Hackbusch [183, Criteria 6.3.35 and 6.3.38].

The easiest proof of the approximation property can be given for Galerkin discretisations. The discretisation, together with the prolongations and restrictions, is defined in §§11.6.3.2–11.6.3.3. The crucial part of the proof of the approximation property is given in §11.6.3.4.

11.6.3.2 Galerkin Discretisation

The boundary value problem is described in the variational form (E.5). Instead of a single finite-dimensional subspace $V_n \subset V$ we consider a hierarchy of subspaces

$$V_0 \subset V_1 \subset \dots \subset V_{\ell-1} \subset V_\ell \subset \dots \subset V,$$

where V_ℓ replaces the notation V_{n_ℓ} ($n_\ell = \dim(V_\ell)$) used in §E.2. Similarly, all mappings $P_n = P_{n_\ell}, \dots$ used in §§E.2–E.6 are now denoted by $P_\ell : V_\ell \rightarrow V, \dots$

11.6.3.3 Canonical Prolongation and Restriction

Section E.6 discusses the relation of Galerkin discretisations using two subspaces $V_{n'} \subset V_n$, now denoted by $V_{\ell-1} \subset V_\ell$. According to Proposition E.15, there are mappings $p : X_{\ell-1} \rightarrow X_\ell$ and $r : X_\ell \rightarrow X_{\ell-1}$ with

$$P_\ell p = P_{\ell-1}, \quad r = p^*, \quad rR_\ell = R_{\ell-1}. \quad (11.64)$$

Since p and r are the natural choice (see the diagram in (E.19)), they are called the *canonical prolongation* and the *canonical restriction*.

Remark 11.33. (a) Using the representation $p = \hat{R}_\ell P_{\ell-1}$ in (E.18) and the bounds in (E.10a,b) and (E.11c), we get the uniform estimates

$$\|p\|_{X_\ell \leftarrow X_{\ell-1}} = \|r\|_{X_{\ell-1} \leftarrow X_\ell} \leq \underline{C}_P \bar{C}_P \quad \text{for all } \ell \geq 1.$$

(b) The matrices A_ℓ and $A_{\ell-1}$ are connected by (11.20):

$$A_{\ell-1} = r A_\ell p \quad \text{for all } \ell \geq 1.$$

⁶ In the case of difference schemes, the theory of *discrete regularity* can be used; cf. Hackbusch [180, 181], [183, §6.3.2.1], [201, §9.2], and Jovanović–Süli [229].

11.6.3.4 Proof of the Approximation Property

Based on the $2m$ -regularity (E.13b), the error estimate (E.14) is proved in §E.5:

$$\|E_\ell\|_{U \leftarrow U} \leq C_E h_\ell^m \quad \text{for all } \ell \geq 1, \quad (11.65a)$$

where $E_\ell := A^{-1} - P_\ell A_\ell^{-1} R_\ell$. $2m$ is the order of the differential operator. The inverse estimate, together with the boundedness of the bilinear form, yields (E.12c):

$$\|A_\ell\|_2 \leq C_K h_\ell^{-2m} \quad \text{for all } \ell \geq 1. \quad (11.65b)$$

The inequality

$$\|\hat{R}_\ell\|_{X_\ell \leftarrow U} = \|\hat{P}_\ell\|_{U \leftarrow X_\ell} \leq \underline{C}_P \quad (11.65c)$$

is mentioned in (E.11c). A last condition for the approximation property is almost identical to the inequality $n_\ell \leq \bar{C}_h n_{\ell-1}$ used in §11.5.3:

$$h_{\ell-1} \leq C_h h_\ell \quad \text{for all } \ell \geq 1. \quad (11.65d)$$

Usually, (11.65d) holds with $C_h = 2$.

Theorem 11.34. *Let A_ℓ be the matrices (E.7b) of the Galerkin discretisation. Choose the canonical p and r . Assume (11.65a–d). Then the approximation property (11.63) holds.*

Proof. Use inequality (11.65a) for ℓ and $\ell - 1$:

$$\begin{aligned} & \|P_\ell A_\ell^{-1} R_\ell - P_{\ell-1} A_{\ell-1}^{-1} R_{\ell-1}\|_{U \leftarrow U} \\ &= \|E_{\ell-1} - E_\ell\|_{U \leftarrow U} \leq C_E (h_\ell^{2m} + h_{\ell-1}^{2m}). \end{aligned}$$

(11.65d) implies $h_{\ell-1}^{2m} \leq C_h^{2m} h_\ell^{2m}$. From

$$h_\ell^{2m} \leq C_K / \|A_\ell\|_2 \quad (\text{cf. (11.65b)}) \quad \text{and} \quad P_\ell = P_{\ell-1} p, \quad R_\ell = r R_\ell^{-1} \quad (\text{cf. (11.64)}),$$

we conclude that

$$\|P_\ell (A_\ell^{-1} - p A_{\ell-1}^{-1} r) R_\ell\|_{U \leftarrow U} \leq C' / \|A_\ell\|_2$$

with $C' := C_E C_K (1 + C_h^{2m})$. Multiplying $P_\ell (A_\ell^{-1} - p A_{\ell-1}^{-1} r) R_\ell$ by \hat{R}_ℓ from the left and by \hat{P}_ℓ from the right and using (E.11b,c), we obtain

$$\begin{aligned} \|A_\ell^{-1} - p A_{\ell-1}^{-1} r\|_2 &\leq \|\hat{R}_\ell\|_{X_\ell \leftarrow U} \|P_\ell (A_\ell^{-1} - p A_{\ell-1}^{-1} r) R_\ell\|_{U \leftarrow U} \|\hat{P}_\ell\|_{U \leftarrow X_\ell} \\ &\leq C' \underline{C}_P^2 / \|A_\ell\|_2, \end{aligned}$$

which is the approximation property with $C_A := C' \underline{C}_P^2$. □

11.6.4 Convergence of the Two-Grid Iteration

As mentioned in §11.2.2, $\rho(M_\ell^{\text{TGM}(\nu_1, \nu_2)}) = \rho(M_\ell^{\text{TGM}(\nu, 0)})$ holds for $\nu = \nu_1 + \nu_2$, so that we may restrict our considerations to $\nu = \nu_1 > 0$, $\nu_2 = 0$. This choice is optimal for statements concerning the contraction number $\|M_\ell^{\text{TGM}(\nu, 0)}\|_2$ with respect to the spectral norm. The following Theorems 11.35 and 11.36 correspond to the cases $\bar{\nu}(h) = \infty$ and $\bar{\nu}(h) < \infty$, respectively.

Theorem 11.35. *Assume the smoothing and approximation properties (11.58a–c), (11.63) with $\bar{\nu}(h) = \infty$. For given $0 < \zeta < 1$, there exists a lower bound $\underline{\nu}$ such that*

$$\|M_\ell^{\text{TGM}(\nu, 0)}\|_2 \leq C_A \eta(\nu) \leq \zeta \quad \text{for all } \nu \geq \underline{\nu}, \ell \geq 1. \quad (11.66)$$

Here, C_A and $\eta(\nu)$ are the quantities in (11.63) and (11.58a,b). By $\zeta < 1$, inequality (11.66) implies convergence of the two-grid iteration. Note that the contraction bound $C_A \eta(\nu)$ is independent of h_ℓ .

Proof. The two-grid iteration matrix can be factorised as follows:

$$M_\ell^{\text{TGM}(\nu, 0)} = (I - pA_{\ell-1}^{-1}rA_\ell)S_\ell^\nu = [A_\ell^{-1} - pA_{\ell-1}^{-1}r] [A_\ell S_\ell^\nu]$$

(cf. Lemma 11.11). Estimating both factors by (11.58a) and (11.63), we obtain the inequality (11.66). \square

Theorem 11.36. *Assume the smoothing and approximation properties (11.58a–c), (11.63), including the case $\bar{\nu}(h) < \infty$. For all $0 < \zeta < 1$, there exist bounds $\bar{h} > 0$ and $\underline{\nu}$ such that (11.66) holds for all $\nu \in [\underline{\nu}, \bar{\nu}(h))$ and all h_ℓ with $h_\ell \leq \bar{h}$, where the interval $[\underline{\nu}, \bar{\nu}(h))$ is not empty (i.e., $\underline{\nu} < \bar{\nu}(h)$).*

Proof. Choose $\underline{\nu}$ as in Theorem 11.35. Because of $\bar{\nu}(h) \rightarrow \infty$ ($h \rightarrow 0$), \bar{h} can be chosen such that $\bar{\nu}(h_\ell) > \underline{\nu}$ for all $h_\ell \leq \bar{h}$. \square

11.6.5 Convergence of the Multigrid Iteration

In Theorem 11.19, the representation $M_\ell^{\text{MGM}(\nu, 0)} = M_\ell^{\text{TGM}(\nu, 0)} - \dots$ of the multigrid iteration matrix is shown. We are exploiting the fact that the perturbation ‘...’ is sufficiently small; hence, two-grid convergence implies multigrid convergence. Besides the smoothing and approximation properties, we require additional conditions, which are easy to satisfy. The first one is

$$\|S_\ell^\nu\|_2 \leq C_S \quad \text{for all } \ell \geq 1, 0 < \nu < \bar{\nu} := \min_{\ell \geq 1} \bar{\nu}(h_\ell) \quad (11.67a)$$

with $\bar{\nu}(h_\ell)$ defined in (11.58c).

Exercise 11.37. Assume $S_\ell := S'_\ell + S''_\ell$. Let (11.67a) hold for S''_ℓ and assume (11.59c): $\lim_{\ell \rightarrow \infty} \|S''_\ell\|_2 = 0$. Prove (11.67a) for S_ℓ (similar to Theorem 11.29).

Exercise 11.38. Prove the inequalities

$$\underline{C}_p^{-1} \|x_{\ell-1}\|_2 \leq \|px_{\ell-1}\|_2 \leq \bar{C}_p \|x_{\ell-1}\|_2 \quad (x_{\ell-1} \in X_{\ell-1}, \ell \geq 1) \quad (11.67b)$$

for the canonical choice (11.64) by using (E.9) with $\underline{C}_p = \bar{C}_p := \underline{C}_P \bar{C}_P$.

The identity $pA_{\ell-1}^{-1}rA_\ell S_\ell^\nu = S_\ell^\nu - [A_\ell^{-1} - pA_{\ell-1}^{-1}r] A_\ell S_\ell^\nu = S_\ell^\nu - M_\ell^{\text{TGM}(\nu,0)}$ implies the next statement.

Lemma 11.39. Let (11.67a,b) be valid. Then

$$\|A_{\ell-1}^{-1}rA_\ell S_\ell^\nu\|_2 \leq \underline{C}_p(C_S + \|M_\ell^{\text{TGM}(\nu,0)}\|_2). \quad (11.67c)$$

Let $\nu = \nu_1 > 0$ and $\nu_2 = 0$ be the numbers of smoothing steps as in §11.6.4. Using (11.67b,c), we can estimate the multigrid iteration matrix in (11.42a,b) by

$$\begin{aligned} \|M_\ell^{\text{MGM}(\nu,0)}\|_2 &\leq \|M_\ell^{\text{TGM}(\nu,0)}\|_2 + C^* \|M_{\ell-1}^{\text{MGM}(\nu,0)}\|_2^\gamma \quad \text{for } \ell \geq 1 \\ \text{with } C^* &:= \underline{C}_p \bar{C}_p (C_S + 1). \end{aligned} \quad (11.68a)$$

Here, ν is assumed to be chosen large enough so that $\|M_\ell^{\text{TGM}(\nu,0)}\|_2 \leq 1$ according to Theorem 11.35 or 11.36. Together with $M_0^{\text{MGM}} = 0$ (cf. (11.42a)), inequality (11.68a) leads to the recursive inequalities (11.68c) for the quantities ζ_ℓ :

$$\zeta_\ell := \|M_\ell^{\text{MGM}(\nu,0)}\|_2 \quad (\ell \geq 0), \quad (11.68b)$$

$$\zeta_0 := 0, \quad \zeta_\ell \leq \zeta + C^* (\zeta_{\ell-1})^\gamma \quad \text{for } \ell \geq 1. \quad (11.68c)$$

ζ is the ℓ -independent bound for the two-grid convergence, whose existence is stated by Theorem 11.35 or 11.36:

$$\|M_\ell^{\text{TGM}(\nu,0)}\|_2 \leq \zeta. \quad (11.68d)$$

Analysing the fixed-point equation $x = \zeta + C^* x^\gamma$, we obtain the next result.

Lemma 11.40. Assume $\gamma \geq 2$, $C^* \gamma > 1$, and $\zeta \leq \frac{\gamma-1}{\gamma} / \sqrt[\gamma-1]{C^* \gamma}$. Then all solutions of the inequalities (11.68c) are bounded by

$$\zeta_\ell \leq \zeta^* \leq \frac{\gamma}{\gamma-1} \zeta < 1 \quad \text{for all } \ell \geq 0. \quad (11.69)$$

Exercise 11.41. For the most interesting case of $\gamma = 2$, prove that

$$\zeta^* = 2\zeta / (1 - \sqrt{1 - 4C^*\zeta}) \quad \text{for } \zeta^* \text{ in (11.69).}$$

Since, by (11.68b), ζ_ℓ are the contraction number bounds, we obtain the desired convergence result.

Theorem 11.42 (multigrid convergence). Assume the smoothing and the approximation properties (11.58a–c) and (11.63), the conditions (11.67a,b), and, in addition, $\gamma \geq 2$. As in Theorems 11.35 and 11.36, for every $0 < \zeta' < 1$ there are $\underline{\nu}$ and $\bar{h} > 0$ such that

$$\|M_\ell^{\text{MGM}(\nu,0)}\|_2 \leq \zeta' < 1 \quad \text{for } \underline{\nu} \leq \nu < \bar{\nu} := \min_{\ell \geq 1} \bar{\nu}(h_\ell),$$

provided that $h_1 \leq \bar{h}$. Here, $\underline{\nu} < \bar{\nu}$ holds. In the case of $\bar{\nu}(h) = \infty$, one may set $\bar{h} := \infty$ (i.e., the choice of the grid size is not restricted).

Proof. Choose $\zeta := \frac{\gamma-1}{\gamma} \zeta'$ small enough, so that ζ fulfils the assumptions of Lemma 11.40. According to Theorem 11.36, $\underline{\nu}$ and \bar{h} have to be chosen in such a way that (11.68d) holds: $\|M_\ell^{\text{TGM}(\nu,0)}\|_2 \leq \zeta$ for $\underline{\nu} \leq \nu < \bar{\nu}$. Lemmata 11.39 and 11.40 give $\zeta_\ell = \|M_\ell^{\text{MGM}(\nu,0)}\|_2 \leq \frac{\gamma}{\gamma-1} \zeta \leq \zeta'$. \square

11.6.6 Case of Weaker Regularity

The proof of the approximation property uses the $2m$ -regularity (cf. (E.13b)) which, in the case of the Poisson equation, is $A^{-1} = -\Delta^{-1} : U = L_2(\Omega) = H^0(\Omega) \rightarrow H^2(\Omega) \cap H_0^1(\Omega)$. This assumption is true for the unit square $\Omega = (0, 1) \times (0, 1)$ as for any convex domain, but it does not hold, e.g., for domains with re-entrant corners. In the general case, one obtains only statements of the form

$$A^{-1} : H^{-\sigma m}(\Omega) \rightarrow H^{(2-\sigma)m}(\Omega) \cap H_0^m(\Omega) \quad \text{for some } \sigma \in (0, 1)$$

(cf. Hackbusch [193, §9.1]). A similar statement may be assumed for A^* . If $\sigma < 1$, the approximation property (11.63) cannot be proved but has to be formulated by the help of other norms.

Let $|\cdot|_t$ for $-1 \leq t \leq 1$ be a discrete analogue of the Sobolev norm $H^{tm}(\Omega)$. We define $U_\ell := (X_\ell, |\cdot|_\sigma)$ and $F_\ell := (X_\ell, |\cdot|_{-\sigma})$. Then

$$\|A_\ell^{-1} - pA_{\ell-1}^{-1}r\|_{U_\ell \leftarrow F_\ell} \leq (C_A / \|A_\ell\|_2)^{1-\sigma} \quad (11.70)$$

can be shown (cf. Hackbusch [183, §6.3.1.3]; cf. (E.15)). For the notation of the norm on the left-hand side, compare with (B.11). For $A_\ell > 0$, the norms can be defined by

$$\|x_\ell\|_{U_\ell} = |x_\ell|_\sigma := \|A_\ell^{\sigma/2}x_\ell\|_2, \quad \|f_\ell\|_{F_\ell} = |f_\ell|_{-\sigma} := \|A_\ell^{-\sigma/2}f_\ell\|_2. \quad (11.71)$$

In the general case, replace the matrix A_ℓ in (11.71) by the positive definite part $A'_\ell := \frac{1}{2}(A_\ell + A_\ell^\text{H}) + c_0 I$ (cf. (11.60a)).

Part (11.58a) of the smoothing property (11.58a–c) has to be adapted to the new norms. Inequality (11.58a) becomes

$$\|A_\ell S_\ell^\nu\|_{F_\ell \leftarrow U_\ell} \leq \eta(\nu) \|A_\ell\|_2^{1-\sigma} \quad \text{for } 0 \leq \nu \leq \bar{\nu}(h_\ell). \quad (11.72)$$

Exercise 11.43. Let \mathcal{S}_ℓ be the Richardson iteration (11.55a,b) and assume $A_\ell > 0$. Using the norms in (11.71), prove for all $\nu \geq 0$ that

$$\|A_\ell S_\ell^\nu\|_{F_\ell \leftarrow U_\ell} = \|A_\ell^{1-\sigma}(I - \Theta A_\ell)^\nu\|_2 \leq \left[\eta_0(\frac{\nu}{1-\sigma}) \|A_\ell\|_2 \right]^{1-\sigma}. \quad (11.73)$$

The two-grid contraction number with respect to $\|\cdot\|_{U_\ell}$ can be concluded from the product of (11.70) and (11.72):

$$\|M_\ell^{\text{TGM}(\nu,0)}\|_{U_\ell \leftarrow U_\ell} \leq \eta(\nu) C_A^{1-\sigma}. \quad (11.74)$$

Similar to §11.6.5, we obtain a corresponding convergence result for the multigrid iteration.

Consider the bound $\eta(\nu) = [\eta_0(\frac{\nu}{1-\sigma}) C_A]^{1-\sigma}$ in (11.73). For the standard case discussed in §§11.6.2–11.6.5, we had $\sigma = 0$ and the bound $\eta(\nu)$ in (11.74) behaved as $\mathcal{O}(\frac{1}{\nu})$. For $0 < \sigma < 1$, the contraction number behaves as $\mathcal{O}(1/\nu^{1-\sigma})$. The value $\sigma = 1$ is not sufficient because $\eta(\nu)$ fails to fulfil (11.58b).

11.7 Symmetric Multigrid Methods

The multigrid analysis above addresses the general (nonsymmetric) case in order to emphasise that multigrid iterations are not restricted to symmetric or even only positive definite problems. However, the symmetric case admits some stronger statements that are covered in this chapter.

11.7.1 Symmetric and Positive Definite Multigrid Algorithms

We consider the two-grid algorithm (11.22b) and the multigrid iteration (11.33). The required symmetry conditions are

$$r = p^*, \quad \nu_1 = \nu_2 = \frac{\nu}{2}, \quad \hat{\mathcal{S}}_\ell = \mathcal{S}_\ell^* \text{ for all } \ell \geq 0 \quad (11.75a)$$

(cf. (11.17)). The second condition requires the post-smoothing $\hat{\mathcal{S}}_\ell$ to be adjoint to the pre-smoothing \mathcal{S}_ℓ . Occasionally, we need the Galerkin product property:

$$A_{\ell-1} = r A_\ell p. \quad (11.75b)$$

The Galerkin product (11.75b), together with $r = p^*$, ensures that $A_{\ell_{\max}} = A_{\ell_{\max}}^H$ implies $A_\ell = A_\ell^H$ for all $0 \leq \ell < \ell_{\max}$. Otherwise, this property must be required explicitly:

$$A_{\ell_{\max}} = A_{\ell_{\max}}^H \implies A_\ell = A_\ell^H \text{ for all } 0 \leq \ell < \ell_{\max}, \quad (11.75c)$$

Lemma 11.44. Let (11.75a,c) be valid. Then the two- and multigrid iterations $\Phi_\ell^{\text{TGM}(\frac{\nu}{2}, \frac{\nu}{2})}$ and $\Phi_\ell^{\text{MGM}(\frac{\nu}{2}, \frac{\nu}{2})}$ are symmetric: $\Phi_\ell^{\text{TGM}(\frac{\nu}{2}, \frac{\nu}{2})}, \Phi_\ell^{\text{MGM}(\frac{\nu}{2}, \frac{\nu}{2})} \in \mathcal{L}_{\text{sym}}$.

Proof. (i) We have to prove that the Hermitian symmetry of $A = A_{\ell_{\max}}$ implies the symmetry of the matrix $N_{\ell_{\max}} = N_{\ell_{\max}}^{\text{MGM}(\frac{\nu}{2}, \frac{\nu}{2})}$ of the second normal form (cf. (5.4)).

(ii) Assume $A_{\ell_{\max}} = A_{\ell_{\max}}^H$ and by (11.75c) that $A_\ell = A_\ell^H$ holds for all levels. First, we prove $\Phi_\ell^{\text{TGM}(\nu/2, \nu/2)} \in \mathcal{L}_{\text{sym}}$. Note that $\Phi_\ell^{\text{CGC}} \in \mathcal{L}_{\text{sym}}$, since $N_\ell^{\text{CGC}} = pA_{\ell-1}^{-1}r$ in Remark 11.6 is symmetric. By Corollary 5.30, the two-grid iteration $\Phi_\ell^{\text{TGM}(\frac{\nu}{2}, \frac{\nu}{2})} = (\mathcal{S}_\ell^*)^\nu \circ \Phi_\ell^{\text{CGC}} \circ \mathcal{S}_\ell^\nu = (\mathcal{S}_\ell^\nu)^* \circ \Phi_\ell^{\text{CGC}} \circ \mathcal{S}_\ell^\nu$ is also symmetric.

(iii) Next, we use the definition (11.42a,b) for an induction on ℓ . Assume that $\Phi_{\ell-1}^{\text{MGM}(\frac{\nu}{2}, \frac{\nu}{2})} \in \mathcal{L}_{\text{sym}}$. Then $(\Phi_{\ell-1}^{\text{MGM}(\frac{\nu}{2}, \frac{\nu}{2})})^\gamma$ is also symmetric and, by Criterion 5.5, the matrix $M_{\ell-1}^{\text{MGM}(\frac{\nu}{2}, \frac{\nu}{2})} A_{\ell-1}^{-1}$ is symmetric. The steps (11.33c–e) define an coarse-grid correction $\hat{\Phi}_\ell^{\text{CGC}}$ with the iteration matrix $M_\ell^{\text{CGC}} := p(M_{\ell-1}^{\text{MGM}})^\gamma A_{\ell-1}^{-1}rA_\ell$. Obviously, $M_\ell^{\text{CGC}} A_\ell^{-1} = p[(M_{\ell-1}^{\text{MGM}})^\gamma A_{\ell-1}^{-1}]r \in \mathcal{L}_{\text{sym}}$ holds, and Criterion 5.5 proves the symmetry of $\hat{\Phi}_\ell^{\text{CGC}}$. As in part (ii), the symmetry of $\Phi_\ell^{\text{MGM}(\nu/2, \nu/2)}$ follows from the representation $\Phi_\ell^{\text{MGM}(\nu/2, \nu/2)} = (\mathcal{S}_\ell^\nu)^* \circ \hat{\Phi}_\ell^{\text{CGC}} \circ \mathcal{S}_\ell^\nu$. \square

The positive definiteness of $\Phi_\ell^{\text{TGM}(\frac{\nu}{2}, \frac{\nu}{2})}$ and $\Phi_\ell^{\text{MGM}(\frac{\nu}{2}, \frac{\nu}{2})}$ is considered next.

Lemma 11.45. Assume (11.75a) and $A_\ell > 0$. Set $M_\ell := M_\ell^{\text{MGM}(\nu/2, \nu/2)}$ and $W_\ell := W_\ell^{\text{MGM}(\nu/2, \nu/2)}$. The following statements also hold for the two-grid case.

(a) Assume that the iteration $\Phi_\ell^{\text{MGM}(\frac{\nu}{2}, \frac{\nu}{2})}$ converges. Then it is positive definite, i.e., $\Phi_\ell^{\text{MGM}(\frac{\nu}{2}, \frac{\nu}{2})} \in \mathcal{L}_{\text{pos}}$, and converges monotonically with respect the energy norm $\|\cdot\|_{A_\ell}$. The transformed iteration matrices $A_\ell^{1/2} M_\ell A_\ell^{-1/2}$ are Hermitian. The matrix W_ℓ of the third normal form is positive definite and fulfills

$$(1 - \rho_\ell)W_\ell \leq A_\ell \leq (1 + \rho_\ell)W_\ell \quad \text{with } \rho_\ell = \rho(M_\ell) = \|A_\ell^{1/2} M_\ell A_\ell^{-1/2}\|_2. \quad (11.76a)$$

If, according to Theorems 11.35 or 11.42, $\rho_\ell \leq \rho < 1$ is h_ℓ -independent, then the condition (11.76b) is also h_ℓ -independent:

$$\kappa(W_\ell^{-1} A_\ell) \leq \frac{1 + \rho_\ell}{1 - \rho_\ell} \leq \frac{1 + \rho}{1 - \rho}. \quad (11.76b)$$

(b) In the case of (11.75b), the inequalities (11.76a,b) can be improved:

$$(1 - \rho_\ell)W_\ell \leq A_\ell \leq W_\ell, \quad \kappa(W_\ell^{-1} A_\ell) \leq 1/(1 - \rho_\ell) \leq 1/(1 - \rho).$$

Proof. The representations (11.23) and (11.42a,b) show that $A_\ell M_\ell A_\ell^{-1} = M_\ell^H$, because $A_\ell \hat{S}_\ell A_\ell^{-1} = S_\ell^H$ according to (5.2b). This proves part (a). Part (b) is obtained from Theorem 3.34c. Part (c) is based on the property $A_\ell^{1/2} M_\ell A_\ell^{-1/2} \geq 0$, which will be proved in (11.87b). \square

11.7.2 Two-Grid Convergence for $\nu_1 > 0, \nu_2 > 0$

The case $\nu_1 = \nu > 0$ and $\nu_2 = 0$ is treated in §11.6. The technique used there can also be applied to the general case $\nu_1 \geq 0, \nu_2 \geq 0, \nu := \nu_1 + \nu_2 > 0$, and especially to $\nu_1 = \nu_2 = \nu/2$.

Exercise 11.46. Assume (11.75a) without the condition $\nu_1 = \nu_2$. Prove

$$\Phi_\ell^{\text{TGM}(\nu_1, \nu_2)} = \left(\Phi_\ell^{\text{TGM}(\nu_2, \nu_1)} \right)^*, \quad (11.77\text{a})$$

$$\Phi_\ell^{\text{TGM}(\nu_1, \nu_2)} = \Phi_\ell^{\text{TGM}(0, \nu_2)} \circ \Phi_\ell^{\text{TGM}(\nu_1, 0)} \text{ in the case of (11.75b).} \quad (11.77\text{b})$$

Under assumption (11.75b), the statements (11.77c,d) for the two-grid iteration matrices $M_\ell(\nu_1, \nu_2) := M_\ell^{\text{TGM}(\nu_1, \nu_2)}$ follow from (11.77a,b):

$$M_\ell(\nu_1, \nu_2) = M_\ell(0, \nu_2)M_\ell(\nu_1, 0) = A_\ell^{-1}M_\ell(\nu_2, 0)^H A_\ell M_\ell(\nu_1, 0), \quad (11.77\text{c})$$

$$A_\ell^{\frac{1}{2}}M_\ell(\nu_1, \nu_2)A_\ell^{-\frac{1}{2}} = \left(A_\ell^{\frac{1}{2}}M_\ell(\nu_2, 0)A_\ell^{-\frac{1}{2}} \right)^H \left(A_\ell^{\frac{1}{2}}M_\ell(\nu_1, 0)A_\ell^{-\frac{1}{2}} \right). \quad (11.77\text{d})$$

For estimating $A_\ell^{1/2}M_\ell(\nu, 0)A_\ell^{-1/2}$, we may use the approximation property (11.78a) and the smoothing property (11.78b):

$$\|A_\ell^{1/2}(A_\ell^{-1} - pA_{\ell-1}^{-1}r)\|_2 \leq \sqrt{C_A / \|A_\ell\|_2}, \quad (11.78\text{a})$$

$$\|A_\ell S_\ell^\nu A_\ell^{-1/2}\|_2 \leq \sqrt{\eta(2\nu) \|A_\ell\|_2}, \quad (11.78\text{b})$$

which correspond to (11.70) and (11.72) for the energy norm $\|\cdot\|_{U_\ell} = \|\cdot\|_{A_\ell}$ and the Euclidean norm $\|\cdot\|_{F_\ell} = \|\cdot\|_2$. Under assumption (11.75b), inequality (11.78a) is equivalent to the approximation property (11.63). In the case of Richardson's iteration (11.55a,b), inequality (11.78b) holds because of

$$\|A_\ell S_\ell^\nu A_\ell^{-1/2}\|_2 = \|A_\ell^{1/2}S_\ell^\nu\|_2^2 = \|A_\ell S_\ell^{2\nu}\|_2 \leq \eta_0(2\nu) \|A_\ell\|_2$$

with $\eta(2\nu) = \eta_0(2\nu)$ (η_0 in (11.56)). The inequalities (11.78a,b) yield the estimate

$$\|M_\ell^{\text{TGM}(\nu, 0)}\|_{A_\ell} = \|A_\ell^{1/2}M_\ell^{\text{TGM}(\nu, 0)}A_\ell^{-1/2}\|_2 \leq \sqrt{\eta(2\nu)C_A}.$$

Using (11.77d), we finally prove the following convergence theorem.

Theorem 11.47. Assume (11.75a,b) without $\nu_1 = \nu_2$. The smoothing and approximation properties (11.78a,b) imply

$$\begin{aligned} \|M_\ell^{\text{TGM}(\nu_1, \nu_2)}\|_{A_\ell} &\leq C_A \sqrt{\eta(2\nu_1)\eta(2\nu_2)}, \\ \|M_\ell^{\text{TGM}(\nu/2, \nu/2)}\|_{A_\ell} &\leq C_A \eta(\nu). \end{aligned}$$

Two-grid convergence follows as in Theorems 11.35–11.36.

As in §11.6.5, multigrid convergence can be concluded from the two-grid convergence. However, it has to be emphasised that the proof technique in §11.6.5 requires $\gamma \geq 2$ and therefore excludes the V-cycle ($\gamma = 1$).

11.7.3 Smoothing Property in the Symmetric Case

In particular, condition (11.75b): $\hat{S}_\ell = S_\ell^*$ is satisfied if $\hat{\mathcal{S}}_\ell = \mathcal{S}_\ell$ is a symmetric smoothing iteration. For symmetric iterations, the proof of the smoothing property is rather easy.

Lemma 11.48. *Let $S_\ell = I - W_\ell^{-1}A_\ell$ be the iteration matrix of a positive definite iteration \mathcal{S}_ℓ and assume that $\gamma W_\ell \leq A_\ell \leq \Gamma W_\ell$ for all $\ell \geq 0$ with $0 \leq \gamma \leq \Gamma < 2$. Then*

$$\|A_\ell S_\ell^\nu\|_2 \leq \|W_\ell\|_2 \max\{\eta_0(\nu), \Gamma |1 - \Gamma|^\nu\}$$

implies the smoothing property (11.58a–c) with $\bar{\nu}(h) = \infty$ if there is some C_W with

$$\|W_\ell\|_2 \leq C_W \|A_\ell\|_2 \quad \text{for all } \ell > 0. \quad (11.79)$$

Proof. Define $Y := W_\ell^{-1/2} R_\ell W_\ell^{-1/2}$ with R_ℓ in $A_\ell = W_\ell - R_\ell$ and note that

$$\|A_\ell S_\ell^\nu\|_2 = \|W_\ell^{1/2} (I - Y) Y^\nu W_\ell^{1/2}\|_2 \leq \|W_\ell^{1/2}\|_2^2 \|(I - Y) Y^\nu\|_2.$$

The first factor is equal to $\|W_\ell\|_2$, the second can be estimated as in Lemma 11.23 by $\max\{\eta_0(\nu), \Gamma |1 - \Gamma|^\nu\}$ because of $(1 - \Gamma)I \leq Y \leq (1 - \gamma)I$. \square

The following variant of the estimate is due to Wittum [403]. The estimate is helpful if good bounds for $\|R_\ell\|_2$ are known.

Lemma 11.49. *In addition to the assumptions of Lemma 11.48, assume $\nu \geq 2$. Define $R_\ell := W_\ell - A_\ell$. Then*

$$\|A_\ell S_\ell^\nu\|_2 \leq \|S_\ell\|_2 \|R_\ell\|_2 \max\{\eta_0(\nu - 2), \Gamma |1 - \Gamma|^{\nu-2}\}.$$

Proof. Define Y as above, estimate $\|A_\ell S_\ell^\nu\|_2$ by $\|W_\ell^{1/2} Y\|_2^2 \|(I - Y) Y^{\nu-2}\|_2$ and use $\|W_\ell^{1/2} Y\|_2^2 = \|W_\ell^{-1/2} R_\ell^2 W_\ell^{-1/2}\|_2 = \rho(W_\ell^{-1/2} R_\ell^2 W_\ell^{-1/2}) = \rho(W_\ell^{-1} R_\ell^2) \leq \|W_\ell^{-1} R_\ell\|_2 \|R_\ell\|_2 = \|S_\ell\|_2 \|R_\ell\|_2$. \square

Exercise 11.50. Prove under the same assumption as in Lemma 11.48 that

$$\|A_\ell S_\ell^\nu A_\ell^{-1}\|_2 \leq \sqrt{\|W_\ell\|_2 \max\{\eta_0(2\nu), \Gamma |1 - \Gamma|^{2\nu}\}}. \quad (11.80)$$

Inequality (11.80) can be regarded as a modification of (11.78b).

The condition $\Gamma < 2$ in $0 < \gamma W_\ell \leq A_\ell \leq \Gamma W_\ell$ coincides with the convergence condition in Theorem 3.34b. However, $\gamma = 0$ is sufficient for the smoothing property, although the convergence rate $\rho(S_\ell)$ becomes worse the smaller γ is. Since damping corresponds to the replacement of W_ℓ by $\vartheta^{-1}W_\ell$, we obtain the following.

Remark 11.51. After a possibly necessary damping, all positive definite iterations satisfy the assumption $\gamma W_\ell \leq A_\ell \leq \Gamma W_\ell$ with $0 \leq \gamma \leq \Gamma < 2$.

11.7.4 Strengthened Two-Grid Convergence Estimates

To simplify the following considerations, the smoothing iteration \mathcal{S}_ℓ is assumed to satisfy the inequality $\gamma W_\ell \leq A_\ell \leq \Gamma W_\ell$ with $0 \leq \gamma \leq \Gamma \leq 1$. As mentioned in Remark 11.51, this assumption can always be achieved by suitable damping. However, the following statements also hold in a somewhat modified form for $0 \leq \gamma \leq \Gamma < 2$. The assumptions are

$$\hat{\mathcal{S}}_\ell = \mathcal{S}_\ell, \quad \hat{S}_\ell = S_\ell = I - W_\ell^{-1} A_\ell, \quad 0 < A_\ell \leq W_\ell. \quad (11.81)$$

The approximation property is required in the form (11.70) with $\|\cdot\|_{U_\ell} := \|\cdot\|_{W_\ell}$, $\|\cdot\|_{F_\ell} := \|\cdot\|_{W_\ell^{-1}}$ (see the following Remark 11.57):

$$\|W_\ell^{1/2}(A_\ell^{-1} - pA_{\ell-1}^{-1}r)W_\ell^{1/2}\|_2 \leq C_A \quad \text{for all } \ell \geq 1. \quad (11.82)$$

Lemma 11.52. *Assume (11.75a,b). The approximation property (11.82) is equivalent to the following inequality:*

$$0 \leq A_\ell^{-1} - pA_{\ell-1}^{-1}r \leq C_A W_\ell^{-1} \quad \text{for all } \ell \geq 1. \quad (11.83)$$

Proof. (C.3f) yields $-C_A I \leq W_\ell^{\frac{1}{2}}(A_\ell^{-1} - pA_{\ell-1}^{-1}r)W_\ell^{\frac{1}{2}} \leq C_A I$. Multiplying by $W_\ell^{-\frac{1}{2}}$ from both sides yields the bounds $\pm C_A W_\ell^{-1}$ for $A_\ell^{-1} - pA_{\ell-1}^{-1}r$. The lower bound $-C_A I$ can be replaced by 0, as can be concluded from Lemma 11.53. \square

We postpone the proof of the modified approximation property (11.82) until Remark 11.57. Now we transform all quantities into a form better suited to symmetry:

$$\begin{aligned} \check{p} &:= A_\ell^{1/2} p A_{\ell-1}^{-1/2}, & \check{r} &:= \check{p}^* = A_{\ell-1}^{-1/2} r A_\ell^{1/2}, & Q_\ell &:= I - \check{p} \check{r}, \\ X_\ell &:= A_\ell^{1/2} W_\ell^{-1} A_\ell^{1/2}, & \check{S}_\ell &:= A_\ell^{1/2} S_\ell A_\ell^{-1/2} = I - X_\ell. \end{aligned}$$

Since (11.75b) can be rewritten as $\check{r} \check{p} = I$, the following lemma can be concluded.

Lemma 11.53. *Under the assumption (11.75a,b), $Q_\ell = I - \check{p} \check{r}$ is an orthogonal projection: $Q_\ell = Q_\ell^\mathsf{H}$. As any orthogonal projection, it fulfills*

$$0 \leq Q_\ell \leq I \quad \text{for all } \ell \geq 1. \quad (11.84a)$$

$Q_\ell \geq 0$ also implies $0 \leq A_\ell^{-1/2} Q_\ell A_\ell^{-1/2} = A_\ell^{-1} - pA_{\ell-1}^{-1}r$, so that the proof of the first inequality in (11.83) is completed. Multiplying (11.83) by $A_\ell^{1/2}$ from both sides yields the next lemma.

Lemma 11.54. *Assume (11.75a,b). The statements (11.82) or (11.83) are equivalent to*

$$0 \leq Q_\ell \leq C_A X_\ell \quad \text{for all } \ell \geq 1. \quad (11.84b)$$

According to (11.23), the transformed two-grid iteration matrix is

$$\check{M}_\ell(\nu_1, \nu_2) := A_\ell^{1/2} M_\ell^{\text{TGM}(\nu_1, \nu_2)} A_\ell^{-1/2} = \check{S}_\ell^{\nu_2} Q_\ell \check{S}_\ell^{\nu_1}. \quad (11.85)$$

In contrast to Theorems 11.35 to 11.55, it is now possible to prove convergence for all $\nu > 0$.

Theorem 11.55 (two-grid convergence). *Assume (11.75a,b), (11.81), and the approximation property (11.82). Then the two-grid iteration converges monotonically with respect to the energy norm $\|\cdot\|_{A_\ell}$:*

$$\begin{aligned} \rho(M_\ell^{\text{TGM}(\nu/2, \nu/2)}) &= \|M_\ell^{\text{TGM}(\nu/2, \nu/2)}\|_{A_\ell} \\ &= \left\| \check{M}_\ell \left(\frac{\nu}{2}, \frac{\nu}{2} \right) \right\|_2 \leq \left\{ \begin{array}{ll} C_A \eta_0(\nu) & \text{if } C_A \leq 1 + \nu \\ (1 - 1/C_A)^\nu & \text{if } C_A > 1 + \nu \end{array} \right\} < 1. \end{aligned} \quad (11.86)$$

Proof. It remains to show the inequality ‘ \leq ’ in (11.86). The inequality (11.87a) following from (11.84a,b) can be inserted into (11.85) and yields (11.87b):

$$0 \leq Q_\ell \leq \alpha C_A X_\ell + (1 - \alpha) I \quad \text{for all } 0 \leq \alpha \leq 1, \quad (11.87a)$$

$$0 \leq \check{M}_\ell \leq \check{S}_\ell^{\nu/2} [\alpha C_A X_\ell + (1 - \alpha) I] \check{S}_\ell^{\nu/2} \quad \text{for all } 0 \leq \alpha \leq 1. \quad (11.87b)$$

Since $\check{S}_\ell = I - X_\ell$, the right-hand side of (11.87b) is a polynomial $f(X_\ell; \alpha)$ with

$$f(\xi; \alpha) := (1 - \xi)^\nu (1 - \alpha + \alpha C_A \xi). \quad (11.87c)$$

For all $0 \leq \alpha \leq 1$, inequality $0 \leq X_\ell \leq I$ (cf. (11.81)) implies the estimate

$$\|\check{M}_\ell\|_2 \leq \|f(X_\ell; \alpha)\|_2 \leq m(\alpha) := \max\{f(\xi; \alpha) : 0 \leq \xi \leq 1\}.$$

In particular, for $\alpha = 1$, we obtain the bound $C_A \eta_0(\nu)$. If $1 + \nu < C_A$, the value $\alpha^* := \frac{\nu}{C_A - 1}$ belongs to $[0, 1]$ and yields the better bound $m(\alpha^*) = (1 - \frac{1}{C_A})^\nu$. \square

Exercise 11.56. Prove the statements of Lemmata 11.52, 11.54 and Theorem 11.55 under the assumption $rA_\ell p \leq A_{\ell-1}$ instead of (11.75b).

It remains to discuss the approximation property (11.82).

Remark 11.57. Assume the approximation property in the original form (11.63): $\|A_\ell^{-1} - p A_{\ell-1}^{-1} r\|_2 \leq C'_A / \|A_\ell\|_2$. Furthermore, let (11.79) be valid: $\|W_\ell\|_2 \leq C_W \|A_\ell\|_2$. Then (11.82) is satisfied by $C_A := C'_A C_W$.

Exercise 11.58. Assume (11.75a,b) without $\nu_1 = \nu_2$, as well as (11.81) and $\nu = \nu_1 + \nu_2 > 0$. Prove that the two-grid iteration $\Phi_\ell^{\text{TGM}(\nu_1, \nu_2)}$ converges monotonically with respect to the energy norm $\|\cdot\|_{A_\ell}$. What is the h_ℓ -independent contraction number? Hint: First, use (11.77d) to estimate $\|\check{M}_\ell(\frac{\nu}{2}, 0)\|_2$ and thereafter apply (11.77d) to $\|\check{M}_\ell(\nu_1, \nu_2)\|_2$.

In the proof of Theorem 11.55, the smoothing property is also used indirectly; however, now it is formulated by the polynomial (11.87c) for arbitrary $0 \leq \alpha \leq 1$ instead of $\alpha = 1$.

11.7.5 V-Cycle Convergence

We apply the technique of §11.7.4 to the multigrid method. Since Theorem 11.42 excludes the V-cycle ($\gamma = 1$), we concentrate on this case. For another proof, see Braess–Hackbusch [65].

Theorem 11.59. *Under the same assumptions (11.75a,b), (11.81), (11.82) as in Theorem 11.55, the V-cycle ($\gamma = 1$) converges monotonically with respect to the energy norm $\|\cdot\|_{A_\ell}$ with the rate*

$$\rho(M_\ell^V(\frac{\nu}{2}, \frac{\nu}{2})) = \|M_\ell^V(\frac{\nu}{2}, \frac{\nu}{2})\|_{A_\ell} \leq \frac{C_A}{C_A + \nu}$$

Proof. For $\gamma = 1$, abbreviate M_ℓ^{MGM} by M_ℓ^V . The recursive equations (11.42a,b) become

$$M_0^V(\nu_1, \nu_2) = 0, \quad M_\ell^V(\nu_1, \nu_2) = M_\ell^{\text{TGM}(\nu_1, \nu_2)} + S_\ell^\nu p M_{\ell-1}^V(\nu_1, \nu_2) A_{\ell-1}^{-1} r A_\ell S_\ell^\nu.$$

Transformation into the symmetric form yields

$$\begin{aligned} \check{M}_\ell^V &:= A_\ell^{1/2} M_\ell^V(\nu_1, \nu_2) A_\ell^{-1/2} = \check{M}_\ell^{\text{TGM}(\nu_1, \nu_2)} + \check{S}_\ell^{\nu_2} \check{p} \check{M}_{\ell-1}^V \check{r} \check{S}_\ell^{\nu_1} & (11.88) \\ (11.85) &= \check{S}_\ell^{\nu_2} \{I - \check{p} [I - \check{M}_{\ell-1}^V] \check{r}\} \check{S}_\ell^{\nu_1} \quad \text{for } \ell \geq 1, \quad \check{M}_0^V = 0. \end{aligned}$$

In the following, choose $\check{M}_\ell^V := \check{M}_\ell^V(\frac{\nu}{2}, \frac{\nu}{2})$, i.e., $\nu_1 = \nu_2 = \frac{\nu}{2}$. Using (11.88), we obtain

$$\check{M}_\ell^V \geq 0$$

by induction: $\check{M}_0^V = 0$ and $I - \check{p}[I - \check{M}_{\ell-1}^V]\check{r} \geq I - \check{p}\check{r} = Q_\ell \geq 0$. Hence, the statements (11.89a) and (11.89b) are equivalent:

$$\|M_\ell^V(\frac{\nu}{2}, \frac{\nu}{2})\|_{A_\ell} = \|\check{M}_\ell^V\|_2 \leq \zeta_\ell \quad (\check{M}_\ell^V := \check{M}_\ell^V(\frac{\nu}{2}, \frac{\nu}{2})), \quad (11.89a)$$

$$0 \leq \check{M}_\ell^V \leq \zeta_\ell I. \quad (11.89b)$$

The induction hypothesis is $0 \leq \check{M}_{\ell-1}^V \leq \zeta_{\ell-1} I$ with $\zeta_{\ell-1} := \frac{C_A}{C_A + \nu}$. Inserting this inequality into (11.88), we arrive at

$$\begin{aligned} 0 \leq \check{M}_\ell^V &\leq \check{S}_\ell^{\nu/2} \{I - (1 - \zeta_{\ell-1})\check{p}\check{r}\} \check{S}_\ell^{\nu/2} = \check{S}_\ell^{\nu/2} \{(1 - \zeta_{\ell-1})Q_\ell + \zeta_{\ell-1}I\} \check{S}_\ell^{\nu/2} \\ (11.87a) \quad &\leq \check{S}_\ell^{\nu/2} \{(1 - \zeta_{\ell-1})[\alpha C_A X_\ell + (1 - \alpha)I] + \zeta_{\ell-1}I\} \check{S}_\ell^{\nu/2} \end{aligned}$$

for all $0 \leq \alpha \leq 1$. For $\alpha \in [0, 1]$, the variable $\beta := (1 - \zeta_{\ell-1})(1 - \alpha) + \zeta_{\ell-1}$ varies in $[\zeta_{\ell-1}, 1]$. Substitution of α by β yields

$$0 \leq \check{M}_\ell^V \leq \check{S}_\ell^{\nu/2} \{(1 - \beta)C_A X_\ell + \beta I\} \check{S}_\ell^{\nu/2} \quad \text{for all } \zeta_{\ell-1} \leq \beta \leq 1.$$

The right-hand side is the polynomial $f(\xi; \beta) := (1 - \xi)^\nu [\beta + (1 - \beta)C_A \xi]$ for $\xi = X_\ell$ and can be estimated by

$$\|f(X_\ell; \beta)\|_2 \leq m(\beta) := \max\{|f(\xi; \beta)| : 0 \leq \xi \leq 1\} \quad (\text{cf. (11.87c,d)}).$$

For $\beta = \zeta_{\ell-1} = C_A/(C_A + \nu)$, one finds $m(\beta) = f(0; \beta) = \beta = C_A/(C_A + \nu)$. Hence, (11.89b) holds with $\zeta_\ell = C_A/(C_A + \nu)$. \square

Exercise 11.60. (a) Under the same assumptions, prove

$$\check{M}_\ell^V(0, \nu_2) \check{M}_\ell^V(\nu_1, 0) = \check{M}_\ell^V(\nu_1, \nu_2)$$

and discuss convergence for $\nu = \nu_1 + \nu_2 > 0$.

(b) Prove the statement of Theorem 11.59 under the weaker condition $rA_\ell p \leq A_{\ell-1}$ instead of (11.75b).

The condition $A_\ell \leq W_\ell$ in (11.81) can be generalised to $A_\ell \leq \sqrt{2} W_\ell$ (cf. Wittum [402, Proposition 4.2.4]).

Obviously, monotone and h_ℓ -independent convergence can also be shown for the W-cycle (more generally, for $\gamma \geq 2$). For this case (assuming $C_A \geq 1$), one finds, e.g., the estimate

$$\|M_\ell^W(\nu/2, \nu/2)\|_{A_\ell} \leq \sqrt{C_A} / (\sqrt{C_A} + \nu).$$

In the case of weaker regularity (cf. §11.6.6) and for $\gamma = 2$ (W-cycle), one can still prove $\|M_\ell^W(\nu/2, \nu/2)\|_{A_\ell} \leq \mathcal{O}(\nu^{\sigma-1}) < 1$ for all $\nu > 0$.

V-cycle convergence without full regularity assumptions is proved by Brenner [79]. See also §12.9.3.

11.7.6 Unsymmetric Multigrid Convergence for all $\nu > 0$

The analysis in §11.6 shows multigrid convergence for sufficiently large $\nu \geq \underline{\nu}$. In the symmetric case, §11.7.5 ensures convergence for all $\nu > 0$ and arbitrarily coarse h_0 . In the general case, we still obtain convergence for all $\nu = \nu_1 + \nu_2 > 0$; however, h_0 must be sufficiently small: $h_0 \leq \bar{h}$. The proof technique is the same as for Theorem 11.29.

Theorem 11.61. *Let the matrices A_ℓ ($\ell \geq 0$) be split into $A_\ell = A'_\ell + A''_\ell$ such that $A'_\ell > 0$. Let S_ℓ and S'_ℓ be the iteration matrices of the corresponding smoothing iterations \mathcal{S}_ℓ and \mathcal{S}'_ℓ . For A''_ℓ and $S''_\ell := S_\ell - S'_\ell$, assume*

$$\|A'^{-1/2} A''_\ell A'^{-1/2}\|_2 \leq C_1 h_\ell^\kappa, \quad \|A'^{1/2} S''_\ell A'^{-1/2}\|_2 \leq C_2 h_\ell^\kappa \quad (11.90a)$$

with $\kappa > 0$. Assume that the following norms are bounded by 1:

$$\|A'^{1/2} S'_\ell A'^{-1/2}\|_2, \|A'^{1/2} p A'^{-1/2}\|_2, \|A'^{-1/2} r A'^{1/2}\|_2 \leq 1 \quad (11.90b)$$

for all $\ell \geq 1$ and that the two- or multigrid method for A'_ℓ (with fixed parameters γ, ν_1, ν_2) converges monotonically with respect to the energy norm $\|\cdot\|_{A'_\ell}$ with the

contraction number ζ' . Further, let

$$\sup\{h_\ell/h_{\ell-1} : \ell \geq 1\} < 1 \quad \text{and} \quad \varepsilon \in (0, 1 - \zeta')$$

be valid. Then the two- and multigrid iterations for A_ℓ also converge monotonically with respect to the energy norm $\|\cdot\|_{A'_\ell}$ with the contraction number $\zeta = \zeta' + \varepsilon < 1$, provided that $h_0 \leq \bar{h}$ holds with sufficiently small \bar{h} .

Proof. First, the two-grid case is considered. The transformed iteration matrix $A_\ell'^{1/2} M'_\ell A_\ell'^{-1/2}$ (of the iteration for $A'_\ell x'_\ell = b'_\ell$) is the product

$$\begin{aligned} & \left[A_\ell'^{1/2} S'_\ell A_\ell'^{-1/2} \right]^{\nu_2} \times \left[A_\ell'^{1/2} (A_\ell'^{-1} - p A_{\ell-1}'^{-1} r) A_\ell'^{1/2} \right] \\ & \times \left[A_\ell'^{-1/2} A'_\ell A_\ell'^{-1/2} \right] \times \left[A_\ell'^{1/2} S'_\ell A_\ell'^{-1/2} \right]^{\nu_1}. \end{aligned}$$

Because of (11.90a), perturbations of S'_ℓ and A'_ℓ in the 1st, 3rd, and 4th factor by S''_ℓ and A''_ℓ , respectively, enlarge the spectral norm only by $\mathcal{O}(h_\ell^\kappa)$. A similar statement holds for the second factor because

$$\begin{aligned} & \left[A_\ell'^{1/2} (A_\ell^{-1} - p A_{\ell-1}'^{-1} r) A_\ell'^{1/2} \right] - \left[A_\ell'^{1/2} (A_\ell'^{-1} - p A_{\ell-1}'^{-1} r) A_\ell'^{1/2} \right] \\ & = A_\ell'^{-1/2} A''_\ell A_\ell^{-1} A_\ell'^{1/2} + A_\ell'^{1/2} p A_{\ell-1}'^{-1} A''_{\ell-1} A_{\ell-1}'^{-1} r A_\ell'^{1/2}. \end{aligned}$$

Let M_ℓ be the two-grid iteration matrix associated with the matrix A_ℓ . The assertion follows from $\left| \|M_\ell\|_{A'_\ell} - \|M'_\ell\|_{A'_\ell} \right| \leq C h_\ell^\kappa \leq C \bar{h}^\kappa$ for the choice $\bar{h} := (\varepsilon/C)^{1/\kappa}$. In the multigrid case, the following recursive estimate holds:

$$\left| \|M_\ell\|_{A'_\ell} - \|M'_\ell\|_{A'_\ell} \right| \leq C_0 h_\ell^\kappa + \left| \|M_{\ell-1}\|_{A'_{\ell-1}} - \|M'_{\ell-1}\|_{A'_{\ell-1}} \right|,$$

which by $h_\ell/h_{\ell-1} \leq C_h < 1$ leads to $\left| \|M_\ell\|_{A'_\ell} - \|M'_\ell\|_{A'_\ell} \right| \leq C h_0^\kappa \leq C \bar{h}^\kappa \leq \varepsilon$. \square

Remark 11.62. The conditions in (11.90b) are satisfied if

$$S'_\ell = I - W_\ell'^{-1} A'_\ell \quad \text{with} \quad 2W_\ell' \geq A'_\ell, \quad r = p^*, \quad r A'_\ell p \leq A'_{\ell-1}$$

(cf. (11.81), Exercises 11.56, and 11.60b). (11.75b) is sufficient for $r A'_\ell p \leq A'_{\ell-1}$.

The statement of Theorem 11.61 is not yet uniform with respect to $\nu = \nu_1 + \nu_2$. In particular, \bar{h} might depend on ν . A ν -independent \bar{h} can be obtained as follows: Theorem 11.42 (modified according to §11.7.2 to the energy norm $\|\cdot\|_{A'_\ell}$) shows convergence for $\nu \geq \underline{\nu}$ as long as $h_0 \leq \bar{h}_0$. For the finitely many $\nu = 1, \dots, \underline{\nu} - 1$, we conclude convergence for $h_0 \leq \bar{h}_\nu$ from Theorem 11.61 with suitable \bar{h}_ν . For $h_0 \leq \bar{h} := \min\{\bar{h}_\nu : 0 \leq \nu \leq \underline{\nu} - 1\}$, we obtain convergence for all $\nu > 0$.

For related results, see Mandel [269] and Bramble–Pasciak–Xu [75].

11.8 Combination of Multigrid Methods with Semi-Iterations

11.8.1 Semi-Iterative Smoothers

So far, only the ν -fold application of a smoothing iteration S_ℓ has been considered as a smoothing step (11.33b,f). An alternative is a semi-iterative smoothing, where S_ℓ^ν is replaced by a polynomial $P_\nu(S_\ell)$ of degree ν with $P_\nu(1) = 1$. However, one should not choose the polynomials that were found to be optimal in §8 because those minimise $\rho(P_\nu(S_\ell))$. Using S_ℓ for smoothing, we do not primarily want to make the error small but smooth. The smoothing property (11.58a) leads us to the following optimisation problem:

$$\text{minimise } \|A_\ell P_\nu(S_\ell)\|_2 \text{ over } P_\nu \in \mathcal{P}_\nu \quad \text{with} \quad P_\nu(1) = 1. \quad (11.91)$$

The semi-iterative Richardson method with $A_\ell > 0$ and $\sigma_M := [0, \|A_\ell\|_2]$ yields the optimisation problem

$$\min_{P_\nu \in \mathcal{P}_\nu, P_\nu(1)=1} \max_{0 \leq \xi \leq \|A_\ell\|_2} \left| \xi P_\nu \left(1 - \frac{\xi}{\|A_\ell\|_2} \right) \right| \quad (11.92)$$

(analogous to (8.23)). The solution reads as follows.

Theorem 11.63. *Let $A_\ell > 0$. The minimiser of (11.92) is a polynomial P_ν derived from the Chebyshev polynomial $T_{\nu+1}$ (cf. Lemma 8.23) by*

$$\tau P_\nu(1 - \tau) = \eta(\nu) T_{\nu+1} \left(\tau - (1 - \tau) \cos \frac{\pi}{2\nu+2} \right) \quad (11.93)$$

$$\text{with } \eta(\nu) = \frac{1}{\nu + 1} \frac{\sin(\pi/(2\nu+2))}{1 + \cos(\pi/(2\nu+2))} \leq \frac{2(\sqrt{2}-1)}{(\nu+1)^2} \quad (\nu \geq 1).$$

P_ν is the product $P_\nu(1 - \tau) = \prod_{\mu=1}^{\nu} (1 - \omega_\mu \tau)$ with

$$\omega_\mu = \left(1 + \cos \frac{\pi}{2\nu+2} \right) / \left(\cos \frac{\pi}{2\nu+2} - \cos \frac{(2\mu+1)\pi}{2\nu+2} \right).$$

The expression (11.92) to be minimised takes the value

$$\|A_\ell P_\nu(I - A_\ell/\|A_\ell\|_2)\|_2 \leq \eta(\nu) \|A_\ell\|_2.$$

Proof. (i) Evaluation of $T_{\nu+1}(\dots)$ at $\tau = 0$ yields $T_{\nu+1}(-\cos \frac{\pi}{2\nu+2})$. Note that $-\cos \frac{\pi}{2\nu+2} = \cos(\pi - \frac{\pi}{2\nu+2}) = \cos(\frac{2\nu+1}{2\nu+2}\pi)$ and therefore

$$T_{\nu+1}(-\cos \frac{\pi}{2\nu+2}) = \cos((\nu+1)\frac{2\nu+1}{2\nu+2}\pi) = \cos((\nu+\frac{1}{2})\pi) = 0.$$

This justifies the factor τ on the left-hand side of (11.93).

(ii) The factor $\eta(\nu)$ is chosen such that $P_\nu(1) = \frac{d}{d\tau} P_\nu(1 - \tau)|_{\tau=0} = 1$ ensures the side condition.

(iii) Since the right-hand side in (11.93) takes the *equi-oscillating* values $\pm\eta(\nu)$ in $[0, 1]$, it is the minimiser of (11.92). \square

We add some comments to the results of Theorem 11.63.

(i) The semi-iterative smoothing achieves an order improvement. While the smoothing factor $\eta(\nu)$ of the stationary Richardson method behaves like $\mathcal{O}(\frac{1}{\nu+1})$, the order becomes $\mathcal{O}(1/(\nu+1)^2)$ in the semi-iterative case.

(ii) The application of the Chebyshev method requires knowledge of the interval $\sigma_M = [a, b]$ containing the spectrum of S_ℓ . Especially, the estimation of $b = 1 - \lambda_\ell / \|A_\ell\|_2$ with $\lambda_\ell = \lambda_{\min}(A_\ell)$ is of decisive importance. An overestimation of the upper bound $\Lambda_\ell = \|A_\ell\|_2$ in $\lambda_\ell I \leq A_\ell \leq \Lambda_\ell I$ is less sensitive (since $\Lambda_\ell/\lambda_\ell$ is the essential quantity). A different situation arises in Theorem 11.63, where we estimate the spectrum of A_ℓ simply by $0 \leq A_\ell \leq \Lambda_\ell I$, $\Lambda_\ell = \|A_\ell\|_2$, i.e., the lower bound λ_ℓ is trivially chosen as $a := 0$. The replacement of $0 \leq A_\ell$ by $\lambda_\ell I \leq A_\ell$ with $\lambda_\ell = \lambda_{\min}(A_\ell)$ would yield only an imperceptible improvement.

(iii) The statements from (ii) clarify the fact that the spectral condition number $\kappa(W_\ell^{-1}A_\ell)$ is not the essential quantity for smoothing.

(iv) The product representation $\prod(1 - \omega_\mu \tau)$ seems to disregard the warnings in §8.3.4 concerning instabilities. The contradiction is solved by the fact that the number ν of smoothing steps should be relatively small according to the discussion in §11.4.3. Choosing, e.g., $\nu \leq 4$, stability problems cannot arise.

For a general positive definite smoothing iteration with $S_\ell = I - W_\ell^{-1}A_\ell$, we obtain analogous results for minimising $\|W_\ell^{-1/2}A_\ell P_\nu(S_\ell)W_\ell^{-1/2}\|_2$, where the norms are chosen as for the approximation property (11.82). Corresponding to the smoothing property (11.78b), the minimisation of

$$\|W_\ell^{-1/2}A_\ell P_\nu(S_\ell)A_\ell^{1/2}\|_2 = \|Y^{1/2}P_\nu(I - Y)\|_2 \quad \text{with } Y := A_\ell^{1/2}W_\ell^{-1}A_\ell^{1/2}$$

is also of interest. The corresponding optimal polynomial can be found in Hackbusch [183, Proposition 6.2.35]. The bound $\mathcal{O}(1/\sqrt{\nu})$ in (11.78b) improves to $\mathcal{O}(1/(2\nu+1))$. The ADI parameters (cf. §8.5.3 and Hackbusch [183, §3.3.4 and Lemma 6.2.36]) have also to be chosen differently for optimising the smoothing effect.

The conjugate gradient method is only conditionally applicable. The standard CG method minimises $\|P_\nu(S_\ell)e_\ell\|_{A_\ell} = \|A_\ell^{1/2}P_\nu(S_\ell)e_\ell\|_2$, where e_ℓ is the error before smoothing and P_ν the corresponding optimal polynomial (cf. Proposition 10.11). However, since not the energy norm but the residual $\|A_\ell P_\nu(S_\ell)e_\ell\|_2$ has to be minimised, the method of the conjugate residuals (cf. §10.3) or the conjugate gradient method for the ‘squared’ equation $A_\ell^H A_\ell x_\ell = A_\ell^H b_\ell$ is better suited. These remarks apply to the pre-smoothing part only. The conjugate gradient methods do not seem to make much sense for the post-smoother. In any case, an nonsymmetric multigrid iteration results. See also Bank–Douglas [27].

The smoothing property of conjugate gradient methods has also been mentioned by Il’ in [226].

11.8.2 Damped Coarse-Grid Corrections

The treatment of nonlinear equations suggests damping the coarse-grid correction as known from gradient methods, in order to obtain a descent method (cf. Hackbusch–Reusken [204]). It turns out that in the linear case, it is also possible to improve convergence. In particular, the V-cycle convergence can be accelerated (cf. Reusken [321], Braess [62]). The optimally damped coarse-grid correction step reads

$$x_\ell^{\text{new}} := x_\ell - \lambda p_\ell \quad \text{with } \lambda := \frac{\langle d_\ell, p_\ell \rangle_\ell}{\langle d_\ell, A_\ell p_\ell \rangle_\ell}, \quad \begin{cases} d_\ell := A_\ell x_\ell - b_\ell, \\ p_\ell := p \tilde{e}_{\ell-1}, \end{cases} \quad (11.94)$$

where $\tilde{e}_{\ell-1}$ is the approximation of the solution of the coarse-grid equation $A_{\ell-1} e_{\ell-1} = d_{\ell-1} := r d_\ell$.

Exercise 11.64. Let $A_\ell > 0$ and $\tilde{e}_{\ell-1}$ as above. Prove that

- (a) $\lambda = 1$ is optimal for the two-grid method.
- (b) If $r = p^*$ and $A_{\ell-1} = r A_\ell p$, λ in (11.94) can be written in the form

$$\lambda = \langle d_{\ell-1}, \tilde{e}_{\ell-1} \rangle_{\ell-1} / \langle A_{\ell-1} \tilde{e}_{\ell-1}, \tilde{e}_{\ell-1} \rangle_{\ell-1}.$$

Another possibility is the damping of the complete multigrid iteration. In the symmetric case, $\check{M}_\ell^{\text{MGM}} \geq 0$ holds (cf. (11.89b)) and implies that $\sigma(M_\ell^{\text{MGM}}) \subset [0, \rho(M_\ell^{\text{MGM}})]$. Extrapolation with $\Theta := 2/(2 - \rho(M_\ell^{\text{MGM}})) \approx 1 + \frac{1}{2}\rho(M_\ell^{\text{MGM}})$ leads to the nearly halved convergence rate $\rho(M_\ell^{\text{MGM}})/(2 - \rho(M_\ell^{\text{MGM}}))$.

11.8.3 Multigrid as Basic Iteration of the CG Method

As shown in §11.7.1, the multigrid method for a positive definite matrix A_ℓ can be designed as a positive definite iteration. The convergence statement $\sigma(M_\ell^{\text{MGM}}) \subset [0, \rho_\ell]$ with $\rho_\ell := \rho(M_\ell^{\text{MGM}}) < 1$ corresponds to the inequalities

$$\gamma W_\ell^{\text{MGM}} \leq A_\ell \leq W_\ell^{\text{MGM}} \quad \text{with } \gamma := 1 - \rho_\ell$$

for the matrix W_ℓ^{MGM} of the third normal form of the multigrid iteration (cf. Theorem 6.10). Applying the CG method to Φ_ℓ^{MGM} , after m steps we obtain an improvement by $2[(\sqrt{\kappa} - 1)/(\sqrt{\kappa} + 1)]^m$, where κ is the condition $\kappa = \frac{\Gamma}{\gamma} = 1/(1 - \rho_\ell)$. A simple rewriting yields

$$2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^m = 2\rho_\ell^m / \left(1 + \sqrt{1 - \rho_\ell} \right)^{2m} \approx 2 \left(\frac{\rho_\ell}{4} + \mathcal{O}((\rho_\ell)^2) \right)^m.$$

Since ρ_ℓ may be assumed to be small (cf. §11.4.2), the convergence rate ρ_ℓ of the multigrid method can be accelerated by using the conjugate gradient method to $\rho_\ell/4$ (cf. Braess [61], Kettler [235]).

However, the use of the conjugate gradient method is of practical interest only if the multigrid convergence rate is relatively unfavourable (e.g., $\rho_\ell > 0.4$). The reason for this are considerations in §11.5.4. In the case of a good convergence rate, the nested iteration (11.44) requires only very few multigrid steps per level. For fast multigrid methods, the iteration number $m = 1$ has proved in §11.5.6 to be sufficient. For this value, the gradient and conjugate gradient methods still coincide. Only for $m \geq 2$ does the conjugate gradient method deserve attention.

11.9 Further Comments

11.9.1 Multigrid Method of the Second Kind

Discretisation of Fredholm's integral equations of the second kind leads to fixed-point equations of the form $x_\ell = K_\ell x_\ell + b_\ell$, i.e.,

$$A_\ell x_\ell = b_\ell \quad \text{with } A_\ell = I - K_\ell.$$

Here, K_ℓ does not characterise a difference operator as in the case of discretised differential equations, but the discrete counterpart of an integral operator $(\mathcal{K}u)(x) = \int_D k(x, y)dy$. Therefore, the *Picard iteration* $x_\ell^{m+1} := K_\ell x_\ell^m + b_\ell$ has a substantially better smoothing property. This implies that the multigrid method (with $\gamma = 2$) has a convergence rate $\rho_\ell = \mathcal{O}(h_\ell^\kappa)$ with a positive exponent κ . Hence, different from the situation considered before, convergence is better the larger the dimension of the problems is. Since the absolute value of ρ_ℓ may be of the size of 10^{-3} to 10^{-6} , the multigrid method is close to a direct solver. The work of the method is still proportional to the work of one Picard iteration.

The application is not restricted to discrete integral equations. In the example in §5.5.1, the equation $Ax = b$ was preconditioned by B , where B as well as A were discretisations of the differential equation. If both differential equations share the same principal part (i.e., if the terms of the highest order of differentiation coincide), the equation $A'x = b' := B^{-1}b$ with $A' := B^{-1}A$ leads to the fast multigrid convergence mentioned above. The application of the multigrid method of the second kind to $A'x = b'$ requires performing the Picard iteration $x^{m+1} = Kx^m + b' = x^m - B^{-1}(Ax - b)$. For example, B could be the five-point formula of the Poisson model problem, whereas A discretises the equation $-\Delta u + c_1 u_x + c_2 u_y + cu = f$.

An exact description and analysis of the multigrid method of the second kind as well as many examples of application can be found in Hackbusch [183, §16], [191], [184], [177], [179], [188].

Since the discrete integral operator K_ℓ is a fully populated matrix, the naive use of the Picard iteration leads to squared complexity. To obtain almost linear complexity the technique of hierarchical matrices can be applied (cf. Appendix D).

11.9.2 Robust Methods

To make our presentation brief, other smoothers than Richardson and chequer-board Gauss–Seidel are mentioned only marginally. In applications to more complicated systems, the problem of robustness arises. Are the convergence rates known for the Poisson model problem uniformly valid in a larger class of problems? The simplest case is an equation $A(\varepsilon)x = b$ depending on one parameter $\varepsilon \in (0, \infty)$. If the convergence rates not only are of the form $\rho(\varepsilon) \leq 1 - C(\varepsilon)h^\tau$ but hold with a constant $C(\varepsilon) = C$ uniformly in $\varepsilon \in (0, \infty)$: $\rho(\varepsilon) \leq 1 - Ch^\tau$, then this iteration is called *robust* with respect to the class of problems. Robust multigrid methods have to satisfy $\rho(\varepsilon) \leq \zeta < 1$ (ζh_ℓ - and ε -independent; cf. Hackbusch [183, §10]).

Good experiences concerning robustness—at least for two spatial dimensions—are observed for ILU smoothers as introduced by Wesseling [394], [393] (cf. Kettler [235], Wittum [401]). Robustness holds for CG methods applied to the modified ILU iteration ($\omega = -1$), as well as for multigrid methods using point- or blockwise ILU iterations as smoother (then with $\omega = 0$ or even $\omega = 1$; cf. Wittum [403], Kettler [235], Oertel–Stüben [295]).

Another approach is the frequency decomposition multigrid method (cf. Hackbusch [186, 190]) which uses not only one but several coarse-grid corrections with different coarse-grid equations. The prolongations from the different coarse grids into the fine grid are constructed in such a way that the corrections cover different frequency intervals.

Constructing the coarse-grid equation at level $\ell - 1$ requires more data than given by the system $A_\ell x_\ell = b_\ell$ for $\ell = \ell_{\max}$. This fact may lead to difficulties when the multigrid iteration is wanted as a black-box solver. Therefore, it is remarkable that there are variants, the so-called *algebraic multigrid methods*, in which the coarse-grid matrix $A_{\ell-1}$ is only constructed by the entries of the matrix A_ℓ (cf. Stüben [359, 360], MacLachlan–Oosterlee [268], Xu–Zikatnov [410]).

11.9.3 History of the Multigrid Method

The first two-grid method was described by Brakhage [69] in 1960. More precisely, it was a two-grid method of the second kind because it was applied to problems mentioned in §11.9.1. In 1961, Fedorenko [131] described a two- and in 1964 a multigrid method for the Poisson model problem (cf. Fedorenko [132]). In 1966, Bakhvalov [23] proved the typical convergence properties for a more complicated situation. Additional early publications were due to Astrachancev [8] (1971), Hackbusch [176] (1976), Bank–Dupont (a report from 1977 was split into [28, 29]), Brandt [77] (1977), and Nicolaides [290] (1977). Further details concerning these and other papers by Frederickson, Wesseling, Hemker, and Braess are mentioned in Hackbusch [183, §2.6.5]. An extensive multigrid bibliography up to 1987 can be found in the proceedings [278].

The progress of multigrid algorithms and theory can be traced in the proceedings of the European Multigrid Conferences: [205, Cologne 1981], [206, Cologne 1985], [207, Bonn 1990], [216, Amsterdam 1993], [210, Stuttgart 1996], [108, Gent 1999]. The proceedings of the later EMG conferences in Hohenwart (2002), Scheveningen (2005), Bad Herrenalb (2008), Ischia (2010), Schwetzingen (2012), and Leuven (2014) can be found in special issues, e.g., of the journal *Comput. Vis. Sci.* Further proceedings in this field are [68, 182, 185, 209, 306].

11.9.4 Frequency Filtering Decompositions

An essential characteristic of the multigrid method, besides the use of a coarser grid, is the product form $\Phi_\ell^{\text{CGC}} \circ S_\ell^\nu$ of two iterations which are active in different frequency intervals. Although many methods can be used for smoothing, the question remains as to whether there exists an alternative to $\Phi_\ell^{\text{CGC}} \circ S_\ell^\nu$. It would be desirable to have a method filtering out the coarse frequencies and needing no hierarchy of grids. Such a method is proposed by Wittum [404, 405] and is based on a sequence of partial steps Φ_ν reducing certain frequency intervals.

First, we describe the standard *blockwise* ILU decomposition. Suppose that A has the block-tridiagonal structure

$$A = A^H = \text{blocktridiag}\{L_i, D_i, L_i^H : i = 1, \dots, N-1\} \quad (11.95a)$$

(cf. (1.8), (A.9)), which, e.g., holds for five- or nine-point formulae. As in (1.2), $N-1 = h^{-1} - 1$ is the number of inner grid points per row. The exact LU decomposition is $A = LD^{-1}L^H$ with

$$L := \text{blocktridiag}\{L_i, T_i, 0\}, \quad D := \text{blockdiag}\{T_i\}, \quad (11.95b)$$

$$T_1 := D_1, \quad T_i := D_i - L_i T_{i-1}^{-1} L_i^H \quad (2 \leq i \leq N-1). \quad (11.95c)$$

Even if the blocks D_i are tridiagonal (cf. (1.8)), the matrices T_i are not sparse. The usual block-ILU decomposition is obtained from (11.95c) by replacing the full inverse of T_i^{-1} with the tridiagonal part $\text{tridiag}\{T_{i-1}^{-1}\}$ of the exact inverse.

Another approach goes back to Axelsson–Polman [15]. Let $t^{(1)}$ and $t^{(2)}$ be two *test vectors*. The matrices T_i are defined in the next lemma.

Lemma 11.65. *Assume that $t^{(1)}, t^{(2)} \in \mathbb{R}^{N-1}$ satisfy*

$$\det\left((t_{i+j}^{(k)})_{j=0,1}^{k=1,2}\right) \neq 0 \quad \text{for all } 1 \leq i \leq N-2.$$

The vectors $c^{(1)}, c^{(2)} \in \mathbb{R}^{N-1}$ may be arbitrary. Then there is a unique symmetric tridiagonal matrix T satisfying the equations $Tt^{(k)} = c^{(k)}$ for $k = 1, 2$.

Hence, we can uniquely define symmetric tridiagonal matrices T_i by

$$T_1 := D_1, \quad T_i t^{(k)} = (D_i - L_i T_{i-1}^{-1} L_i^H) t^{(k)} \quad (k = 1, 2) \quad (11.96)$$

for $i = 2, \dots, N-1$. The matrices T_i inserted into (11.95b) yield a new incomplete blockwise triangular decomposition $A = LD^{-1}L^H - C$. Definition (11.96) means

that T is exact with respect to the *test subspace* $\text{span}\{t^{(1)}, t^{(2)}\}$. The corresponding iteration is

$$\varPhi(x, b; t^{(1)}, t^{(2)}) := x - L^{-H} D L^{-1} (Ax - b)$$

with L, D defined in (11.95b) and (11.96).

Wittum [404, 405] proposes taking the sine functions e^ν in (11.25b) with different frequencies ν as test vectors:

$$\varPhi_\nu := \varPhi(\cdot, \cdot; e^\nu, e^{\nu+1}) \quad \text{with } \nu \in [1, N-2].$$

Choosing a factor $\alpha > 1$, which, e.g., may be chosen as $\alpha = 2$, a geometrical sequence of frequencies is selected (for [...] see Footnote 5 on page 297):

$$\nu_1 := 1, \quad \nu_{i+1} := \max\{\nu_i + 2, \lfloor \alpha \nu_i \rfloor\} \quad \text{as long as } \nu_{i+1} \leq N-2, \quad (11.97)$$

Let k be the number of frequencies selected in (11.97). Obviously, this number is equal to $k = \mathcal{O}(\log N) = \mathcal{O}(|\log h|) = \mathcal{O}(\log n)$. The iteration of the frequency filtering decomposition is defined by the product:

$$\varPhi_\alpha^{\text{ffd}} := \varPhi_{\nu_k} \circ \dots \circ \varPhi_{\nu_2} \circ \varPhi_{\nu_1} \quad (\alpha > 1 \text{ with } \nu_i \text{ in (11.97)}).$$

The work of one iteration $\varPhi_\alpha^{\text{ffd}}$ amounts to $\mathcal{O}(n \log n)$. The numerical results (cf. Wittum [404, 405]) demonstrate the very fast convergence of this iteration. Its efficacy can even exceed that of the standard multigrid methods.

The convergence is analysed for the case of a nine-point formula $A > 0$ with constant coefficients $D_i = D_{i+1}$, $L_i = L_{i+1} = L_i^H$ (cf. Wittum [405]). The first step of the proof concerns the monotone convergence of \varPhi_ν with respect to the energy norm for all ν . However, the more characteristic step is a *neighbourhood property*. According to its definition, \varPhi_ν eliminates error components in $\text{span}\{e^\nu, e^{\nu+1}\}$. It is essential that \varPhi_ν yields a uniform and h -independent contraction number for all frequencies in the interval $\nu \leq \mu \leq \alpha\nu$, i.e., that \varPhi_ν also acts efficiently in a certain neighbourhood of the *gauge frequency* ν .

The idea of frequency filtering decompositions can also be generalised to nonsymmetric or even nonlinear problems (cf. Wittum [405], Wagner [385, 386]). See also Weiler–Wittum [390], Wagner–Wittum [387], and Buzdin–Wittum [91].

Table 11.10 shows the iteration error $\|e^m\|_2 = \|x^m - x\|_2$ of the frequency filtering decomposition method $\varPhi_\alpha^{\text{ffd}}$ for $\alpha = 2$ applied to the Poisson model problem. The number k of partial steps ranges from 3 to 6. After 2 to 3 steps, machine precision is reached. One observes that with decreasing h , the convergence speed is bounded from above and therefore h -independently bounded.

m	$h = 1/8, k = 3$	$h = 1/16, k = 4$	$h = 1/32, k = 5$	$h = 1/64, k = 6$
	$\ e^m\ _2 \quad \rho_{m+1,m}$	$\ e^m\ _2 \quad \rho_{m+1,m}$	$\ e^m\ _2 \quad \rho_{m+1,m}$	$\ e^m\ _2 \quad \rho_{m+1,m}$
0	6.3_{10}^{-01}	7.0_{10}^{-01}	7.4_{10}^{-01}	7.6_{10}^{-01}
1	$2.6_{10}^{-07} \quad 4.1_{10}^{-7}$	$1.8_{10}^{-06} \quad 2.6_{10}^{-6}$	$1.4_{10}^{-05} \quad 1.9_{10}^{-5}$	$5.6_{10}^{-05} \quad 7.3_{10}^{-5}$
2	$1.0_{10}^{-12} \quad 3.9_{10}^{-6}$	$2.8_{10}^{-06} \quad 1.5_{10}^{-5}$	$1.5_{10}^{-09} \quad 1.0_{10}^{-5}$	$2.2_{10}^{-08} \quad 3.9_{10}^{-4}$
3	$4.1_{10}^{-13} \quad (4.0_{10}^{-1})$	$6.7_{10}^{-13} \quad (2.3_{10}^{-2})$	$1.2_{10}^{-12} \quad 8.2_{10}^{-4}$	$9.3_{10}^{-12} \quad 4.1_{10}^{-4}$

Table 11.10 Iteration of the frequency filtering decomposition for the Poisson model problem.

11.9.5 Nonlinear Systems

Although this monograph is devoted to systems of linear equations, the solution of nonlinear systems is of great importance. There are two principle approaches. In §11.9.5.1 we consider the Newton method, while in §11.9.5.2 proper nonlinear iterations are described.

The nonlinear system is of the form⁷

$$\mathcal{A}(x) = 0, \quad (11.98)$$

where the function $\mathcal{A} : \mathbb{K}^I \rightarrow \mathbb{K}^I$ is assumed to be continuously differentiable. We denote the derivative by

$$A(x) := \frac{d}{du} \mathcal{A}(x) \in \mathbb{K}^{I \times I}.$$

Let $x^* \in \mathbb{K}^I$ be the solution of (11.98) and define

$$A := A(x^*). \quad (11.99)$$

We require A to be regular. Then x^* is the unique solution in a neighbourhood \mathcal{X} of x^* and $A(x)$ is regular for all $x \in \mathcal{X}$. If the problem (11.98) is derived by discretising a nonlinear partial differential equation, we expect the same sparse structure of the matrices $A(x)$ as usual.

11.9.5.1 Newton's Method

The Newton method is the standard technique to transfer the solution of a nonlinear system into a sequence of linear problems. Starting with $x^0 \in \mathcal{X}$, the exact Newton method yields the sequence

$$x^{m+1} := x^m - A(x^m)^{-1} \mathcal{A}(x^m). \quad (11.100)$$

If the neighbourhood \mathcal{X} is small enough, the described sequence converges quadratically to x^* (cf. Quarteroni–Sacco–Saleri [314, §7.1]). Having in mind large-scale problems, the linear system

$$A(x^m) \delta = \mathcal{A}(x^m)$$

for the correction $\delta = x^m - x^{m+1}$ should not be computed directly. Instead any of the linear iterations described in this book can be applied to solve for δ .

Here the following comments apply:

- The derivative $A(x^m)$ has to be computed either analytically or by numerical differentiation. Since this may be costly, often $A(x^m)$ is replaced with an

⁷ In the nonlinear case, without loss of generality, the right-hand side can be defined by zero.

approximation. For instance, only $A(x^0)$ is computed and the later $A(x^m)$ are replaced by $A(x^0)$.

- If the iteration for the linear problem requires a larger amount of work for initialisation, this cost is required for each step of the Newton method. This is another reason for replacing $A(x^m)$ by a fixed matrix \tilde{A} .
- If, as above, $A(x^m)$ is replaced by some \tilde{A} , quadratic convergence is lost and the convergence of

$$x^{m+1} := x^m - \tilde{A}^{-1} \mathcal{A}(x^m)$$

depends on $A(x^m) - \tilde{A}$.

- In the case of the true matrix $A(x^m)$, the stopping criterion for the iteration applied to $A(x^m)\delta = \mathcal{A}(x^m)$ should produce approximations for δ_m for δ such that the error $\delta_m - \delta$ is comparable with the error of $x^{m+1} - x^*$. A too accurate solution of δ in the beginning does not pay, whereas a too rough approximation for later m prevents quadratic convergence.
- Since $A(x)$ is continuous and regular for $x \in \mathcal{X}$, the matrices of the family

$$\{A(x^m) : m \in \mathbb{N}_0\}$$

are spectrally equivalent. Therefore, in principle, the same preconditioner can be used for all linear systems that arise.

The usual convergence behaviour of (11.100) shows two phases. In a pre-asymptotic first phase only linear convergence is observed (say for $0 \leq m < m_0$). Later, for $m \geq m_0$, proper quadratic convergence occurs and only a few additional steps are needed. Above, the neighbourhood \mathcal{X} is chosen so that iteration (11.100) converges. Proper quadratic behaviour requires iterates in an even smaller neighbourhood $\mathcal{X}_{\text{quad}}$. It would be desirable to find a starting value in $\mathcal{X}_{\text{quad}}$ instead of \mathcal{X} .

A good strategy for this purpose is the (nonlinear) nested iteration. This requires defining nonlinear systems at all discretisation levels $\ell = 0, \dots, \ell_{\max}$, where the system at level ℓ_{\max} coincides with the original system (11.98):

$$\mathcal{A}_\ell(x_\ell) = b_\ell \quad (0 \leq \ell \leq \ell_{\max}), \quad (11.101)$$

where $b_{\ell_{\max}} = 0$. The nonlinear nested iteration takes the following form:

```

 $\tilde{x}_0 :=$  somehow computed approximation of  $\mathcal{A}_0(x^*) = 0$ ;
for  $\ell := 1$  to  $\ell_{\max}$  do
  begin  $\tilde{x}_\ell := \tilde{p}\tilde{x}_{\ell-1}$ ;  $\tilde{b}_{\ell-1} := \mathcal{A}_{\ell-1}(\tilde{x}_{\ell-1})$ ;
    apply an iterative solver starting with  $\tilde{x}_\ell$  delivering a new value  $\tilde{x}_\ell$ 
  end;

```

(11.102)

The data \tilde{b}_ℓ ($0 \leq \ell \leq \ell_{\max} - 1$) will be used later. Although \tilde{x}_ℓ is only an approximation, it is the exact solution of $\mathcal{A}_\ell(\tilde{x}_\ell) = \tilde{b}_\ell$.

11.9.5.2 Nonlinear Iterations

\varPhi in Definition 2.1 can be generalised to the nonlinear problem (11.98) by a nonlinear mapping

$$x^{m+1} = \varPhi(x^m, \mathcal{A}).$$

For instance, the nonlinear analogue of the Richardson iteration (3.4) is

$$x^{m+1} = \varPhi_{\text{nonl}}^{\text{Rich}}(x^m, \mathcal{A}) := x^m - \Theta \mathcal{A}(x^m).$$

Rewriting x^m by $x^* + e^m$ and assuming a small error e^m , we obtain the Taylor expansion

$$\mathcal{A}(x^m) = \mathcal{A}(x^*) + Ae^m + o(e^m) = Ae^m + o(e^m)$$

with A in (11.99) and therefore

$$x^{m+1} = x^m - \Theta Ae^m + o(e^m) \approx x^m - \Theta (Ax^m - b) = \varPhi^{\text{Rich}}(x^m, b, A)$$

with $b := Ax^*$. This proves that

$$\varPhi_{\text{nonl}}^{\text{Rich}}(x, \mathcal{A}) \rightarrow \varPhi^{\text{Rich}}(x, b, A) \quad \text{as } x \rightarrow x^*.$$

The nonlinear analogue of the Gauss–Seidel method replaces each step in (3.9) by solving the i -th equation in the system $\mathcal{A}(x) = 0$ with respect to the $x[i]$. The scalar nonlinear equations that arise can be solved, e.g., by Newton’s method. In the same way, the nonlinear Jacobi iteration and the nonlinear SOR can be performed (cf. Törnig [365, §§8.2–8.4]).

More involved algebraic linear iterations as the ILU iteration are hard to transfer into a nonlinear counterpart since it requires the (incomplete) decomposition of the derivative A .

The linear iteration $\varPhi_{\text{lin}}(x, b, A) = x - N(Ax - b)$ has the obvious nonlinear counterpart $\varPhi_{\text{nonl}}(x, \mathcal{A}) := x - N\mathcal{A}(x)$. In all these cases, the asymptotic convergence speed of the nonlinear iteration \varPhi_{nonl} coincides with the convergence speed of the linear iteration \varPhi_{lin} applied to the linearised system $Ax - b$ with A in (11.99).

11.9.5.3 Nonlinear Two- and Multigrid Iteration

The multigrid iteration has a very natural generalisation to nonlinear systems. The underlying reason is that the method requires not the derivative $A(x) = d\mathcal{A}(x)/du$ but only a directional derivative.

Instead of (11.98), we consider the family (11.101) of systems at all levels ℓ . We start with the two-grid iteration involving the levels ℓ and $\ell - 1$. We assume that the nested iteration is already used for levels below ℓ , so that a good starting value for x_ℓ , the approximate solution $\tilde{x}_{\ell-1}$ at level $\ell - 1$ and its defect $\tilde{b}_{\ell-1}$ are known.

The real number s used in lines 3 and 5 will be explained below.

```

function  $\Phi_\ell^{\text{NTGM}}(x_\ell, b_\ell);$  {solution of  $\mathcal{A}_\ell(x_\ell) = b_\ell$  desired}
begin  $x_\ell := \mathcal{S}_\ell^{\nu_1}(x_\ell, b_\ell);$  {pre-smoothing}
     $d_{\ell-1} := r(\mathcal{A}_\ell(x_\ell) - b_\ell); d_{\ell-1} := \tilde{b}_{\ell-1} + s \cdot d_{\ell-1};$ 
     $\xi_{\ell-1} := \mathcal{A}_{\ell-1}^{-1}(d_{\ell-1});$  {coarse-grid solve}
     $x_\ell := x_\ell - p(\xi_{\ell-1} - \tilde{x}_{\ell-1})/s;$  {coarse-grid correction}
     $\Phi_\ell^{\text{NTGM}} := \mathcal{S}_\ell^{\nu_2}(x_\ell, b_\ell);$  {post-smoothing}
end;

```

The pre- and post-smoothing iterations \mathcal{S}_ℓ may, e.g., be the nonlinear Richardson or Jacobi iteration. \mathcal{S}_ℓ^ν denotes the ν -fold application.

Let x_ℓ^* be the solution of $\mathcal{A}_\ell(x_\ell^*) = 0$. We recall the neighbourhood \mathcal{X}_ℓ of x_ℓ^* , in which x_ℓ^* is the unique solution. Hence, the function $\mathcal{A}_\ell : \mathcal{X}_\ell \rightarrow \mathcal{Y}_\ell := \mathcal{A}_\ell(\mathcal{X}_\ell)$ is bijective. This allows us to define the inverse function \mathcal{A}_ℓ^{-1} on \mathcal{Y}_ℓ . The function Φ_ℓ^{NTGM} uses $\mathcal{A}_{\ell-1}^{-1}$ for solving a coarse-grid equation $\mathcal{A}_{\ell-1}\xi_{\ell-1} = d_{\ell-1}$. This requires that $d_{\ell-1} \in \mathcal{Y}_{\ell-1}$. Since $\mathcal{Y}_{\ell-1}$ is a neighbourhood of zero, $d_{\ell-1}$ must be small enough. Since, by definition, $\tilde{x}_{\ell-1}$ is a good approximation of $x_{\ell-1}^*$, the defect $\tilde{b}_{\ell-1}$ is small enough. Choosing the number s small enough, $d_{\ell-1} = \tilde{b}_{\ell-1} - s \cdot d_{\ell-1}$ also belongs to $\mathcal{Y}_{\ell-1}$.

To understand the correction $x_\ell := x_\ell + p(\xi_{\ell-1} - \tilde{x}_{\ell-1})/s$, rewrite the bracket as

$$\xi_{\ell-1} - \tilde{x}_{\ell-1} = \mathcal{A}_{\ell-1}^{-1}(d_{\ell-1}) - \mathcal{A}_{\ell-1}^{-1}(\tilde{b}_{\ell-1}) \approx \left(\frac{d}{dy} \mathcal{A}_{\ell-1}^{-1} \right) (d_{\ell-1} - \tilde{b}_{\ell-1}).$$

The derivative of the inverse function $\mathcal{A}_{\ell-1}^{-1}(y)$ is

$$\left(\frac{d}{dx} \mathcal{A}_{\ell-1}(x) \right)^{-1} = A_{\ell-1}^{-1} \quad \text{for } x = \mathcal{A}_{\ell-1}^{-1}(y).$$

Together with $d_{\ell-1} - \tilde{b}_{\ell-1} = s \cdot d_{\ell-1}$, we obtain $\xi_{\ell-1} - \tilde{x}_{\ell-1} = s A_{\ell-1}^{-1} d_{\ell-1}$ and the correction step yields asymptotically $x_\ell - p(\xi_{\ell-1} - \tilde{x}_{\ell-1})/s \approx x_\ell - p A_{\ell-1}^{-1} d_{\ell-1}$ with the restricted defect $d_{\ell-1} = r(\mathcal{A}_\ell(x_\ell) - b_\ell)$. This is the same expression as in (11.21b-d) and proves that the nonlinear two-grid iteration has an asymptotic convergence speed which coincides with the convergence speed of the linear two-grid iteration applied to the linearised system.

The recursive application of Φ_ℓ^{NTGM} yields the nonlinear multigrid iteration. Note that the application of Φ_ℓ^{NTGM} is interwoven with the nonlinear nested iteration (11.102) in which the solver is the m -fold application of Φ_ℓ^{NTGM} . This implies that, when Φ_ℓ^{NTGM} is called, the quantities \tilde{x}_k and \tilde{b}_k are known for all lower levels $k < \ell$. In addition, we need a nonlinear function $\tilde{\Phi}_0(x_0, b_0)$ returning a good approximation of $\mathcal{A}_0^{-1}(b_0)$. This may be a Newton method. The number γ has the same meaning as in the linear case: $\gamma = 1$ is the V-cycle, $\gamma = 2$ is the W-cycle. The numbers $s = s(d_{\ell-1})$ play the same role as in the two-grid iteration. It can be chosen such that $s \cdot d_{\ell-1}$ is of the same size as $\tilde{b}_{\ell-1}$.

```

function  $\Phi_\ell^{\text{NMGM}}(x_\ell, b_\ell)$ ; {solution of  $\mathcal{A}_\ell(x_\ell) = b_\ell$  desired}
begin  $x_\ell := \mathcal{S}_\ell^{\nu_1}(x_\ell, b_\ell)$ ; {pre-smoothing}
     $d_{\ell-1} := r(\mathcal{A}_\ell(x_\ell) - b_\ell)$ ;  $d_{\ell-1} := \tilde{b}_{\ell-1} + s \cdot d_{\ell-1}$ ;
     $\xi_{\ell-1} := \tilde{x}_{\ell-1}$ ; for  $i := 1$  to  $\gamma$  do  $\xi_{\ell-1} := \Phi_{\ell-1}^{\text{NMGM}}(\xi_{\ell-1}, d_{\ell-1})$ ;
     $x_\ell := x_\ell - p(\xi_{\ell-1} - \tilde{x}_{\ell-1}) / s$ ; {coarse-grid correction}
     $\Phi_\ell^{\text{NMGM}} := \mathcal{S}_\ell^{\nu_2}(x_\ell, b_\ell)$ ; {post-smoothing}
end;

```

Again the nonlinear multigrid iteration has an asymptotic convergence speed which coincides with the convergence speed of the linear multigrid iteration applied to the linearised system. Details about the convergence proof can be found in [194, §9.5].

If one applies Φ_ℓ^{NMGM} to the linear problem $\mathcal{A}_\ell(x_\ell) = A_\ell x_\ell - b_\ell$, the auxiliary data $(\tilde{x}_{\ell-1}, \tilde{b}_{\ell-1})$ can be chosen as $(0, 0)$ and the algorithm coincides with the linear multigrid iteration.

There are different nonlinear multigrid versions using other reference data $(\tilde{x}_{\ell-1}, \tilde{b}_{\ell-1})$ and other factors s . A comparison with numerical examples is given in Hackbusch [189].

Chapter 12

Domain Decomposition and Subspace Methods

Abstract Domain decomposition is an umbrella term collecting various methods for solving discretised boundary value problems in a domain Ω by means of a decomposition of Ω . Often this approach is chosen to support parallel computing. After general remarks in Section 12.1, the algorithm using overlapping subdomains is described in Section 12.2. In the case of nonoverlapping subdomains, one needs more involved methods (cf. Section 12.3). In particular the so-called FETI method described in §12.3.2 is very popular. The Schur complement method in Section 12.4 gives rise to many variants of iterations. The more abstract view of domain decomposition methods replaces the subdomain by a subspace. Section 12.5 formulates the setting of subspace iterations. Here we distinguish between the additive and multiplicative subspace iteration as explained in the corresponding Sections 12.6 and 12.7. Illustrations follow in Section 12.8. Interestingly, multigrid iterations can also be considered as subspace iterations as analysed in Section 12.9.

12.1 Introduction

Various iterative methods can be classified as *subspace methods*. If the subspaces correspond to discretisations of boundary value problems, these methods are called *domain decomposition methods (DDM)*. Another name is *Schwarz iteration* since a prototype of this iteration is due to Hermann A. Schwarz [336] (1870). It took about hundred years that this class of algorithms attracted the interest of numerical analysts (cf. Babuška [21]).

In principle, the method can be applied to nonsymmetric or indefinite and even nonlinear problems, but the convergence analysis is usually restricted to the positive definite case.¹

¹ For a more general analysis see Cai–Widlund [92], Xu [408], and Dryja–Hackbusch [113].

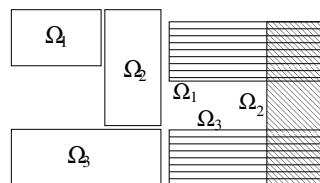


Fig. 12.1 Disjoint (left) and overlapping (right) subdomains.

Let the system of equations $Ax = b$ represent discretisation of a boundary value problem in the domain Ω (cf. §1.2). The naming characteristic of the domain decomposition method is a decomposition of the complete problem into smaller systems of equations corresponding to boundary value problems in subdomains $\Omega_\nu \subset \Omega$.

The choice of subdomains can be caused by different motivations.

1. The decomposition of a complicated domain may lead to subdomains of simpler type. In the past when fast methods as, e.g., the multigrid iteration had not yet been (sufficiently) known and only fast direct solvers for Poisson problems on rectangular domains were available (cf. Buneman [87]), one tried to decompose complex domains into disjoint rectangles (cf. left part of Fig. 12.1) or overlapping rectangles (cf. Fig. 12.1, right).
2. A second reason for a domain decomposition might be that the original problem is a composed problem coupling² subproblems of different physical nature. For instance, the coefficient σ in the differential operator $L = -\operatorname{div}(\sigma \operatorname{grad})$ is often a material constant. A combination of different materials lead to a discontinuous and piecewise constant σ . Choosing the subdomains corresponding to identical materials, we obtain subproblems with constant coefficients σ , i.e., $L = -\sigma \Delta$. This shows that the subproblems may behave more regular than the overall problem. Note that this argument only holds for nonoverlapping domain decompositions.
3. A third argument for a decomposition is the use of parallel computers³ (cf. Smith–Bjørstad–Gropp [343, §3.6.3]). Decomposing the entire problem into many separate subproblems, we obtain a number of subtasks which can be solved in parallel. In this case, the subproblems should be similar in size because of the load-balancing of processors.

It must be emphasised that the solution of subproblems can never solve the complete problem, but only represents a partial step. The algorithm has also to establish the correct coupling of subproblems. If parallelisation is the main argument, the major part of computational work should be consumed by the solution of subproblems. The characteristic feature of the multigrid method, the coarse-grid correction, will reappear in the domain decomposition method (cf. §12.8.2).

The *capacitance matrix method* and the *method of fictitious domains* also belong to the class of domain decomposition methods (cf. page 334).

In the course of the development of the domain decomposition method, the term ‘subdomain’ has been generalised to ‘subspace’, in particular, to a subspace of the Galerkin method. If the subspace is spanned by those finite element functions that

² In general we assume that a discretisation of the global problem is given, which may be subdivided into part for defining the domain decomposition method. However, even the global (undiscretised) problem may be divided into subproblems which are discretely coupled in very weak form (see, e.g., the mortar method in [100, §7] or Quarteroni–Valli [315, §2.5.1]).

³ The use of parallel computers and the construction of the iterative method are in principle independent. For instance, there are parallel implementations of the multigrid method (cf. Reiter et al. [320]).

differ from zero only in a subdomain $\Omega' \subset \Omega$, the terms ‘subdomain’ and ‘subspace’ coincide. Other subspaces, however, can also be constructed and deserve practical interest.

The progress in domain decomposition methods is documented in the proceedings of the regular DDM conferences: [150, Paris 1987], [93, Los Angeles 1988], [94, Houston 1989], [151, Moscow 1990], [236, Norfolk 1991], [313, Como 1992], [237, Penn State 1993], [153, Beijing 1995], [50, Bergen 1996], [271, Boulder 1997], [254, Greenwich 1998], [95, Chiba 1999], [106, Lyon 2000], [217, Cocoyoc 2002], [239, Berlin 2003], [399, New York 2005], [258, St. Wolfgang/Strobl 2006], [43, Jerusalem 2008], [225, Zhanjiajie 2009], [31, San Diego 2011], [122, Rennes 2012], [109, Lugano 2013].

Monographs on domain decomposition are by Mathew [276], Quarteroni–Valli [315], Smith–Bjørstad–Gropp [343], and Toselli–Widlund [366].

12.2 Overlapping Subdomains

The ideal domain decomposition method is characterised by the exact solution of the subproblems. The following techniques will be of this kind. For instance, we may assume that all matrices corresponding to the subproblems are LU factorised (or use a Cholesky decomposition in the symmetric case).

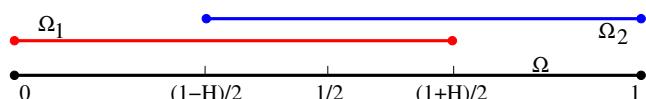
The practical application, however, will often involve an iterative solution of the subdomains. This leads to a composed iteration (cf. §5.5).

12.2.1 Introductory Example

The overlapping domain decomposition corresponds to the original approach of Schwarz [336] and can be illustrated by a one-dimensional example:

$$-u'' = f \text{ in } \Omega = (0, 1), \quad u(0) = u(1) = 0. \quad (12.1)$$

For simplicity, we apply the approach to the undiscretised problem. The domain Ω is decomposed into the depicted subdomains (subintervals):



$$\Omega_1 = (0, \frac{1+H}{2}) \text{ and } \Omega_2 = (\frac{1-H}{2}, 1) \quad \text{for some } 0 < H < 1.$$

The overlap $\Omega_1 \cap \Omega_2$ has the length H .

The following iteration determines functions u_{Ω_ν} defined on Ω_ν ($\nu = 1, 2$). Let $u_{\Omega_\nu}^0$ be the starting value (in fact only $u_{\Omega_2}^0$ is needed). The first partial step solves the boundary value problem in Ω_1 with the Dirichlet value at $\frac{1+H}{2}$ taken from $u_{\Omega_2}^0$:

$$u_{\Omega_1}^1 \text{ solution of } -u'' = f \text{ in } \Omega_1 \text{ with } u(0) = 0, u\left(\frac{1+H}{2}\right) = u_{\Omega_2}^0\left(\frac{1+H}{2}\right).$$

The second partial step solves for $u_{\Omega_2}^1$ with the left Dirichlet boundary value taken from $u_{\Omega_1}^1$:

$$u_{\Omega_2}^1 \text{ solution of } -u'' = f \text{ in } \Omega_2 \text{ with } u\left(\frac{1-H}{2}\right) = u_{\Omega_2}^0\left(\frac{1-H}{2}\right), u(1) = 0.$$

The iterates $u_{\Omega_\nu}^m$ should converge to $u^*|_{\Omega_\nu}$ which is the restriction of the solution u^* of (12.1) to the subdomain Ω_ν . To study the convergence behaviour, we consider problem (12.1) with $f = 0$. The solutions of $-u'' = 0$ are affine functions. The Dirichlet values in (12.1) yield the solution $u^* = 0$. We start with the initial value $u_{\Omega_2}^0(x) = 1 - x$. Since $u_{\Omega_2}^0\left(\frac{1+H}{2}\right) = \frac{1-H}{2}$, the next iterate is

$$u_{\Omega_1}^1(x) = \frac{1-H}{1+H}x.$$

Its value at $\frac{1-H}{2}$ determines

$$u_{\Omega_2}^1(x) = \left(\frac{1-H}{1+H}\right)^2 (1-x).$$

Obviously, the general solution is

$$u_{\Omega_1}^m(x) = \left(\frac{1-H}{1+H}\right)^{2m-1} x \quad \text{and} \quad u_{\Omega_2}^m(x) = \left(\frac{1-H}{1+H}\right)^{2m} (1-x).$$

Note that the iterate $u_{\Omega_\nu}^m$ is also the error because of $u^* = 0$. Hence, the convergence rate is

$$\eta = \left(\frac{1-H}{1+H}\right)^2 < 1.$$

Figure 12.2 illustrates the iteration.

The same approach can be applied to the discretised boundary value problem and yields the same convergence rate η . Note that $\eta = \left(\frac{1-H}{1+H}\right)^2 < 1$ is independent of the step size h . However, this holds only for a fixed value of H .

The minimal overlap is one step size h . In this case, the rate

$$\eta = \left(\frac{1-h}{1+h}\right)^2 = 1 - 4h + \mathcal{O}(h^2)$$

deteriorates with increasing dimension.

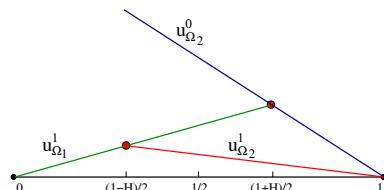


Fig. 12.2 Overlapping DDM.

12.2.2 Many Subdomains

This approach can be generalised to many subdomains. Let K be the number of subdomains, set $\delta := 1/K$ and choose some overlap size $0 < H < \delta$. Define

$$\Omega_\nu = ((\nu - 1)\delta - H, \nu\delta + H) \cap (0, 1) \quad \text{for } \nu = 1, \dots, K.$$

Then Ω_ν and $\Omega_{\nu+1}$ overlap in $(\nu\delta - H, \nu\delta + H)$. If we want to solve the subdomains problems exactly, the subdomains should be sufficiently small, which implies that there must be many subdomains; i.e., K should be large.

However, the convergence cannot be independent of the number K of subdomains since $K \rightarrow \infty$ implies $H \rightarrow 0$ and therefore a deteriorating rate $\eta \rightarrow 1$.

A two-dimensional example of an overlapping domain decomposition follows in §12.8.1.

12.3 Nonoverlapping Subdomains

12.3.1 Dirichlet–Neumann Method

Consider the decomposition of $\Omega = (0, 1)$ into $\Omega_1 = (0, \frac{1}{2})$ and $\Omega_2 = [\frac{1}{2}, 1]$. Now it is not enough to require identical Dirichlet values $u_{\Omega_1}(\frac{1}{2}) = u_2(\frac{1}{2})$ for the respective solutions of (12.1) in Ω_1 and Ω_2 . In addition, we have to ensure that the Neumann data coincide:

$$\frac{\partial u_{\Omega_1}}{\partial n_{\Omega_1}}\left(\frac{1}{2}\right) = -\frac{\partial u_{\Omega_2}}{\partial n_{\Omega_2}}\left(\frac{1}{2}\right)$$

(the opposite signs of the normal derivatives correspond to the fact that the normal directions with respect to Ω_1 and Ω_2 satisfy $n_{\Omega_1} = -n_{\Omega_2}$).

A possible iteration is the following with a constant $\vartheta \in (0, 1)$:

- $u_{\Omega_1}^{m+1}$: solution of (12.1) with $u(0) = 0$ and $u(\frac{1}{2}) = \vartheta u_{\Omega_2}^m(\frac{1}{2}) + (1 - \vartheta)u_{\Omega_1}^m(\frac{1}{2})$,
- $u_{\Omega_2}^{m+1}$: solution of (12.1) with $u(1) = 0$ and $\frac{\partial}{\partial n_{\Omega_2}}u(\frac{1}{2}) = -\frac{\partial}{\partial n_{\Omega_1}}u_{\Omega_1}^{m+1}(\frac{1}{2})$.

In this special case, the convergence analysis is easy since the solutions $u_{\Omega_\nu}^m$ are characterised by only one parameter:

$$u_{\Omega_1}^m(x) = \alpha_m x, \quad u_{\Omega_2}^m(x) = \beta_m(1 - x).$$

One verifies that $\alpha_{m+1} = \frac{1}{2}(\vartheta\beta_m + (1 - \vartheta)\alpha_m)$ and $\beta_{m+1} = -\alpha_{m+1}$, i.e.,

$$\begin{bmatrix} \alpha_{m+1} \\ \beta_{m+1} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 - \vartheta & \vartheta \\ -1 + \vartheta & -\vartheta \end{bmatrix} \begin{bmatrix} \alpha_m \\ \beta_m \end{bmatrix}.$$

Since the matrix has the eigenvalues 0 and $1 - 2\vartheta \in (-1, 1)$, the iteration converges to the exact solution $u^* = 0$ with the rate $\eta = |1 - 2\vartheta|$. For $\vartheta = 1/2$, the iterates for $m = 2$ are already exact.

Discretisation of (12.1) yields the matrix $A = h^{-2} \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & -1 & 2 & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{bmatrix} \in \mathbb{K}^{\{1, \dots, n\} \times \{1, \dots, n\}}$

$\mathbb{K}^{\{1, \dots, n\} \times \{1, \dots, n\}}$. Assume that n is odd. The index $q := (n+1)/2$ corresponds to the nodal point $x = 1/2$. We write $u^{I,m} \in \mathbb{K}^{\{1, \dots, q\}}$ instead of $u_{\Omega_1}^{m+1}$, and $u^{II,m} \in \mathbb{K}^{\{q, \dots, n\}}$ for $u_{\Omega_2}^{m+1}$.

The discrete problem for $u^{I,m+1}$ is the system $A^I u^{I,m+1} = b^I$ with

$$A^I = \begin{bmatrix} A|_{\{1, \dots, q-1\} \times \{1, \dots, q\}} \\ 0 & \dots & 0 & 1 \end{bmatrix} \in \mathbb{K}^{\{1, \dots, q\} \times \{1, \dots, q\}},$$

$$b^I = (\vartheta u_q^{II,m} + (1-\vartheta)u_q^{I,m}) e^q,$$

where e_q is the q -th unit vector. The second vector $u^{II,m+1}$ is the solution of $A^{II} u^{II,m+1} = b^{II}$ using the restriction $A|_{\{q, \dots, n\} \times \{q, \dots, n\}}$ modified in the first row:

$$A^{II} = h^{-2} \begin{bmatrix} 1 & -1 & \dots \\ -1 & 2 & \ddots \\ \vdots & \ddots & \ddots & -1 \\ -1 & 2 & & \end{bmatrix}.$$

The right-hand side is $b^{II} = h^{-2} (u_q^{I,m} - u_{q-1}^{I,m}) e_q$.

We recall Remark E.8b: the matrix A^{II} cannot be obtained from A without knowledge of the underlying partial differential problem. Therefore the corresponding method is not algebraic.

Because of the combination of a Dirichlet problem for $u_{\Omega_1}^{m+1}$ and a Neumann problem for $u_{\Omega_2}^{m+1}$, the iteration is called the Dirichlet–Neumann method.

12.3.2 Lagrange Multiplier Based Methods

The following approach is related to methods like in §12.3. It does not yield algebraic iterations since it requires the finite element formulation for the subdomains (cf. Remark E.8b).

For illustration, we assume that a domain $\Omega \subset \mathbb{R}^d$ is split into two nonoverlapping domains Ω_1 and Ω_2 ; i.e., $\overline{\Omega} = \overline{\Omega_1} \cup \overline{\Omega_2}$ and $\Omega_1 \cap \Omega_2 = \emptyset$. Let the boundary value problem be defined by a symmetric and coercive bilinear form $a(u, v) = \int_{\Omega} \dots dx : V \rightarrow V$ with $V = H_0^1(\Omega)$. Restricting the integrals to Ω_1 and Ω_2 , we obtain the bilinear forms $a_1(u, v) = \int_{\Omega_1} \dots dx : V_1 \times V_1 \rightarrow \mathbb{K}$ and $a_2 : V_2 \times V_2 \rightarrow \mathbb{K}$ (cf. §E.3). The spaces are defined by $V_i = \{v \in H^1(\Omega_i) : v|_{\partial\Omega \cap \partial\Omega_i} = 0\}$. Hence the functions in V_1 and V_2 can take arbitrary values on

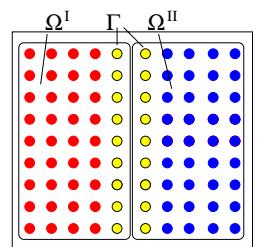


Fig. 12.3 Two subdomains and two copies of Γ .

the interior boundary $\Gamma := \partial\Omega_1 \cap \partial\Omega_2$. We recall that the minimisation of $J_i(u_i) := \frac{1}{2}a_i(u_i, u_i) - f_i(u_i)$ ($i = 1, 2$) yields the solutions of the differential equations $Lu_i = f_i$ in Ω_i with Dirichlet data $u_i = 0$ on $\partial\Omega \cap \partial\Omega_i$ and Neumann data $\frac{\partial}{\partial n}u_i = 0$ on Γ (cf. Remark E.3). The minimisation of

$$J_0(u_1, u_2) := J_1(u_1) + J_2(u_2)$$

over $u_1 \in V_1$ and $u_2 \in V_2$ is identical to minimising each $a_i(u_i, u_i) - f(u_i)$ separately. In general, the Dirichlet values $u_i|_\Gamma$ are different. Using $u_1|_\Gamma = u_2|_\Gamma$ as a side condition, we obtain the minimisation problem

$$\min \{J_0(u_1, u_2) : u_i \in V_i \text{ subject to } u_1|_\Gamma = u_2|_\Gamma\},$$

which is equivalent to the original problem $\min\{\frac{1}{2}a(u, u) - f(u) : u \in V\}$. The reason is that $u_1|_\Gamma = u_2|_\Gamma$ implies that there is a function $u \in V$ with $u_i = u|_{\Omega_i}$. The side condition will be coupled by a Lagrange parameter (function on Γ).

The discrete problem in Ω_1 is described by the minimisation of $J_1(u_1)$ over all $u_1 \in V_h^1$, where $V_h^1 \subset V_1$ is a finite element subspace containing functions $u_1 = \sum_{\alpha \in I^1} x_\alpha \phi_\alpha$ (cf. (E.6)). The components x_α of $x^1 \in \mathbb{K}^{I^1}$ correspond to nodal points in $\Omega_1 \cup \Gamma$. We distinguish $\alpha \in I_\Omega^1$ with nodal points in Ω_1 from $\alpha \in I_\Gamma^1$ with nodal points on Γ . Correspondingly, the vector $x^1 \in V_h^1$ can be decomposed into $x^1 = \begin{bmatrix} x_\Omega^1 \\ x_\Gamma^1 \end{bmatrix}$ with $x_\Omega^1 \in \mathbb{K}^{I_\Omega^1}$ and $x_\Gamma^1 \in \mathbb{K}^{I_\Gamma^1}$.

The solution of $\min J_1(u_1)$ over V_h^1 is equivalent to the system

$$A^1 x^1 = b^1, \quad \text{or in block form: } \begin{bmatrix} A_{\Omega\Omega}^1 & A_{\Omega\Gamma}^1 \\ A_{\Gamma\Omega}^1 & A_{\Gamma\Gamma}^1 \end{bmatrix} \begin{bmatrix} x_\Omega^1 \\ x_\Gamma^1 \end{bmatrix} = \begin{bmatrix} b_\Omega^1 \\ b_\Gamma^1 \end{bmatrix}.$$

An analogous statement holds for the second domain. Since $u_1|_\Gamma$ and $u_2|_\Gamma$ are treated independently, x_Γ^1 and x_Γ^2 contain different values at the Γ -nodal points (see the two copies of Γ in Fig. 12.3). We form the vectors $x := \begin{bmatrix} x^1 \\ x^2 \end{bmatrix}$, $b := \begin{bmatrix} b^1 \\ b^2 \end{bmatrix}$, and the matrix $A = \begin{bmatrix} A^1 & 0 \\ 0 & A^2 \end{bmatrix}$.

The side condition $x_\Gamma^1 = x_\Gamma^2$ can be written as $Mx = 0$, involving the matrix $M = [0 \ I \ 0 \ -I]$. The Lagrange parameter $\lambda \in \mathbb{K}^{I_\Gamma}$ is used to couple the minimisation of $J_0(u_1, u_2)$ with the side condition $Mx = 0$. The first variation yields the system

$$\begin{bmatrix} A & M^\top \\ M & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}.$$

The iterative solution of this (indefinite) system is the basis of the so-called FETI method (cf. Farhat–Roux [127], Brenner [78], Mathew [276, §4], and Toselli–Widlund [366, §6]). The interpretation of this method within the framework of the subspace method is discussed by Brenner [80].

12.4 Schur Complement Method

12.4.1 Nonoverlapping Domain Decomposition with Interior Boundary

Consider the two-dimensional Poisson model problem with the grid $\Omega_h \subset (0, 1)^2$ (cf. (1.3)). On the left-hand side of Figure 12.4 the domain is now decomposed into *three* parts: two proper subdomains $\Omega_h^1 := \Omega_h \cap (0, \frac{1}{2}) \times (0, 1)$ and $\Omega_h^2 := \Omega_h \cap (\frac{1}{2}, 1) \times (0, 1)$ and the interior boundary $\Gamma := \Omega_h \cap \{\frac{1}{2}\} \times (0, 1)$. In the right part of Figure 12.4 we have four proper subdomains $\Omega_h^1, \dots, \Omega_h^4$ and the interior boundary Γ which now contains a so-called *cross-point* at $(1/2, 1/2)$.

In the following the indices 1 to $k-1$ are associated with the subdomains while k belongs to Γ . The matrix $A \in \mathbb{K}^{\Omega_h \times \Omega_h}$ can be decomposed into $A_{\ell\ell} := A|_{\Omega_h^\ell \times \Omega_h^\ell}$, $A_{kk} := A|_{\Gamma \times \Gamma}$, $A_{\ell k} := A|_{\Omega_h^\ell \times \Gamma}$, and $A_{k\ell} := A|_{\Gamma \times \Omega_h^\ell}$ for $1 \leq \ell \leq k-1$. The interior boundary is chosen such that there is no interaction between Ω_h^ℓ and Ω_h^k for $\ell \neq k$; i.e., $A_{ij} = 0$ holds for all $i \in \Omega_h^\ell$ and $j \in \Omega_h^k$. Hence, A has the block structure

$$A = \begin{bmatrix} A_{11} & \mathbf{O} & A_{1,k} \\ \ddots & \ddots & \vdots \\ \mathbf{O} & A_{k-1,k-1} & A_{k-1,k} \\ A_{1,k}^H & \dots & A_{k-1,k}^H & A_{k,k} \end{bmatrix}. \quad (12.2a)$$

We define the index sets $I_I := \cup_{\ell=1}^{k-1} \Omega_h^\ell$ (interior indices) and $I_B := \Gamma$ (boundary indices) and obtain the 2×2 block matrix

$$A = \begin{bmatrix} A_{II} & A_{IB} \\ A_{BI} & A_{kk} \end{bmatrix}, \quad \begin{cases} A_{II} = \text{blockdiag}\{A_{\ell\ell} : 1 \leq \ell \leq k-1\}, \\ A_{BI} := \begin{bmatrix} A_{1,k}^H, \dots, A_{k-1,k}^H \end{bmatrix} = A_{IB}^H. \end{cases} \quad (12.2b)$$

12.4.2 Direct Solution

The Schur complement related to (12.2b) is described in §C.6. A reformulation in terms of the blocks in (12.2a) shows that the system $Ax = b$ is equivalent to

$$\begin{bmatrix} A_{11} & \mathbf{O} & A_{1,k} \\ \ddots & \ddots & \vdots \\ \mathbf{O} & A_{k-1,k-1} & A_{k-1,k} \\ 0 & \dots & 0 & S \end{bmatrix} \begin{bmatrix} x^1 \\ \vdots \\ x^{k-1} \\ x^k \end{bmatrix} = \begin{bmatrix} b^1 \\ \vdots \\ b^{k-1} \\ \hat{b}^k \end{bmatrix},$$

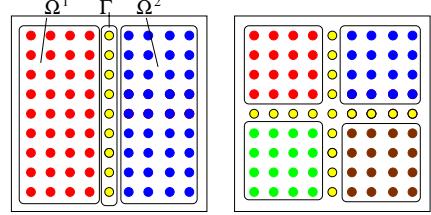


Fig. 12.4 Subdomains and interior boundary Γ .

where the Schur complement⁴ S is

$$S := A_{kk} - \sum_{\ell=1}^{k-1} A_{\ell,k}^H A_{\ell\ell}^{-1} A_{\ell,k}, \quad \hat{b}^k := b^k - \sum_{\ell=1}^{k-1} A_{\ell,k}^H A_{\ell\ell}^{-1} b^\ell. \quad (12.3)$$

Obviously, the system is decoupled. As soon as $Sx^k = \hat{b}^k$ is solved (or approximated), all subproblems $A_{\ell\ell}x^\ell = b^\ell - A_{\ell k}x^k$ ($1 \leq \ell \leq k-1$) can be treated in parallel.

The size of the matrix S is $n_k \times n_k$, where $n_k = \mathcal{O}(h^{1-d}k^{1/d})$ is a typical size. Here, d is the spatial dimension of the domain $\Omega \subset \mathbb{R}^d$.

The solution of $Sx^k = \hat{b}^k$ is rather uncomfortable since S and S^{-1} are fully populated matrices.

Remark 12.1. (a) Computing S explicitly by (12.3) requires the availability of the inverse matrices $A_{\ell\ell}^{-1}$ in order to perform matrix-matrix multiplications in (12.3). In the exact form, this approach is rather costly. Since S is fully populated, Cholesky decomposition is again costly. Using approximations, the hierarchical matrix technique (cf. Appendix D) can be applied successfully (cf. §D.3 and [192]).
(b) A possible approach to $A_{\ell\ell}^{-1}$ is the recursive application of the Schur complement method to the subdomains Ω_ℓ . This leads to the *nested dissection method* of George [149].

(c) If only a subroutine for the exact or approximate realisation of $z^\ell \mapsto A_{\ell\ell}^{-1}z^\ell$ is available, one can perform the map $z \mapsto Sz$, but the inversion of $Sx^k = \hat{b}^k$ is still a problem. Since the entries of S are not available, algebraic linear iterations (e.g., the Gauss–Seidel iteration) cannot be applied.

(d) Since $A > 0$ implies $S > 0$ (cf. Proposition C.64), the CG method can be applied for solving $Sx^k = \hat{b}^k$.

(e) Formulating the Schur complement problem requires nothing other than the matrix data. Hence algebraic iterations solving the Schur complement problem yield again algebraic methods for solving the original problem.

A consequence of Remark C.63b is the next statement.

Remark 12.2. In principle, the equation $Sz = c$ can be solved by setting $b := \begin{bmatrix} 0 \\ c \end{bmatrix}$ and solving $Ax = b$. Then $x = \begin{bmatrix} * \\ z \end{bmatrix}$ contains the desired solution z as the second block in x .

In the present case, this remark does not make sense since we are looking for methods solving $Ax = b$.

According to Remark 12.1d, we may use the CG method. In this case, the spectral condition number $\kappa(S)$ is of interest. In the case of the Poisson model problem, $\kappa(S) = \mathcal{O}(h^{-1})$ can be shown.⁵ Hence, the CG method (applied to the Richardson

⁴ The Schur complement S is also known under the name *capacitance matrix*.

⁵ The matrix S can be regarded as the discretisation of a certain boundary integral equation, whose operator is spectrally equivalent to $(-\Delta)^{1/2}$.

iteration) yields the still unfavourable convergence rate $1 - \mathcal{O}(h^{1/2})$. One can try to find a suitable preconditioning for the matrix S . In the case of an interior boundary Γ without cross-points (as on the left in Fig. 12.4), Dryja [111] proposes the square root of the second differences on the grid line Γ , which is computable by using the fast Fourier transform⁶ (cf. also Mróz [286]).

A variant of the capacitance matrix method termed the ‘method of fictitious domains’ is based on the embedding of domain Ω into a larger domain Ω' . For example, one may choose Ω' as a rectangle in order to obtain an easily solvable discretisation of the differential equation in Ω' . For this subject, we refer the reader to Proskurowski–Widlund [312], Astrachanov [9], Börgers–Widlund [53], and Quarteroni–Valli [315, §1.6].

12.4.3 Preconditioners of the Schur Complement

Instead of looking directly for a preconditioner for S , one may try to precondition S by the Schur complement of a related problem. The first approach is the *Neumann–Dirichlet method* (cf. Widlund [397] and Bjørstad–Widlund [52]; therein references to earlier literature). Let the domain Ω be decomposed into disjoint subdomains Ω_ℓ , as depicted on the left in Figure 12.1. The interior grid points (nodal points) are the sets denoted by I_ℓ for $1 \leq \ell \leq k - 1$.

For the sake of simplicity, we assume $k = 3$ (i.e., two subdomains; see left example in Figure 12.4). The preconditioning of the Schur complement S can be constructed using Galerkin discretisation in Ω_1 :

$$\text{find } v \in V^h \text{ with } a_{\Omega_1}(v, w) = (f, w)_U \quad \text{for all } w \in V^h \quad (12.4)$$

(cf. §E.3). (12.4) yields the (discrete) solution of the differential equation on Ω_1 with the natural boundary condition on $\partial\Omega_1 \cap \partial\Omega_2$. Problem (12.4) involves only the index sets I_1 and I_3 . The system of equations has the block structure

$$\begin{bmatrix} A_{11} & A_{13} \\ A_{13}^H & A'_{33} \end{bmatrix} \begin{bmatrix} x^1 \\ x^3 \end{bmatrix} = \begin{bmatrix} c^1 \\ c^3 \end{bmatrix} \quad (12.5)$$

with a different matrix $A'_{33} \neq A_{33}$. The corresponding Schur complement

$$S' := A'_{33} - A_{13}^H A_{11}^{-1} A_{13} \quad (12.6)$$

is used as a preconditioning matrix for S . The proof of $\kappa(S'^{-1} S) = \mathcal{O}(1)$ requires tools from functional analysis (continuation theorems; cf. Widlund [397]). The solution of a system with the matrix

$$A' := \begin{bmatrix} A_{11} & 0 & A_{13} \\ 0 & A_{22} & A_{23} \\ A_{13}^H & A_{23}^H & A'_{33} \end{bmatrix}$$

⁶ The fast Fourier transform is first described by Gauss [146] in 1805 (cf. Heideman et al. [215]).

starts with the subproblem corresponding to blocks x^1, x^3 using the matrix of (12.5) and is completed by solving for x^2 . If S' is a good preconditioner for S , then A' is also a good preconditioner for A as stated below (see also Mandel [270]).

- Exercise 12.3.** (a) Show that $S' < S$ for S and S' in (12.3) and (12.6).
 (b) Let $\gamma \leq 1 \leq \Gamma$. Prove the equivalence of $\gamma S' \leq S \leq \Gamma S'$ and $\gamma A' \leq A \leq \Gamma A'$.

The solution of $A'x = c$ can be obtained from the exact solution of the subproblems (12.5) and $A_{22}x^2 = c^2$. The exact solution can be approximated by replacing A' by A'' . According to Lemma 7.55, we estimate $\kappa(A''^{-1}A)$ by $\kappa(A''^{-1}A')\kappa(A'^{-1}A) = \kappa(A''^{-1}A')\kappa(S'^{-1}S) \leq \text{const } \kappa(A''^{-1}A')$.

The second approach uses the fact that under certain conditions concerning $S = S' + S''$ the matrices S^{-1} and

$$S'^{-1} + S''^{-1}$$

are spectrally close. Here, S' is defined as above, while $S'' := A'_{33} - A'_{23}A_{22}^{-1}A_{23}$ is the Schur complement of the second subdomain problem with a_{Ω_2} in (12.4). Now Remark 12.2 can be used to perform $z' \mapsto S'^{-1}z'$ and $z'' \mapsto S''^{-1}z''$. This approach is called the *Neumann–Neumann method* (cf. Smith–Bjørstad–Gropp [343, §4.2.2]).

12.4.4 Multigrid-like Domain Decomposition Methods

Domain decomposition can be performed in the following recursive manner: First, the index set I is decomposed into disjoint sets: $I = I_\ell \cup I'_\ell$, then we decompose $I'_\ell = I_{\ell-1} \cup I'_{\ell-1}$, $I'_{\ell-1} = I_{\ell-2} \cup I'_{\ell-2}, \dots$ and end up with the disjoint splitting $I = I_\ell \cup I_{\ell-1} \cup \dots \cup I_1 \cup I_0$. For instance, the indices I_μ may correspond to the grid-point set $\Omega_\mu \setminus \Omega_{\mu-1}$ ($\Omega_{-1} := \emptyset$), where $\Omega_0 \subset \Omega_1 \subset \dots \subset \Omega_\ell$ is a grid hierarchy associated with the step size sequence $h_\mu = 2^{-\mu}h_0$. The index decomposition $I = I_\ell \cup I'_\ell$ corresponds to a partitioning of the matrix $A = A^{(\ell)}$. This block matrix can be written as the product

$$A^{(\ell)} = \begin{bmatrix} A_{11}^{(\ell)} & A_{12}^{(\ell)} \\ A_{12}^{(\ell)\text{H}} & A_{22}^{(\ell)} \end{bmatrix} = L^{(\ell)} \begin{bmatrix} A_{11}^{(\ell)} & 0 \\ 0 & S^{(\ell)} \end{bmatrix} L^{(\ell)\text{H}}, \quad L^{(\ell)} = \begin{bmatrix} I & 0 \\ A_{12}^{(\ell)\text{H}} A_{11}^{(\ell)-1} & I \end{bmatrix},$$

where $S^{(\ell)}$ denotes again the Schur complement. Following Exercise 12.3, preconditioning A needs a good preconditioner of $S^{(\ell)}$ by $S'^{(\ell)}$. Since $S^{(\ell)}$ corresponds to a coarse-grid correction, $S'^{(\ell)}$ can be defined recursively on ℓ . Such approaches can be found in Axelsson [11], Axelsson–Vassilevski [16, 17], Vassilevski [377], Kuznetsov [251, 252, 253]. They lead to methods that are similar to the hierarchical basis multigrid method described in §12.9.4. Occasionally, these methods are termed *algebraic multilevel iteration* (AMLI) although it is not an algebraic iteration in the sense of Definition 2.2b.

12.5 Subspace Iteration

12.5.1 General Construction

Let $X = \mathbb{K}^I$ be the linear space containing the solution x of $Ax = b$. The subproblems, which we index by $\kappa \in J$, correspond to lower-dimensional problems represented by vectors $x^\kappa \in X_\kappa = \mathbb{K}^{I_\kappa}$. The solution x of the system $Ax = b$ is composed of the partial solutions x^κ . For that purpose, we choose linear and injective mappings, which may be called *prolongations*:

$$p_\kappa : X_\kappa \rightarrow X \quad (\kappa \in J). \quad (12.7)$$

The true solution $x = A^{-1}b$ should be expressed in the form

$$x = \sum_{\kappa \in J} p_\kappa x^\kappa. \quad (12.8)$$

This is possible in general if and only if

$$\sum_{\kappa \in J} \text{range}(p_\kappa) = X \quad (12.9)$$

holds. Here, $\text{range}(p_\kappa) = \{p_\kappa x^\kappa : x^\kappa \in X_\kappa\}$ denotes the image space of p_κ ; furthermore, the sum of subspaces $V_\kappa \subset X$ denotes the space

$$\sum_{\kappa \in J} V_\kappa = \text{span}\{V_\kappa : \kappa \in J\} = \{x \in X : x = \sum_{\kappa \in J} v^\kappa : v^\kappa \in V_\kappa\}.$$

p_κ in (12.7) is represented by a rectangular matrix. Its Hermitian transposed matrix p_κ^H describes a mapping from X onto X_κ :

$$r_\kappa := p_\kappa^H : X \rightarrow X_\kappa \quad (\text{restriction}). \quad (12.10)$$

For each $\kappa \in J$, we define the matrices (Galerkin products)

$$A_\kappa := r_\kappa A p_\kappa \quad (\kappa \in J). \quad (12.11)$$

Their size is $n_\kappa \times n_\kappa$, where

$$n_\kappa := \dim X_\kappa. \quad (12.12)$$

The lower-dimensional subproblems are equations of the form

$$A_\kappa y^\kappa = c^\kappa \quad (\kappa \in J, \quad y^\kappa, c^\kappa \in X_\kappa). \quad (12.13)$$

For the present, we assume that problems of the form (12.13) can be solved exactly. The regularity of A_κ is not guaranteed without additional conditions on A . A sufficient condition is given below.

Exercise 12.4. Assume that $A > 0$ and p_κ is injective for all $\kappa \in J$. Prove $A_\kappa > 0$.

We associate the prolongations p_κ with the restrictions r_κ in (12.10) and the projections

$$P_\kappa := p_\kappa A_\kappa^{-1} r_\kappa A = p_\kappa A_\kappa^{-1} r_\kappa A : X \rightarrow X. \quad (12.14a)$$

Exercise 12.5. Prove that (a) P_κ are projections onto $\text{range}(p_\kappa)$ for $\kappa \in J$.
 (b) Let $A > 0$. P_κ is an A -orthogonal projection, i.e., P_κ is selfadjoint with respect to the A -scalar product (9.8a):

$$\langle P_\kappa x, y \rangle_A = \langle x, P_\kappa y \rangle_A \quad \text{for all } x, y \in X. \quad (12.14b)$$

12.5.2 The Prolongations

Let n_κ in (12.12) be the partial dimension and $n = \dim(X)$ the entire one. A first classification is given by the alternatives (12.15a) and (12.15b):

$$\sum_{\kappa \in J} n_\kappa = n \quad (\text{'disjoint subdomains'}), \quad (12.15a)$$

$$\sum_{\kappa \in J} n_\kappa > n \quad (\text{'overlapping subdomains'}). \quad (12.15b)$$

Note that $\sum n_\kappa < n$ is excluded because of (12.9).

Remark 12.6. In case (12.15a), each $x \in X$ allows a unique decomposition (12.8); in case (12.15b), more than one representation (12.8) is possible.

A particularly simple situation of the form (12.15a) arises from the block representation of x . Let $\{I_\kappa : \kappa \in J\}$ describe the block structure (A.7): $I = \dot{\cup}_{\kappa \in J} I_\kappa$. We choose

$$X_\kappa = \mathbb{K}^{I_\kappa}, \quad (p_\kappa)_{\alpha\beta} = \delta_{\alpha\beta} \quad \text{for } \alpha \in I, \beta \in I_\kappa \quad (12.16a)$$

(cf. (1.11)). Therefore, $\sum_{\kappa \in J} p_\kappa x^\kappa$ is only another notation of the vector $x = (x^\kappa)_{\kappa \in J}$ composed of the blocks x^κ . In the case of ordered indices, the matrix p_κ takes the form

$$p_\kappa = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ I \\ 0 \\ \vdots \\ 0 \end{bmatrix} \} \text{ block of index } \kappa, \quad r_\kappa = [0, \dots, 0, I, 0, \dots, 0]. \quad (12.16b)$$

Remark 12.7. If p_κ is defined according to (12.16a,b), the matrices A_κ in (12.11) coincide with the diagonal block $A^{\kappa\kappa}$ of the matrix A (cf. §A.4).

In the case of (12.15b), p_κ can still be defined by (12.16a), but the subsets $I_\kappa \subset I$ are no longer disjoint and therefore do not describe a block decomposition. A property still remaining is mentioned next.

Remark 12.8. If the index set I is decomposed into $I = \cup_{\kappa \in J} I_\kappa$ (possibly not disjointly) and the prolongations p_κ are defined by (12.16a), the matrices A_κ in (12.11) represent the principal submatrices of the matrix A belonging to the index subset I_κ .

12.5.3 Multiplicative and Additive Schwarz Iterations

Here, the subspace iterations are called Schwarz iteration (see historical remarks in §12.1). The projections P_κ in (12.14a) are associated with the (partial) linear iterations

$$\Phi_\kappa(x, b) := x - p_\kappa A_\kappa^{-1} r_\kappa(Ax - b) \quad (\kappa \in J). \quad (12.17)$$

Note the similarity to the coarse-grid correction in (11.19).

According to Definition 5.12 and Exercise 12.5, the iteration Φ_κ is an A -orthogonal projection, provided that $A > 0$.

Exercise 12.9. Prove that the iteration matrix belonging to (12.17) is $M_\kappa = I - P_\kappa$. Like P_κ , the iteration matrix M_κ is a projection. In the case of $A > 0$, M_κ is A -selfadjoint (cf. Exercise 12.5). Φ_k is a positive definite iteration: $\Phi_k \in \mathcal{L}_{\text{pos}}$.

Let the index set $J = \{1, \dots, k\}$ be ordered. The *multiplicative Schwarz iteration* is the k -fold product

$$\Phi^{\text{multSI}} := \Phi_k \circ \Phi_{k-1} \circ \dots \circ \Phi_2 \circ \Phi_1 \quad (\Phi_j \text{ in (12.17)}).$$

In contrast, the *additive Schwarz iteration* (with a damping factor Θ) is defined by

$$\begin{aligned} \Phi_\Theta^{\text{addSI}} &:= \Theta \cdot \sum_{\kappa \in J} \Phi_\kappa, \quad \text{i.e.,} \\ \Phi_\Theta^{\text{addSI}}(x, b) &= x - \Theta \sum_{\kappa \in J} p_\kappa A_\kappa^{-1} r_\kappa(Ax - b), \end{aligned}$$

where the index set J need not be ordered. We recall the algebra of linear iterations discussed in §5 with the product in §5.4 and the summation in §5.3.

Lemma 12.10. (a) The iteration matrices of Φ^{multSI} and $\Phi_\Theta^{\text{addSI}}$ are

$$\begin{aligned} M^{\text{multSI}} &= (I - P_\kappa)(I - P_{\kappa-1}) \dots (I - P_1), \\ M_\Theta^{\text{addSI}} &= I - \Theta \left(\sum_{\kappa \in J} p_\kappa A_\kappa^{-1} r_\kappa \right) A = I - \Theta \sum_{\kappa \in J} P_\kappa. \end{aligned}$$

(b) Let $A > 0$. The matrix of the second normal form of $\Phi_\Theta^{\text{addSI}}$ is

$$N_\Theta^{\text{addSI}} = \Theta N^{\text{addSI}} \quad \text{with} \quad N^{\text{addSI}} = \sum_{\kappa \in J} p_\kappa A_\kappa^{-1} r_\kappa. \quad (12.18a)$$

Under assumption (12.9), N^{addSI} is regular, so that the matrix of the third normal form $W_\Theta^{\text{addSI}} = \Theta^{-1} W^{\text{addSI}}$ with $W^{\text{addSI}} = (N^{\text{addSI}})^{-1}$ exists. It satisfies

$$A \leq kW^{\text{addSI}} \quad (k := \#J = \text{number of 'subdomains'}). \quad (12.18b)$$

Proof. (i) To prove part (b) set $N := N^{\text{addSI}}$. $Nx = 0$ implies $r_\kappa x = 0$ because of $\langle x, Nx \rangle = \langle A_\kappa^{-1} r_\kappa x, r_\kappa x \rangle$. Since $\ker(r_\kappa) = \text{range}(p_\kappa)^\perp$, the orthogonality $x \perp \text{range}(p_\kappa)$ follows for all $\kappa \in J$. From (12.9), we conclude $x = 0$.

(ii) $A^{1/2} p_\kappa A_\kappa^{-1} r_\kappa A^{1/2}$ is $\leq I$, since it is an orthogonal projection. Summation yields $A^{1/2} N A^{1/2} \leq kI$. According to (C.3g), inequality (12.18b) follows from $N \leq kA^{-1}$. \square

Exercise 12.11. Let $A > 0$. Prove the following: (a) The iteration adjoint to Φ^{multSI} is $\Phi_1 \circ \dots \circ \Phi_k$. The corresponding symmetric iteration (cf. (5.13)) is

$$\Phi^{\text{symmultSI}} := \Phi_1 \circ \dots \circ \Phi_{k-1} \circ \Phi_k \circ \Phi_{k-1} \circ \dots \circ \Phi_1.$$

(b) $\Phi_\Theta^{\text{addSI}} \in \mathcal{L}_{\text{pos}}$, provided that $\Theta > 0$.

A first (nonquantitative) convergence statement follows from Lemma 7.5. The iterations Φ_κ are nonexpansive. Hence convergence holds if and only if (5.10) is satisfied. Since $\ker(N_\kappa) = \ker(r_\kappa)$ follows from the injectivity of p_κ , convergence is equivalent to (5.10):

$$(5.10) \quad \Leftrightarrow \quad \bigcap_{\kappa} \ker(r_\kappa) = \{0\} \quad \Leftrightarrow \quad \left(\bigcap_{\kappa} \ker(r_\kappa) \right)^\perp = \{0\}^\perp.$$

Conclusion A.33 yields the identity $\left(\bigcap_{\kappa} \ker(r_\kappa) \right)^\perp = \sum_{\kappa} \ker(r_\kappa)^\perp$. Together with the complements $\ker(r_\kappa)^\perp = \text{range}(p_\kappa)$ and $\{0\}^\perp = X$, we obtain the equivalence with $\sum_{\kappa} \text{range}(p_\kappa) = X$ which is (12.9). The second part of the next proposition follows from Proposition 5.23.

Proposition 12.12. Condition (12.9) implies convergence of Φ^{multSI} . For suitable Θ , the additive version $\Phi_\Theta^{\text{addSI}}$ converges under the same condition.

12.5.4 Interpretation as Gauss–Seidel and Jacobi Iteration

Assume the case (12.15a) ('disjoint subdomains'). Moreover, let p_κ be given by (12.16a). Then Φ_k in (12.17) describes the solution of equation $Ax = b$ with respect to the block of index $\kappa \in J$. This proves part (a) of the next lemma.

Lemma 12.13. (a) Under the assumptions (12.15a) and (12.16a), the multiplicative Schwarz iteration coincides with the block-Gauss–Seidel iteration, whereas the additive Schwarz iteration coincides with the damped block-Jacobi iteration.

(b) Let $A > 0$. The multiplicative Schwarz iteration converges. The additive Schwarz iteration converges for sufficiently small $\Theta > 0$ ($0 < \Theta < 2k$ is always sufficient).

Proof. For part (b), use (12.18b) and Theorems 3.36, 3.39, and 3.50. \square

12.5.5 Classical Schwarz Iteration

A situation not comparable with the Gauss–Seidel or Jacobi iteration arises in case (12.15b). We consider the classical Schwarz iteration described in §12.2.1. The differential equation (12.1) is replaced by the standard difference scheme $A = h^{-2}\text{tridiag}\{-1, 2, -1\} \in \mathbb{K}^{I \times I}$ with the step size $h = \frac{1}{n+1}$ and the index set $I = \{1, \dots, n\}$. The integers n_1 and n_2 are defined by $\frac{1+H}{2} = (n_1 + 1)h$ and $\frac{1-H}{2} = (n_2 - 1)h$. Then the iteration in §12.2.1 corresponds to the choice

$$\begin{aligned} J &= \{1, 2\}, & p_\kappa \text{ in (12.16a),} \\ I_1 &= \{1, \dots, n_1\}, & I_2 = \{n_2, \dots, n\}. \end{aligned}$$

The projection property of the iterations Φ_κ ($\kappa = 1, 2$) implies that

$$r_\kappa A \Phi_\kappa(x, b) = r_\kappa b, \quad \text{i.e., } (\Phi_\kappa(x, b))_\nu = b_\nu \quad \text{for all } \nu \in I_\kappa,$$

while $(\Phi_\kappa(x, b))_\nu = x_\nu$ for $\nu \notin I_\kappa$. Hence, for $\kappa = 1$, the difference equations are exactly solved for $1 \leq \nu \leq n_1$ using the value x_{n_1+1} at $\xi = \frac{1+H}{2}$ as boundary value.

12.5.6 Approximate Solution of the Subproblems

In the case of blockwise Jacobi methods, one chooses the block structure (and thus the block diagonal $D = \text{blockdiag}\{D_\kappa : \kappa \in J\}$) such that equations of the form $D_\kappa y^\kappa = c^\kappa$ are easy to solve. Even if the additive Schwarz iteration can partially be interpreted as a block-Jacobi method, this does not mean that the subproblems (12.13): $A_\kappa y^\kappa = c^\kappa$ are also easily solved. The reason is that, in general, (12.13) also discretises the same differential equation (in a subdomain).

If no direct solver for the system $A_\kappa y^\kappa = c^\kappa$ in (12.13) is available, this subproblem must be solved again by some iteration $\Phi^{(\kappa)}$. Then a *composed iteration* arises, which differs from the one studied in §5.5 only by the fact that k subproblems are iteratively approximated during each outer iteration step. Denote the matrix of

the third normal form of $\Phi^{(\kappa)}$ by W_κ . We only consider symmetric iterations with

$$\delta_\kappa W_\kappa \leq A_\kappa := r_\kappa A p_\kappa \leq \Delta_\kappa W_\kappa \quad (\kappa \in J, 0 < \delta_\kappa \leq \Delta_\kappa). \quad (12.19)$$

Often, $\Delta_\kappa < 2$ is required, i.e., $\Phi^{(\kappa)}$ be convergent (cf. Theorem 3.34a).

As in §5.5, we can construct the iteration $\Phi_{(m)}$ for solving $Ax = b$ by the additive or multiplicative Schwarz iteration, where the exact solution of the subproblems (12.13), $A_\kappa y^\kappa = c^\kappa$, is replaced by $m > 0$ applications⁷ of $\Phi^{(\kappa)}$. Let $W_\kappa^{(m)}$ be the matrix of the third normal form of the m -fold iteration $(\Phi^{(\kappa)})^m$. For simplicity, we write $W_\kappa := W_\kappa^{(m)}$. Then the projection $P_\kappa = p_\kappa A^{-1} r_\kappa A$ in (12.14a) becomes

$$\Pi_\kappa = p_\kappa W_\kappa^{-1} r_\kappa A \quad (\kappa \in J). \quad (12.20a)$$

$I - \Pi_\kappa$ is the iteration matrix of the iteration Φ_κ replacing (12.17):

$$\Phi_\kappa(x, b) := x - p_\kappa W_\kappa^{-1} r_\kappa (Ax - b) \quad (\kappa \in J). \quad (12.20b)$$

Note that Π_κ is no longer a projection. Some of its properties are gathered below.

Remark 12.14. (a) The mapping Π_κ is A -adjoint, i.e., $\Pi_\kappa^H = A \Pi_\kappa A^{-1}$ and $\langle \Pi_\kappa x, y \rangle_A = \langle x, \Pi_\kappa y \rangle_A$ hold (cf. (12.14b)).

(b) $\sigma(\Pi_\kappa) \subset \sigma(W_\kappa^{-1} A_\kappa) \cup \{0\}$, where the equality sign ‘=’ holds instead of ‘ \subset ’, if $n_\kappa = \dim X_\kappa < n = \dim X$.

(c) $\rho(\Pi_\kappa) \leq \Delta_\kappa$ is equivalent to the last inequality in (12.19):

$$A_\kappa \leq \Delta_\kappa W_\kappa \quad (\kappa \in J) \quad (12.21)$$

and gives rise to the estimate (12.23) instead of (12.18b).

The iteration matrices of the multiplicative and additive Schwarz iteration with approximate subspace solvers (12.20b) are

$$M^{\text{multSI}} = (I - \Pi_k)(I - \Pi_{k-1}) \dots (I - \Pi_1), \quad (12.22a)$$

$$M_\Theta^{\text{addSI}} = I - \Theta \left(\sum_{\kappa \in J} p_\kappa W_\kappa^{-1} r_\kappa \right) A = I - \Theta \sum_{\kappa \in J} \Pi_\kappa. \quad (12.22b)$$

In the following, we always assume that approximate subspace solvers are applied. The case of exact subspace solutions considered in §12.5.3 can be regarded as the special choice $W_\kappa = A_\kappa$ leading to $\delta_\kappa = \Delta_\kappa = 1$ in (12.19).

Exercise 12.15. Assume (12.19) and prove the following.

- (a) The additive Schwarz iteration based on (12.20b) is a positive definite iteration.
- (b) Define $N^{\text{addSI}} = \sum_{\kappa \in J} p_\kappa W_\kappa^{-1} r_\kappa$ and $W^{\text{addSI}} := (N^{\text{addSI}})^{-1}$ according to (12.18a). The estimate (12.18b) now becomes

$$A \leq k \max_{\kappa} \{\Delta_\kappa\} W^{\text{addSI}} \quad (k := \#J). \quad (12.23)$$

⁷ The numbers $m = m_\kappa$ may also depend on κ .

12.5.7 Strengthened Estimate $A \leq \Gamma W$

We say that two indices $\kappa, \lambda \in J$ (or the respective subdomains) are connected if

$$\langle A p_\kappa x^\kappa, p_\lambda y^\lambda \rangle \neq 0 \quad \text{for suitable } x^\kappa \in X_\kappa, y^\lambda \in X_\lambda. \quad (12.24)$$

Exercise 12.16. Prove that if there are subsets $I_\kappa, I_\lambda \subset I$ such that p_κ and p_λ satisfy (12.16a), then (12.24) holds if and only if the graph $G(A)$ contains an edge between some vertices $\alpha \in I_\kappa$ and $\beta \in I_\lambda$.

In Figure 12.1, the subdomains indexed by 1 and 3 are not connected because of the sparsity of A . If, as in the block-tridiagonal case, the block-index i is only connected to $i \pm 1$, the set $J = \{1, 2, \dots, k\}$ can be split into $J_1 = \{1, 3, \dots\}$ and $J_2 = \{2, 4, \dots\}$ satisfying the assumptions of the following lemma with $K = 2$.

Lemma 12.17. Let $A > 0$ and assume that J can be decomposed into K subsets J_1, \dots, J_K , so that property (12.24) only applies to indices $\kappa \neq \lambda$ from different sets J_i, J_j ($i \neq j$). Then (12.23) holds in the strengthened form

$$A \leq K \max_{\kappa} \{\Delta_\kappa\} W^{\text{addSI}}. \quad (12.25)$$

Proof. Write $N := N^{\text{addSI}}$ in (12.18a) as a sum $N_1 + \dots + N_K$ with

$$N_i := \sum_{\kappa \in J_i} p_\kappa W_\kappa^{-1} r_\kappa \leq \Delta N'_i, \quad \text{where } N'_i := \sum_{\kappa \in J_i} p_\kappa A_\kappa^{-1} r_\kappa$$

and $\Delta := \max_{\kappa} \{\Delta_\kappa\}$. By definition, (12.24) does not hold for indices $\kappa, \lambda \in J_i$ with $\kappa \neq \lambda$; hence, $\text{range}(p_\kappa) \perp_A \text{range}(p_\lambda)$. This proves that N'_i is an A -orthogonal projection and therefore, as in the proof of Lemma 12.10, satisfies $N'_i \leq A^{-1}$. Summation of $N_i \leq \Delta N'_i \leq \Delta A^{-1}$ yields $N \leq K \Delta A^{-1}$, implying (12.25). \square

Another bound Γ in $A \leq \Gamma W^{\text{addSI}}$ can be derived from the symmetric matrix

$$E = (\varepsilon_{\kappa\lambda})_{k,\lambda \in J} \in \mathbb{R}^{J \times J}, \quad (12.26a)$$

whose entries are the smallest bounds in

$$|\langle p_\kappa x^\kappa, p_\lambda y^\lambda \rangle_A| \leq \varepsilon_{\kappa\lambda} \|x^\kappa\|_W \|y^\lambda\|_W \quad (x^\kappa \in X_\kappa, y^\lambda \in X_\lambda). \quad (12.26b)$$

where $\|\cdot\|_W$ denotes the following norms on X_κ :

$$\|x^\kappa\|_W := \|W_\kappa^{1/2} x^\kappa\|_2 = \sqrt{\langle W_\kappa x^\kappa, x^\kappa \rangle} \quad \text{for } x^\kappa \in X_\kappa. \quad (12.26c)$$

Lemma 12.18. (a) In the case of $W_\kappa = A_\kappa$ (exact subspace solution), the W -norm coincides with the energy norm: $\|x^\kappa\|_W = \|p_\kappa x^\kappa\|_A$.

(b) Under assumption (12.21), $\|p_\kappa x^\kappa\|_A \leq \sqrt{\Delta_\kappa} \|x^\kappa\|_W$ holds.

Proof. For part (a), use $\|x^\kappa\|_W^2 = \langle W_\kappa x^\kappa, x^\kappa \rangle = \langle A_\kappa x^\kappa, x^\kappa \rangle = \langle p_\kappa^H A p_\kappa x^\kappa, x^\kappa \rangle = \langle A p_\kappa x^\kappa, p_\kappa x^\kappa \rangle = \langle p_\kappa x^\kappa, p_\kappa x^\kappa \rangle_A = \|p_\kappa x^\kappa\|_A^2$. \square

The standard Cauchy–Schwarz inequality is $|\langle p_\kappa x^\kappa, p_\lambda y^\lambda \rangle_A| \leq \|x^\kappa\|_W \|y^\lambda\|_W$ with $W_\kappa = A_\kappa$. A strict inequality $\varepsilon_{\kappa\lambda} < 1$ describes a *strengthened* Cauchy–Schwarz inequality. It determines the minimal angle between the subspaces $p_\kappa X_\kappa$ and $p_\lambda X_\lambda$. In particular, $\varepsilon_{\kappa\kappa} = 1$ holds for all $\kappa \in J$.

Remark 12.19. The Cauchy–Schwarz inequality implies $0 \leq \varepsilon_{\kappa\lambda} \leq \sqrt{\Delta_\kappa \Delta_\lambda}$. The number $\varepsilon_{\kappa\lambda}$ vanishes if and only if the respective ranges $p_\kappa X_\kappa$ and $p_\lambda X_\lambda$ are A -orthogonal. By definition, this holds if the indices κ and λ are not connected.

Abbreviate $N := N^{\text{addSI}}$ and estimate NAx by

$$\begin{aligned} \|NAx\|_A^2 &= \left\langle \sum_\kappa p_\kappa W_\kappa^{-1} r_\kappa Ax, \sum_\lambda p_\lambda W_\lambda^{-1} r_\lambda Ax \right\rangle_A \\ &\leq \sum_{\kappa, \lambda} |\langle p_\kappa W_\kappa^{-1} r_\kappa Ax, p_\lambda W_\lambda^{-1} r_\lambda Ax \rangle_A| \\ (12.26c) \quad &\leq \sum_{\kappa, \lambda} \varepsilon_{\kappa\lambda} \|W_\kappa^{-1} r_\kappa Ax\|_W \|W_\lambda^{-1} r_\lambda Ax\|_W \\ &\leq \rho(E) \sum_\kappa \|W_\kappa^{-1} r_\kappa Ax\|_W^2 \end{aligned}$$

using the symmetry of $E = (\varepsilon_{\kappa\lambda})_{\kappa, \lambda}$. The inequality

$$\begin{aligned} \sum_\kappa \|W_\kappa^{-1} r_\kappa Ax\|_W^2 &= \sum_\kappa \langle r_\kappa Ax, W_\kappa^{-1} r_\kappa Ax \rangle = \sum_\kappa \langle Ax, p_\kappa W_\kappa^{-1} r_\kappa Ax \rangle \\ &= \langle Ax, NAx \rangle = \langle x, NAx \rangle_A \\ &\leq \|x\|_A \|NAx\|_A \end{aligned}$$

yields $\|NAx\|_A^2 \leq \rho(E) \|x\|_A \|NAx\|_A$ and hence $\|NAx\|_A \leq \rho(E) \|x\|_A$. This inequality is equivalent to

$$\|A^{1/2} NA^{1/2}\|_2 \leq \rho(E) \Leftrightarrow A^{1/2} NA^{1/2} \leq \rho(E) I \Leftrightarrow N \leq \rho(E) A^{-1},$$

and finally to $A \leq \rho(E) W^{\text{addSI}}$. This proves the next theorem.

Theorem 12.20. $A \leq \Gamma W^{\text{addSI}}$ holds with $\Gamma = \rho(E)$ and E in (12.26a,b):

$$A \leq \rho(E) W^{\text{addSI}}. \quad (12.27)$$

The trivial estimate $\varepsilon_{\kappa\lambda} \leq \sqrt{\Delta_\kappa \Delta_\lambda}$ in Remark 12.19 for connected κ, λ , together with $\rho(E) \leq \|E\|_\infty \leq \max_\kappa \{\Delta_\kappa\} K$, leads us back to (12.23).

12.6 Properties of the Additive Schwarz Iteration

12.6.1 Parallelism

Actual interest in the additive Schwarz iteration is due to its parallelism, which makes the method well-suited for parallel computing. Therefore, we consider the single steps of the algorithm in detail.

1. After computing the partitioned defect $d^\kappa := p_\kappa^H(Ax^m - b)$ for all $\kappa \in J$, the steps

$$d^\kappa \mapsto A_\kappa^{-1}d^\kappa = A_\kappa^{-1}r_\kappa(Ax^m - b) \mapsto \delta x^\kappa := p_\kappa A_\kappa^{-1}d^\kappa = p_\kappa A_\kappa^{-1}r_\kappa(Ax^m - b)$$

are completely independent of each other and can be computed by different processors without any communication.

2. Even the correction step

$$x^{m+1} := x^m - \Theta \sum_{\kappa \in J} \delta x^\kappa \quad (\delta x^\kappa \text{ as defined above})$$

can be executed in parallel if (12.15a) and (12.16a) hold. In this case, one uses the processors of index $\kappa \in J$ for storing the block of index $\kappa \in J$. Then the correction simplifies to $(x^{m+1})^\kappa = (x^m)^\kappa - \Theta(\delta x^\kappa)^\kappa$, and therefore requires only local quantities. In the overlapping case (12.15b), an additional communication is necessary.

3. If according to (ii), the blocks $\{(x^m)^\kappa : \kappa \in J\}$ are distributed over the processors, computing $d^\kappa := p_\kappa^H(Ax^m - b)$ requires communication with all processors of indices $\lambda \in J$ that are connected to κ (definition in (12.24)).

12.6.2 Condition Estimates

A general assumption for following considerations is $A > 0$. Let the matrix $W = W^{\text{addSI}} = \Theta W_\Theta^{\text{addSI}}$ be defined as in Lemma 12.10b. We recall (B.14): the spectral condition number $\kappa(W^{-1}A)$ is the ratio Γ/γ of the optimal bounds in

$$\gamma W^{\text{addSI}} \leq A \leq \Gamma W^{\text{addSI}}. \quad (12.28)$$

Note that the spectral condition number is independent of the choice of Θ . In (12.23), (12.25), and (12.27) we found $\Gamma = k \max_\kappa \{\Delta_\kappa\}$, $K \max_\kappa \{\Delta_\kappa\}$, and $\rho(E)$, respectively, as upper bounds. The following theorem enables us to arrive at an explicit description of a lower bound γ .

Theorem 12.21. Assume that $A > 0$ and $W_\kappa = A_\kappa$. Let C be a constant such that for each $x \in X$, a representation $x = \sum_{\kappa \in J} p_\kappa x^\kappa$ ($x^\kappa \in X_\kappa$) exists with

$$\sum_{\kappa \in J} \langle Ap_\kappa x^\kappa, p_\kappa x^\kappa \rangle \leq C \langle Ax, x \rangle. \quad (12.29)$$

Then the first inequality $\gamma W^{\text{addSI}} \leq A$ in (12.28) holds with $\gamma = 1/C$.

This statement is first mentioned in 1986 by Nepomnyashich [288]. Nevertheless, this theorem is often called the ‘Lemma of P. Lions’ since the later publication [266] contains this statement indirectly. In the explicit form one finds the theorem in Widlund [398].

In the case of (12.15a) (‘disjoint subdomains’), the decomposition of the vector x into $x = \sum_{\kappa \in J} p_\kappa x^\kappa$ is unique (cf. Remark 12.6); in the opposite case (12.15b) (‘overlapping subdomains’), one can choose an appropriate representation $x = \sum_{\kappa \in J} p_\kappa x^\kappa$ for (12.29) from infinitely many possibilities.

The proof of Theorem 12.21 can be omitted, since it is the special case $W_\kappa = A_\kappa$ of Theorem 12.24 (cf. Lemma 12.18a).

The estimate stated in Theorem 12.21 is sharp, as shown by the following result of Xu [407].

Corollary 12.22. If C is the smallest possible constant in (12.29), then $\gamma = 1/C$ is the largest constant in $\gamma W^{\text{addSI}} \leq A$.

Proof. Given x , define $y := A^{-1}W^{\text{addSI}}x$. We may choose the decomposition $x^\kappa := A_\kappa^{-1}r_\kappa Ay \in X_\kappa$, because $\sum p_\kappa x^\kappa = \sum p_\kappa A_\kappa^{-1}r_\kappa Ay = N^{\text{addSI}}Ay = x$. Now

$$\begin{aligned} \sum_{\kappa \in J} \langle Ap_\kappa x^\kappa, p_\kappa x^\kappa \rangle &\stackrel{(12.14a)}{=} \sum_{\kappa \in J} \langle AP_\kappa y, P_\kappa y \rangle \stackrel{(12.14b)}{=} \sum_{\kappa \in J} \langle AP_\kappa^2 y, y \rangle \\ &= \sum_{\kappa \in J} \langle AP_\kappa y, y \rangle = \langle AN^{\text{addSI}}Ay, y \rangle \\ &= \langle N^{\text{addSI}}Ay, Ay \rangle = \langle x, W^{\text{addSI}}x \rangle \stackrel{A \geq \gamma W}{\leq} \frac{1}{\gamma} \langle x, Ax \rangle \end{aligned}$$

proves (12.29) with $C = 1/\gamma$. □

A corresponding result of Bjørstad–Mandel [51] exists for the reverse direction.

Remark 12.23. Let $A > 0$ and $W_\kappa = A_\kappa$. If for all $x \in X$ and *all* decompositions $x = \sum_{\kappa \in J} p_\kappa x^\kappa$ ($x^\kappa \in X_\kappa$), the inequality $\sum_{\kappa \in J} \langle Ap_\kappa x^\kappa, p_\kappa x^\kappa \rangle \geq C \langle Ax, x \rangle$ holds with $C > 0$, then (12.28) is satisfied by $\Gamma = \frac{1}{C}$.

Inequality (12.29) can also be written as $\sum_\kappa \|p_\kappa x^\kappa\|_A^2 \leq C \|x\|_A^2$. Replacing the energy norm $\|\cdot\|_A$ with the W -norm introduced in (12.26c), we obtain a generalisation of Theorem 12.21 to $W_\kappa \neq A_\kappa$.

Theorem 12.24. Assume $A > 0$, $W_\kappa > 0$ ($\kappa \in J$). Let C' be a constant so that for any $x \in X$ a decomposition $x = \sum_{\kappa \in J} x^\kappa$ ($x^\kappa \in X_\kappa$) exists with

$$\sum_{\kappa \in J} \|x^\kappa\|_W^2 \leq C' \|x\|_A^2. \quad (12.30)$$

Then the first inequality $\gamma W^{\text{addSI}} \leq A$ in (12.28) holds with $\gamma = 1/C'$.

Proof. Squaring the inequality

$$\begin{aligned} \|x\|_A^2 &= \langle x, x \rangle_A = \left\langle x, \sum p_\kappa x^\kappa \right\rangle_A = \sum \langle Ax, p_\kappa x^\kappa \rangle \\ &= \sum \langle r_\kappa Ax, x^\kappa \rangle = \sum \left\langle W_\kappa^{-1/2} r_\kappa Ax, W_\kappa^{1/2} x^\kappa \right\rangle \\ &\leq \sum \|W_\kappa^{-1/2} r_\kappa Ax\|_2 \|W_\kappa^{1/2} x^\kappa\|_2 \\ &\leq \sqrt{\sum \|W_\kappa^{-1/2} r_\kappa Ax\|_2^2} \sqrt{\sum \|W_\kappa^{1/2} x^\kappa\|_2^2} \\ &= \sqrt{\sum \|W_\kappa^{-1/2} r_\kappa Ax\|_2^2} \sqrt{\sum \|x^\kappa\|_W^2} \stackrel{(12.30)}{\leq} \\ &\leq \sqrt{\sum \|W_\kappa^{-1/2} r_\kappa Ax\|_2^2} \sqrt{C' \|x\|_A^2} \end{aligned}$$

and cancelling the factor $\|x\|_A^2$ yields $\langle Ax, x \rangle \leq C' \sum \|W_\kappa^{-1/2} r_\kappa Ax\|_2^2$. Since $\|W_\kappa^{-1/2} r_\kappa Ax\|_2^2 = \sum \langle A(p_\kappa W_\kappa^{-1} r_\kappa) Ax, x \rangle = \langle ANAx, x \rangle$, we arrive at the inequality

$$A \leq C' ANA \quad (N = N^{\text{addSI}}),$$

which is equivalent to $A^{-1} \leq C' N = C'(W^{\text{addSI}})^{-1}$ and $A \geq \frac{1}{C'} W^{\text{addSI}}$. \square

If inequality (12.29) can be verified more easily than (12.30), the use of Theorem 12.24 can be avoided as shown next.

Exercise 12.25. Assume that (12.29) is valid with the constant C . Prove that (12.30) holds with $C' := C/\delta$, where $\delta := \min\{\delta_\kappa\}$ involves the lower bounds in (12.19): $\delta_\kappa W_\kappa \leq A_\kappa$.

It would be desirable if the bounds γ and Γ in (12.28) were h -independent. Even if the number k of subdomains is independent of h , k might be a large number (depending on the number of available parallel processors), so that the k -independence of γ and Γ also seems to be desirable. Therefore, the bound $\Gamma = k$ in (12.18b) is not optimal. However, Lemma 12.17 already yields a criterion for $\Gamma = K$, where K does not depend on the number k of the subdomains, but only on the degree of their mutual connectivity. Moreover, Theorem 12.20 may help, if $\rho(E)$ is independent of the parameters. An h - or k -independent lower bound γ can be obtained from Theorem 12.21 or Theorem 12.24 if the respective constant C or C' used there is h - or k -independent.

12.6.3 Convergence Statements

The additive Schwarz iteration $\Phi_\Theta^{\text{addSI}}$ yields a convergent iteration, provided that suitable damping is applied. According to (3.25), the optimal damping factor is $\Theta := 2/(\gamma + \Gamma)$ with γ and Γ in (12.28). This leads us to the contraction number

$$\rho(M_\Theta^{\text{addSI}}) = \|M_\Theta^{\text{addSI}}\|_A \leq (\Gamma - \gamma)/(\Gamma + \gamma).$$

The same rate holds for the gradient method with $\Phi_\Theta^{\text{addSI}}$ as the basic iteration. The best convergence rate $(\sqrt{\Gamma} - \sqrt{\gamma})/(\sqrt{\Gamma} + \sqrt{\gamma})$ is given by the CG method applied to $\Phi_\Theta^{\text{addSI}}$. In any case, a small ratio Γ/γ is favourable. In the latter cases, the value of Θ does not matter. The choice $\Theta = 1$ leads to $W_\Theta^{\text{addSI}} = W^{\text{addSI}}$ and $N_\Theta^{\text{addSI}} = N^{\text{addSI}}$.

A simply analysed situation is the case of two disjoint subdomains (the 2-cyclic case).

Theorem 12.26. *Assume $A > 0$ and (12.15a) with $k = 2$ (i.e., two disjoint domains).*

(a) *Then the optimal bounds γ and Γ in (12.28) have the form*

$$\begin{aligned} \gamma &= 1 - \delta, \quad \Gamma = 1 + \delta \quad \text{with} \\ \delta &:= \|A^{-1/2} p_1^\top A p_2 A^{-1/2}\|_2 < 1 \quad (A_\kappa \text{ in (12.11)}). \end{aligned} \quad (12.31a)$$

(b) *$\Theta = 1$ is the optimal damping factor of the additive Schwarz iteration and yields the convergence rate $\rho(M_1^{\text{addSI}}) = \|M_1^{\text{addSI}}\|_A = \delta$. The CG method applied to $\Phi_\Theta^{\text{addSI}}$ has the asymptotic rate $\delta/(1 + \sqrt{1 - \delta^2})$.*

(c) *The number δ in (12.31a) is also the best bound in the strengthened Cauchy–Schwarz inequality*

$$|\langle x, y \rangle_A| \leq \delta \|x\|_A \|y\|_A \quad \text{for all } x \in \text{range}(p_1), y \in \text{range}(p_2). \quad (12.31b)$$

Proof. (i) Inserting $x = p_1 x^1$ and $y = p_2 x^2$ in (12.31b) with $x^\kappa \in X_\kappa$ and exploiting

$$\|p_\kappa x^\kappa\|_A = \sqrt{\langle A p_\kappa x^\kappa, p_\kappa x^\kappa \rangle} = \sqrt{\langle A_\kappa x^\kappa, x^\kappa \rangle} = \|A_\kappa^{1/2} x^\kappa\|_2,$$

we prove that the optimal δ in (12.31b) coincides with the norm in (12.31a).

(ii) The coincidence of the additive Schwarz iteration with the block-Jacobi iteration leads to the equivalence of inequality (12.28) to $\gamma D \leq A \leq \Gamma D$ with $D := \text{blockdiag}\{A_1, A_2\}$. Also

$$(\gamma - 1)D \leq A - D \leq (\Gamma - 1)D$$

is an equivalent inequality. Lemma 4.8 applied to $B := D^{-1/2}(A - D)D^{-1/2}$ yields $\lambda_{\max}(B) = \delta$ and $\lambda_{\min}(B) = -\delta$. The remaining statements follow from $\gamma - 1 = -\delta$ and $\Gamma - 1 = \delta$. \square

The constant δ in (12.31b) coincides with ε_{12} in (12.26a,b).

A corresponding analysis for two overlapping subdomains is given by Bjørstad–Mandel [51]. Different from the nonoverlapping case in Theorem 12.26, $\Gamma = 2$ cannot be improved. For the proof, choose $0 \neq x \in \text{range}(p_1) \cap \text{range}(p_2)$, which is possible because of (12.15b). Hence there are $x^\kappa \in \text{range}(p_\kappa)$ with $x = p_\kappa x^\kappa$. The identity

$$p_\kappa A_\kappa^{-1} r_\kappa Ax = p_\kappa A_\kappa^{-1} (r_\kappa A p_\kappa) x^\kappa = p_\kappa x^\kappa = x$$

for $k = 1, 2$ leads to $NAx = 2x$ or $Ax = 2Wx$ and implies $\Gamma \geq 2$. On the other hand, (12.18b) ensures that $\Gamma \leq k = 2$. A multiple eigenvalue $\Gamma = 2$ hardly troubles the CG method, but it influences Φ^{addSI} as well as the gradient method based on Φ^{addSI} .

Let $\Phi_{(m)}$ be the additive Schwarz iteration combined with an m -fold application of the solver $\Phi^{(\kappa)}$ for the subproblem $A_\kappa y^\kappa = c^\kappa$. Above we analysed the case $m = 1$. The case $m > 1$ is discussed below in detail. Note that now the matrix W_κ belongs to the third normal form of $\Phi^{(\kappa)}$ (and not of $(\Phi^{(\kappa)})^m$).

Remark 12.27. Assume $A > 0$ and $\Phi^{(\kappa)} \in \mathcal{L}_{\text{sym}}$ for $\kappa \in J$. Let $\Phi_{(m)}$ be the composed iteration defined above.

(a) Then $\Phi_{(m)} \in \mathcal{L}_{\text{sym}}$ holds for any $m \in \mathbb{N}$ and the corresponding matrix $N_{(m)}$ of the second normal form has the representation

$$N_{(m)} = \sum_{\kappa \in J} p_\kappa \left[I - (I - W_\kappa^{-1} A_\kappa)^m \right] A_\kappa^{-1} p_\kappa^H.$$

We rewrite (12.19) in the form

$$\gamma_\kappa W_\kappa \leq A_\kappa \leq \Gamma_\kappa W_\kappa \quad (\gamma_\kappa > 0, \kappa \in J). \quad (12.32)$$

This implies that $\Phi^{(\kappa)} \in \mathcal{L}_{\text{pos}}$. If m is odd or $\Gamma_\kappa \leq 2$, then $\Phi_{(m)} \in \mathcal{L}_{\text{semi}}$ (cf. Definition 5.11). $\Phi_{(m)} \in \mathcal{L}_{\text{pos}}$ is ensured if m is odd or $\Gamma_\kappa < 2$.

(c) Assume that $\gamma W^{\text{addSI}} \leq A \leq \Gamma W^{\text{addSI}}$ holds in the case of the exact solution of the subproblems (i.e., $A_\kappa = W_\kappa$), and that either m is odd or $\gamma_\kappa \geq 0$. Then the matrix $W_{(m)}$ of the third normal form of $\Phi_{(m)}$ satisfies

$$\gamma \min_{\kappa} \{1 - (1 - \gamma_\kappa)^m, 1 - (1 - \Gamma_\kappa)^m\} W_{(m)} \leq A \leq \Gamma W_{(m)}.$$

Proof. We rewrite $N_{(m)}$ of part (a) as

$$N_{(m)} = \sum_{\kappa \in J} p_\kappa A_\kappa^{-1/2} \left[I - (I - X_\kappa)^m \right] A_\kappa^{-1/2} p_\kappa^H \quad \text{with} \quad X_\kappa := A_\kappa^{1/2} W_\kappa^{-1} A_\kappa^{1/2}.$$

$\Phi^{(\kappa)} \in \mathcal{L}_{\text{sym}}$ implies $X_\kappa = X_\kappa^H$ and $N_{(m)} = N_{(m)}^H$, so that $\Phi_{(m)} \in \mathcal{L}_{\text{sym}}$. Since $A_\kappa \leq \Gamma_\kappa W_\kappa$ in (12.32) together with $\Gamma_\kappa \geq \gamma_\kappa > 0$ implies $W_\kappa > 0$, $\Phi^{(\kappa)} \in \mathcal{L}_{\text{pos}}$ follows.

Inequality (12.32) can be written as $\sigma(X_\kappa) \subset [\gamma_\kappa, \Gamma_\kappa]$ and yields the enclosure $\sigma(I - (I - X_\kappa)^m) \subset [a_{m,\kappa}, b_{m,\kappa}]$. Define $f(x) := 1 - (1 - x)^m$. The values of

$a_{m,\kappa}$, $b_{m,\kappa}$ depend on the parity of m . For odd m , we have

$$[a_{m,\kappa}, b_{m,\kappa}] := [f(\gamma_\kappa), f(\Gamma_\kappa)],$$

whereas for odd m

$$[a_{m,\kappa}, b_{m,\kappa}] := \begin{cases} [f(\gamma_\kappa), f(\Gamma_\kappa)] & \text{if } \Gamma_\kappa \leq 1, \\ [\min\{f(\gamma_\kappa), f(\Gamma_\kappa)\}, 1] & \text{if } \gamma_\kappa \leq 1 \leq \Gamma_\kappa, \\ [f(\Gamma_\kappa), f(\gamma_\kappa)] & \text{if } \gamma_\kappa \geq 1 \end{cases} \subset [a_{m,\kappa}, 1].$$

The first requirements in the part (b) imply $a_{m,\kappa} \geq 0$ and therefore $\Phi_{(m)} \in \mathcal{L}_{\text{semi}}$, while the stronger assumption ensures $a_{m,\kappa} > 0$ and therefore $\Phi_{(m)} \in \mathcal{L}_{\text{pos}}$.

The matrix $N_{(m)}$ satisfies $aN \leq N_{(m)} \leq bN$ with $N = \sum_{\kappa \in J} p_\kappa A_\kappa^{-1} p_\kappa^H$ and $a = \min_\kappa \{a_{m,\kappa}\}$, $b = \max_\kappa \{b_{m,\kappa}\}$. The latter inequality is equivalent to $aW_{(m)} \leq W \leq bW_{(m)}$ involving the inverse matrices $W = W^{\text{addSI}}$, and so on. The combination with $\gamma W^{\text{addSI}} \leq A \leq \Gamma W^{\text{addSI}}$ yields

$$\gamma a W_{(m)} \leq A \leq \Gamma b W_{(m)}.$$

Under the conditions of part (c), we get the desired estimate. \square

Instead of the iterative solution of $A_\kappa y^\kappa = c^\kappa$, one should also take a semi-iterative treatment into consideration. If the convergence rates of $\Phi^{(\kappa)}$ are clearly different, one should use different iteration numbers m_κ such that all $b_{m,\kappa}$ are similar in size.

12.7 Analysis of the Multiplicative Schwarz Iteration

12.7.1 Convergence Statements

The case of $k = 2$ nonoverlapping domains is very easy to analyse.

Exercise 12.28. Prove that under the assumptions of Theorem 12.26, the multiplicative Schwarz iteration has the convergence rate δ^2 . Hint: use Remark 4.21.

The following general analysis for $k > 1$ is based on the presentation of Xu [409] (details in Xu–Zikatnov [410]; see also Bramble–Pasciak–Wang–Xu [74], Griebel–Oswald [170], Oswald [304]) and private communications with H. Yserentant. From the beginning we consider the case that the subproblems are solved approximately using W_i ($i \in J$) (cf. §12.5.6). The choice $A_i = W_i$ corresponds to the exact solution.

The convergence will be based on two inequalities similar to (12.30) and (12.26b). As discussed in Remark 12.19, the constants $\varepsilon_{\kappa\lambda}$ in (12.26b) are smaller the smaller the subspaces X_κ are. Therefore, we introduce the set $\{Y_j : 1 \leq j \leq k\}$ of subspaces with the following properties:

$$Y_j \subset X_j, \quad \sum_{j=1}^k p_j Y_j = X,$$

where $p_j Y_j$ is the range of p_j restricted to Y_j . The second property ensures that every $x \in X$ has a representation $x = \sum p_j y^j$ with $y^j \in Y_j$. In the overlapping case (12.15b), one may, e.g., choose Y_j such that $\{Y_j\}_{1 \leq j \leq k}$ is nonoverlapping.

Inequality (12.30) is now required in the following form. There is a bound C_1 such that for each $x \in X$ we have

$$\sum_{j=1}^k \|y^j\|_W^2 \leq C_1 \|x\|_A^2 \quad \text{for a suitable decomposition } x = \sum p_j y^j \text{ with } y^j \in Y_j. \quad (12.33)$$

If $Y_j = X_j$, (12.33) and condition (12.30) in Theorem 12.24 are identical; otherwise, condition (12.33) is stronger, since there may be fewer decompositions $x = \sum p_j y^j$ with $y^j \in Y_j$ than $x = \sum p_j x^j$ with $x^j \in X^j$ as in (12.30).

Besides (12.33), we need estimates similar to strengthened Cauchy inequalities. Let ε_{ij}^{XY} be the smallest numbers with

$$\begin{aligned} |\langle p_i x^i, p_j y^j \rangle_A| &\leq \varepsilon_{ij}^{XY} \|x^i\|_W \|y^j\|_W \\ \text{for all } x^i \in X_i, y^j \in Y_j, \text{ and } i < j. \end{aligned} \quad (12.34a)$$

For $i \geq j$, we define $\varepsilon_{ij}^{XY} := 0$. If $X_j = Y_j$ and $i < j$, (12.34a) is identical with (12.26b), i.e., $\varepsilon_{ij}^{XY} = \varepsilon_{ij}$; otherwise, $\varepsilon_{ij}^{XY} \leq \varepsilon_{ij}$ may become a strict inequality. We form the strictly upper triangular matrix

$$E_{XY} := (\varepsilon_{ij}^{XY})_{i,j=1,\dots,k} \quad (12.34b)$$

and denote its spectral norm by

$$C_2 := \|E_{XY}\|_2. \quad (12.34c)$$

An estimate of this $k \times k$ matrix is the subject of the next exercise.

Exercise 12.29. Prove that $\varepsilon_{ij}^{XY} \leq \Delta$ with $\Delta := \max_\kappa \Delta_\kappa$ in (12.21) for all $i < j$ and that $C_2 \leq \Delta \sqrt{k(k-1)/2} < k\Delta$. Hint: $\|E_{XY}\|_2^2 \leq \|E_{XY}^\top E_{XY}\|_\infty$.

The following theorem which will be proved in §12.7.2 yields the first convergence result.

Theorem 12.30. Assume (12.21): $A_j \leq \Delta W_j$ with $\Delta < 2$. Let C_1 and C_2 be the numbers defined in (12.33) and (12.34c). Then the multiplicative Schwarz iteration converges monotonically with respect to the energy norm. The contraction number can be estimated by

$$\|M^{\text{multSI}}\|_A \leq \sqrt{1 - \frac{2 - \Delta}{C_1 (1 + C_2)^2}} \quad (12.35)$$

Inserting the bound in Exercise 12.29, we obtain the k -dependent rate $\|M^{\text{multSI}}\|_A \leq 1 - \mathcal{O}(k^{-2})$. If, however, the bounds C_1, C_2 are k -independent, the convergence is also. If the subspace problems are solved exactly ($A_j = W_j$), the factor $2 - \Delta$ in (12.35) becomes 1.

The second convergence statement replaces the estimates (12.34a) by

$$\|x\|_A^2 \leq C_3 \sum_{j=2}^k \frac{\|y^j\|_W^2}{j-1} \quad \text{for any } x = \sum_{j=2}^k p_j y^j \text{ with } y^j \in Y_j. \quad (12.36)$$

The relation to (12.34a–c) can be seen from the sufficient condition stated below.

Lemma 12.31. *Let δ_{ij} ($1 \leq i, j \leq k$) be the smallest numbers satisfying*

$$\langle p_i y^i, p_j y^j \rangle_A \leq \delta_{ij} \|y^i\|_W \|y^j\|_W \quad \text{for all } y^i \in Y_i, y^j \in Y_j. \quad (12.37a)$$

Define the symmetric $k \times k$ matrix E_{YY} by

$$E_{YY} := \left(\sqrt{i-1} \sqrt{j-1} \delta_{ij} \right)_{i,j=1,\dots,k}. \quad (12.37b)$$

Then estimate (12.36) follows with $C_3 := \rho(E_{YY})$.

Proof. The norm of $x = \sum_{j=2}^k p_j y^j$ can be estimated by

$$\begin{aligned} \|x\|_A^2 &= \sum_{i,j=2}^k \langle p_i y^i, p_j y^j \rangle_A \stackrel{(12.37a)}{\leq} \sum_{i,j=2}^k \delta_{ij} \|y^i\|_W \|y^j\|_W \\ &\stackrel{(12.37b)}{=} \sum_{i,j=2}^k (E_{YY})_{ij} \left[(i-1)^{-1/2} \|y^i\|_W \right] \left[(j-1)^{-1/2} \|y^j\|_W \right] \\ &\leq \rho(E_{YY}) \left[\sum_{j=2}^k (j-1)^{-1} \|y^j\|_W^2 \right]. \end{aligned}$$

Note that $\delta_{ij} \leq \varepsilon_{ij}^{XY}$ with ε_{ij}^{XY} in (12.34a) for $i < j$ and $W_i = A_i$, since both arguments y^i and y^j belong to possibly smaller subspaces. The weights $\sqrt{i-1}$ in (12.37b) demonstrate that the ordering of the substeps in the multiplicative approach is essential. Unfortunately, C_3 is not k -independent as demonstrated in part (b) of the next exercise.

Exercise 12.32. (a) Prove that (12.37a) is always true for the choice $\delta_{ij} = \Delta$ ($\Delta := \max_\kappa \Delta_\kappa$ in (12.21)), which leads to $C_3 = \Delta k(k-1)/2$. Furthermore, one may use the estimate $C_3 \leq (k-1)\rho((\delta_{ij})_{i,j=1,\dots,k})$.

(b) For $W_i = A_i$, prove $C_3 \geq k-1$.

The convergence result involving inequality (12.36) is the following.

Theorem 12.33. *Assume (12.21): $A_j \leq \Delta W_j$ with $\Delta < 2$. Let C_1 and C_3 be the numbers defined in (12.33) and (12.36). Then the multiplicative Schwarz iteration converges monotonically with respect to the energy norm. The contraction number can be estimated by*

$$\|M^{\text{multSI}}\|_A \leq \sqrt{1 - \frac{2-\Delta}{C_1 (1 + \sqrt{\Delta C_3})^2}} < 1. \quad (12.38)$$

It may happen that C_2 in (12.34c) is larger than $\mathcal{O}(1)$ because of the first k_0 rows in the matrix E_{XY} . Then

$$C_{2,k_0} := \rho(E_{XY,k_0}), \quad E_{XY,k_0} := (\varepsilon_{ij}^{XY})_{i,j=k_0+1,\dots,k} \quad (12.39)$$

might be smaller than $C_2 = C_{2,0}$. The following result is due to Dryja–Widlund [114] (proof in §12.7.2).

Corollary 12.34. Assume (12.21): $A_\kappa \leq \Delta W_\kappa$ with $\Delta < 2$. Let C_1 be defined by (12.33) and C_{2,k_0} by (12.39) for some $k_0 \in \{1, \dots, k\}$. Then

$$\|M^{\text{multSI}}\|_A \leq \sqrt{1 - \frac{2 - \Delta}{C_1 \left[1 + \sqrt{C_{2,k_0}^2 + \Delta k_0 \left(1 + \frac{1}{C_1} \right) \left(1 + \frac{\Delta(k_0+1)}{2} \right)} \right]^2}}.$$

12.7.2 Proofs of the Convergence Theorems

The products

$$M_i := (I - \Pi_i)(I - \Pi_{i-1}) \cdot \dots \cdot (I - \Pi_1) \quad (1 \leq i \leq k)$$

with $\Pi_i = p_i W_i^{-1} r_i A$ in (12.20a) are the iteration matrices corresponding to the products $\Phi_i \circ \dots \circ \Phi_1$ of the first i substeps. As usual, the empty product is

$$M_0 = I.$$

According to (12.22a), we have $M^{\text{multSI}} = M_k$. Summing the identities

$$M_{i-1} - M_i = \Pi_i M_{i-1} \quad (1 \leq i \leq k), \quad (12.40a)$$

we obtain

$$I - M_i = \sum_{j=1}^i \Pi_j M_{j-1} \quad (1 \leq i \leq k). \quad (12.40b)$$

Lemma 12.35. Let Δ_j be defined by (12.21).

- (a) $\|\Pi_j x\|_A^2 \leq \Delta_j \langle \Pi_j x, x \rangle_A$ holds for all $x \in X$ and $j \in J$.
- (b) $\|p_j y^j\|_A \leq \sqrt{\Delta_j} \|y^j\|_W$ for all $y^j \in Y_j$.

Proof. (a) Since $A_j \leq \Delta_j W_j$, we have

$$\begin{aligned} \|\Pi_j x\|_A^2 &= \langle A p_j W_j^{-1} r_j A x, p_j W_j^{-1} r_j A x \rangle = \langle A p_j W_j^{-1} r_j A p_j W_j^{-1} r_j A x, x \rangle \\ &= \langle A p_j W_j^{-1} A_j W_j^{-1} r_j A x, x \rangle \leq \langle A p_j W_j^{-1} r_j A x, x \rangle \end{aligned}$$

proving $\|\Pi_j x\|_A^2 \leq \Delta_j \langle \Pi_j x, x \rangle_A$.

- (b) Use $\|p_j y^j\|_A^2 = \langle A p_j y^j, p_j y^j \rangle = \langle r_j A p_j y^j, y^j \rangle \stackrel{(12.11)}{=} \langle A_j y^j, y^j \rangle \stackrel{(12.21)}{\leq} \Delta_j \langle W_j y^j, y^j \rangle \stackrel{(12.26c)}{=} \Delta_j \|y^j\|_W^2$. \square

Lemma 12.36. Let $\Delta := \max_j \Delta_j$. For all $x \in X$, we have

$$\|x\|_A^2 - \|M^{\text{multSI}}x\|_A^2 \geq (2 - \Delta) \sum_{i=1}^k \langle \Pi_i M_{i-1}x, M_{i-1}x \rangle_A. \quad (12.41)$$

Proof. (12.40a) and Lemma 12.35b show that

$$\begin{aligned} \|M_{i-1}x\|_A^2 - \|M_i x\|_A^2 &= \|M_{i-1}x\|_A^2 - \|(M_{i-1} - \Pi_i M_{i-1})x\|_A^2 \\ &= 2 \langle \Pi_i M_{i-1}x, M_{i-1}x \rangle_A - \|\Pi_i M_{i-1}x\|_A^2 \geq (2 - \Delta_j) \langle \Pi_i M_{i-1}x, M_{i-1}x \rangle_A. \end{aligned}$$

Summation over i yields (12.41), since $M_0 = I$ and $M_k = M^{\text{multSI}}$. \square

Lemma 12.37. Let C_1 be the bound in (12.33). Then, for all $x, x_j \in X$, the decomposition $x = \sum p_j y^j$ in (12.33) with $y^j \in Y_j$ is such that

$$\sum_{j=1}^k \langle p_j y^j, x_j \rangle_A \leq \sqrt{C_1} \|x\|_A \sqrt{\sum_{j=1}^k \langle \Pi_j x_j, x_j \rangle_A}. \quad (12.42)$$

Proof. Choose the decomposition according to (12.33). Summation of

$$\langle p_j y^j, x_j \rangle_A = \langle y^j, r_j A x_j \rangle = \left\langle W_j^{\frac{1}{2}} y^j, W_j^{-\frac{1}{2}} r_j A x_j \right\rangle \leq \|y^j\|_W \|W_j^{-\frac{1}{2}} r_j A x_j\|_2,$$

together with $\|W_j^{-1/2} r_j A x_j\|_2^2 = \langle A p_j W_j^{-1} r_j A x_j, x_j \rangle = \langle \Pi_j x_j, x_j \rangle_A$, yields

$$\sum \langle p_j y^j, x_j \rangle_A \leq \sum \|y^j\|_W \sqrt{\langle \Pi_j x_j, x_j \rangle_A} \leq \sqrt{\sum \|y^j\|_W^2} \sqrt{\sum \langle \Pi_j x_j, x_j \rangle_A}.$$

Applying (12.33) to the first square root, we arrive at (12.42). \square

Proof of Theorem 12.30. For a given vector $x \in X$, we choose a decomposition $x = \sum p_j y^j$ with $y^j \in Y_j$ satisfying the estimate (12.33). We write $\|x\|_A^2$ as

$$\|x\|_A^2 = \sum_j \langle x, p_j y^j \rangle_A = \sum_j \langle M_{j-1}x, p_j y^j \rangle_A + \sum_j \langle (I - M_{j-1})x, p_j y^j \rangle_A. \quad (12.43a)$$

To estimate the first term on the right-hand side, apply Lemma 12.37 with $x_j := M_{j-1}x$:

$$\sum_j \langle M_{j-1}x, p_j y^j \rangle_A \leq \sqrt{C_1} \|x\|_A \sqrt{\sum_{j=1}^k \langle \Pi_j M_{j-1}x, M_{j-1}x \rangle_A}. \quad (12.43b)$$

For the second term in (12.43a), use (12.40b):

$$\langle (I - M_{j-1})x, p_j y^j \rangle_A = \sum_{i=1}^{j-1} \langle \Pi_i M_{i-1}x, p_j y^j \rangle_A.$$

Since $\Pi_i M_{i-1} x = p_i W_i^{-1} r_i A M_{i-1} x \in p_i X_i$, (12.34a) can be applied and yields

$$\begin{aligned} \sum_{j=1}^k \langle (I - M_{j-1})x, p_j y^j \rangle_A &\leq \sum_{1 \leq i < j \leq k} \langle \Pi_i M_{i-1} x, p_j y^j \rangle_A \\ &\leq \sum_{1 \leq i < j \leq k} \varepsilon_{ij}^{XY} \|W_i^{-1} r_i A M_{i-1} x\|_W \|y^j\|_W = \langle E_{XY} \alpha, \beta \rangle, \end{aligned} \quad (12.43c)$$

where the vectors $\alpha, \beta \in \mathbb{R}^k$ have the components $\alpha_i = \|W_i^{-1} r_i A M_{i-1} x\|_W$ and $\beta_j = \|y^j\|_W$. From $\langle E_{XY} \alpha, \beta \rangle \leq \|E_{XY}\|_2 \|\alpha\|_2 \|\beta\|_2 \leq C_2 \|\alpha\|_2 \|\beta\|_2$ and

$$\begin{aligned} \|\alpha\|_2^2 &= \sum_i \|W_i^{-1} r_i A M_{i-1} x\|_W^2 = \sum_i \langle r_i A M_{i-1} x, W_i^{-1} r_i A M_{i-1} x \rangle \\ &= \sum_i \langle A p_i W_i^{-1} r_i A M_{i-1} x, M_{i-1} x \rangle = \sum_i \langle \Pi_i M_{i-1} x, M_{i-1} x \rangle_A, \end{aligned}$$

we conclude that

$$\sum_{j=1}^k \langle (I - M_{j-1})x, p_j y^j \rangle_A \leq C_2 \sqrt{\sum_j \|y^j\|_W^2} \sqrt{\sum_i \langle \Pi_i M_{i-1} x, M_{i-1} x \rangle_A}.$$

Inequality (12.33) proves that

$$\sum_{j=1}^k \langle (I - M_{j-1})x, p_j y^j \rangle_A \leq C_2 \sqrt{C_1} \|x\|_A \sqrt{\sum_i \langle \Pi_i M_{i-1} x, M_{i-1} x \rangle_A}. \quad (12.43d)$$

Together, (12.43a, b, d) yield

$$\|x\|_A^2 \leq \sqrt{C_1}(1 + C_2) \|x\|_A \sqrt{\sum_i \langle \Pi_i M_{i-1} x, M_{i-1} x \rangle_A}. \quad (12.43e)$$

Therefore, $\sum_i \langle \Pi_i M_{i-1} x, M_{i-1} x \rangle_A$ is bounded from below by

$$\sum_i \langle \Pi_i M_{i-1} x, M_{i-1} x \rangle_A \geq \|x\|_A^2 / [C_1(1 + C_2)^2].$$

This estimate and Lemma 12.36 show that

$$\begin{aligned} \|M^{\text{multSI}} x\|_A^2 &\leq \|x\|_A^2 - (2 - \Delta) \sum_i \langle \Pi_i M_{i-1} x, M_{i-1} x \rangle_A \\ &\leq \|x\|_A^2 - (2 - \Delta) \|x\|_A^2 / [C_1(1 + C_2)^2]. \end{aligned} \quad (12.43f)$$

Since $x \in X$ is arbitrary, the last inequality implies the estimate (12.35) and ends the proof of Theorem 12.30. \square

Proof of Theorem 12.33. The proof differs only with respect to the estimation of the left-hand side in (12.43c). Using (12.40b), we conclude as follows:

$$\begin{aligned}
& \sum_{j=1}^k \langle (I - M_{j-1})x, p_j y^j \rangle_A = \sum_{1 \leq i < j \leq k} \langle \Pi_i M_{i-1} x, p_j y^j \rangle_A \\
&= \sum_{i=1}^{k-1} \sum_{j=i+1}^k \langle \Pi_i M_{i-1} x, p_j y^j \rangle_A = \sum_{i=1}^{k-1} \left\langle \Pi_i M_{i-1} x, \sum_{j=i+1}^k p_j y^j \right\rangle_A \\
&\leq \sum_{i=1}^{k-1} \|\Pi_i M_{i-1} x\|_A \left\| \sum_{j=i+1}^k p_j y^j \right\|_A \\
&\leq \sqrt{\sum_{i=1}^{k-1} \|\Pi_i M_{i-1} x\|_A^2} \sqrt{\sum_{i=1}^{k-1} \left\| \sum_{j=i+1}^k p_j y^j \right\|_A^2}. \tag{12.43g}
\end{aligned}$$

Applying (12.36) to $x = \sum_{j=i+1}^k p_j y^j$ (i.e., $y^2 = \dots = y^i = 0$), we conclude that

$$\begin{aligned}
\sum_{i=1}^{k-1} \left\| \sum_{j=i+1}^k p_j y^j \right\|_A^2 &\leq \sum_{i=1}^{k-1} C_3 \sum_{j=i+1}^k \frac{1}{j-1} \|y^j\|_W^2 \\
&= C_3 \sum_{j=2}^k \sum_{i=1}^{j-1} \frac{1}{j-1} \|y^j\|_W^2 = C_3 \sum_{j=2}^k \|y^j\|_W^2.
\end{aligned}$$

Lemma 12.35b shows that $\|\Pi_i M_{i-1} x\|_A^2 \leq \Delta_i \langle \Pi_i M_{i-1} x, M_{i-1} x \rangle_A$. Together, we obtain

$$\begin{aligned}
\sum_{j=1}^k \langle (I - M_{j-1})x, p_j y^j \rangle_A &\leq \sqrt{\Delta C_3} \sqrt{\sum_{j=2}^k \|y^j\|_W^2} \sqrt{\sum_{i=1}^{k-1} \langle \Pi_i M_{i-1} x, M_{i-1} x \rangle_A} \\
&\stackrel{(12.33)}{\leq} \sqrt{\Delta C_1 C_3} \|x\|_A \sqrt{\sum_{i=1}^{k-1} \langle \Pi_i M_{i-1} x, M_{i-1} x \rangle_A}
\end{aligned}$$

for $\Delta = \max_i \Delta_i$. A comparison with (12.43d) shows that C_2 in (12.43d) is replaced by $\sqrt{\Delta C_3}$. Substituting C_2 in (12.35) by $\sqrt{\Delta C_3}$, we prove (12.38). \square

Proof of Corollary 12.34. The sum in (12.43c) over $1 \leq i < j \leq k$ can be split into a first part Σ_I with $i \leq k_0$ and a second part Σ_{II} with $k_0 < i < j \leq k$. The estimate of Σ_{II} is obtained by replacing the lower index bound 1 by $k_0 + 1$:

$$\sum_{j=k_0+1}^k \langle (I - M_{i-1}) x, p_j y^j \rangle_A \leq C_{2,k_0} \sqrt{C_1} \|x\|_A \sqrt{\sum_{i=k_0+1}^k \langle \Pi_i M_{i-1} x, M_{i-1} x \rangle_A}.$$

The first sum with $i \leq k_0$ reads

$$\Sigma_I := \sum_{j=1}^k \sum_{i=1}^{\min\{j-1, k_0\}} \langle \Pi_i M_{i-1} x, p_j y^j \rangle_A$$

$$= \sum_{i=1}^{k_0} \left\langle \Pi_i M_{i-1} x, \sum_{j=i+1}^k p_j y^j \right\rangle_A = \sum_{i=1}^{k_0} \left\langle \Pi_i M_{i-1} x, x - \sum_{j=1}^i p_j y^j \right\rangle_A.$$

Applying the Cauchy–Schwarz inequality, we arrive at

$$\Sigma_I^2 \leq \left[\sum_{i=1}^{k_0} \langle \Pi_i M_{i-1} x, \Pi_i M_{i-1} x \rangle_A \right] \times \left[\sum_{i=1}^{k_0} \|x - \sum_{j=1}^i p_j y^j\|_A^2 \right].$$

The first factor can be estimated by $\Delta_i \sum_{i=1}^{k_0} \langle \Pi_i M_{i-1} x, M_{i-1} x \rangle_A$ (cf. Lemma 12.35a). Using $\|p_j y^j\|_A^2 \leq \sqrt{\Delta_j} \|y^j\|_W^2$ (cf. Lemma 12.35b) and Cauchy–Schwarz, we obtain that

$$\begin{aligned} \|x - \sum_{j=1}^i p_j y^j\|_A^2 &\leq \left(\|x\|_A + \sum_{j=1}^i \|p_j y^j\|_A \right)^2 \leq \left(\|x\|_A + \sqrt{\Delta} \sum_{j=1}^i \|y^j\|_W \right)^2 \\ &\leq (1 + i\Delta) \left[\|x\|_A^2 + \sum_{j=1}^i \|y^j\|_W^2 \right] \stackrel{(12.33)}{\leq} (1 + i\Delta) (1 + C_1) \|x\|_A^2. \end{aligned}$$

Since $\sum_{i=1}^{k_0} (1 + i\Delta) = k_0 + \Delta k_0 (k_0 + 1)/2$, the summation $\sum_{i=1}^{k_0}$ yields

$$\sum_{i=1}^{k_0} \|x - \sum_{j=1}^i p_j y^j\|_A^2 \leq k_0 (1 + C_1) \left(1 + \frac{\Delta}{2} (k_0 + 1) \right) \|x\|_A^2.$$

Therefore, the final result for Σ_I^2 is

$$\Sigma_I^2 \leq \Delta k_0 (1 + C_1) \left(1 + \frac{\Delta}{2} (k_0 + 1) \right) \|x\|_A^2 \sum_{i=1}^{k_0} \langle \Pi_i M_{i-1} x, M_{i-1} x \rangle_A.$$

Σ_I and Σ_{II} are of the form $\Sigma_I = c_I \omega_I$ and $\Sigma_{II} = c_{II} \omega_{II}$ with

$$\begin{aligned} c_I^2 &= C_{2,k_0}^2 C_1, & \omega_I^2 &= \|x\|_A^2 \sum_{i=k_0+1}^k \langle \Pi_i M_{i-1} x, M_{i-1} x \rangle_A, \\ c_{II}^2 &= \Delta k_0 (1 + C_1) \left[1 + \frac{\Delta}{2} (k_0 + 1) \right], & \omega_{II}^2 &= \|x\|_A^2 \sum_{i=1}^{k_0} \langle \Pi_i M_{i-1} x, M_{i-1} x \rangle_A. \end{aligned}$$

Hence $\Sigma_I + \Sigma_{II} \leq \sqrt{c_I^2 + c_{II}^2} \sqrt{\omega_I^2 + \omega_{II}^2}$, where the second root coincides with $\|x\|_A \sqrt{\sum_{i=1}^k \langle \Pi_i M_{i-1} x, M_{i-1} x \rangle_A}$. The product in (12.43d) contains this factor and $C_2 \sqrt{C_1}$. Replacing $C_2 \sqrt{C_1}$ by $\sqrt{c_I^2 + c_{II}^2}$, we obtain

$$\|x\|_A^2 \leq \left(\sqrt{C_1} + \sqrt{c_I^2 + c_{II}^2} \right) \|x\|_A \sqrt{\sum_{i=1}^k \langle \Pi_i M_{i-1} x, M_{i-1} x \rangle_A}$$

instead of (12.43e). The statement of Corollary 12.34 follows as in (12.43f). \square

12.8 Examples

12.8.1 Schwarz Method With Proper Domain Decomposition

Let the square $\Omega \subset (0, 1) \times (0, 1)$ be partitioned as in Figure 12.5 into disjoint smaller squares Ω_κ ($\kappa \in J$) of the side length H , so that $\overline{\Omega} = \cup_{\kappa \in J} \overline{\Omega}_\kappa$. In order to obtain overlapping subdomains, which are typical for the Schwarz iteration, we enlarge Ω_κ to the square Ω'_κ by elongating the sides in the left and upper direction by the factor αH (cf. Fig. 12.5). Close to the boundary, Ω'_κ has to be restricted to Ω . The grid Ω_h of the Poisson model problem with step size h has to be compatible with the domain decomposition, i.e., H/h and $\alpha H/h$ must be integers.

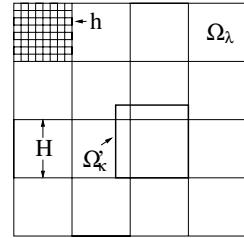


Fig. 12.5 h , H , Ω_κ , Ω'_κ .

The entire index set is $I = \Omega_h$. The subsets I_κ are $I_\kappa := I \cap \Omega'_\kappa$. Vectors $x^\kappa \in X_\kappa$ can be regarded as grid functions on Ω'_κ . For p_κ , we make the trivial choice (12.16a). The matrices $A_\kappa = r_\kappa A p_\kappa$ are again the matrices of the Poisson model case (but on a small square).

The corresponding multiplicative variant of the Schwarz iteration is the classical Schwarz method; however the number of subdomains Ω'_κ generalises from 2 to H^{-2} . For the convergence analysis of the additive variant, one has to determine the quantities γ, Γ in (12.28). Lemma 12.17 can be applied as follows. Decompose the set of squares Ω'_κ analogously to the four-colour numbering (3.13) into four classes J_1, \dots, J_4 . Two squares $\Omega'_\kappa, \Omega'_\lambda$ with $\kappa \neq \lambda$, $\kappa \in J_i$, $\lambda \in J_j$, $i \neq j$ have distance $(1 - \alpha)H > h$, provided that $\alpha < 1 - H/h$. Therefore, κ and λ are not connected (cf. (12.24)); hence, Lemma 12.17 proves $\Gamma = 4$.

The bound γ does prove to be h - but not H -independent: $\gamma = \mathcal{O}(H^2)$ (cf. Widlund [397]). The reason can be understood by Theorem 12.21. A smooth grid function as the constant function $y = 1$ cannot be decomposed into $x = \sum p_\kappa x^\kappa$ with similarly smooth subgrid functions $p_\kappa x^\kappa$. Figure 12.6 shows that the functions x^1 and x^2 have a gradient of size $1/H$ in the overlap region of length H . However, for smooth x and nonsmooth $p_\kappa x^\kappa$, the product $\langle Ax, x \rangle$ would be small compared with $\langle Ap_\kappa x^\kappa, p_\kappa x^\kappa \rangle$, so that $C = 1/\gamma$ in (12.29) becomes large.

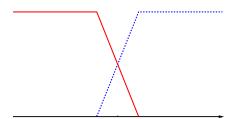


Fig. 12.6 Overlapping functions with sum 1.

The deterioration of condition Γ/γ for decreasing H can be recognised as unavoidable by taking the limit $H = h$. Then the (open) square Ω'_κ contains exactly one grid point. Therefore, the additive Schwarz iteration is the classical pointwise (possibly damped) Jacobi iteration, for which $\Gamma/\gamma = \mathcal{O}(h^{-2}) = \mathcal{O}(H^{-2})$ holds. The same convergence order holds for the multiplicative variant (proper Schwarz iteration), which coincides with the classical pointwise Gauss-Seidel iteration.

12.8.2 Additive Schwarz Iteration with Coarse-Grid Correction

To overcome the unfavourable convergence results in §12.8.1, a coarse-grid correction can be added (cf. Dryja [112] and Dryja–Widlund [114, 115]). The set $J = \{1, \dots, H^{-2}\}$ introduced in §12.8.1 associated with the subsquares Ω'_κ ($\kappa \in J$) is enlarged by the index 0. The new index set $I_0 = \Omega_H$ consists of all (interior) grid points corresponding to the coarser step size H . The prolongation $p_0 : X_0 = \mathbb{R}^{I_0} \rightarrow X = \mathbb{R}^I$ belonging to $0 \in J$ is defined differently from (12.16a). It is sufficient to explain the application of p_0 to unit vectors. Let $e_{\xi,\eta} \in X_0$ be the vector with the value 1 at the grid point $(\xi, \eta) \in \Omega_H = I_0$ and 0 elsewhere. A first approach to p_0 is piecewise bilinear interpolation in $I = \Omega_h$:

$$(p_0 e_{\xi,\eta})(x, y) = (1 - |\xi - x|/H)(1 - |\eta - y|/H)$$

for $(x, y) \in \Omega_h$ with $|\xi - x|, |\eta - y| < H$,

$$(p_0 e_{\xi,\eta})(x, y) = 0 \quad \text{otherwise.}$$

By adding $0 \in J$, according to Lemma 12.17, the previous bound $\Gamma = 4$ can increase at most to $\Gamma = 5$. However, concerning γ , now better estimates can be expected, since instead of the grid function x , only the remainder $x - p_0 x^0$ of interpolation [we choose $x^0(\xi, \eta) := x(\xi, \eta)$ for $(\xi, \eta) \in \Omega_H$] has to be represented by $\sum p_\kappa x^\kappa$. The spectral condition number Γ/γ proves to be $\mathcal{O}(1 + \log(H/h))$. It becomes h - and H -independent if p_0 describes the orthogonal projection (with respect to the Euclidean norm) (cf. Dryja–Widlund [114, 115], Dryja [112]).

A similar idea with nonoverlapping subdomains $\Omega_\kappa = \Omega'_\kappa$ is also the basis of the method of Bramble–Pasciak–Schatz [70, 71, 72, 73].

12.8.3 Formulation in the Case of Galerkin Discretisation

Let the boundary value problem be described in the variational form (E.2) and discretised by (E.5). The subspace $V_n \subset V$ is now denoted by V_h . The bijective relation $v_h = P_h x = \sum_{\alpha \in I} x_\alpha \phi_\alpha$ (ϕ_α : basis functions of V_h ; cf. (E.6)) connects the functions $v_h \in V_h$ and the coefficient vectors $x \in X$.

The subdomain $\Omega_\kappa \subset \Omega$ corresponds to the subspace

$$V_{h,\kappa} := \{v_h \in V_h : v_h(\xi) = 0 \quad \text{for } x \in \Omega \setminus \Omega_\kappa\},$$

i.e., $v_h \in V_{h,\kappa}$ has its support in Ω_κ . For usual finite elements, each component x_α of $x \in X$ corresponds to the function value of v_h in a corresponding *nodal point* $x_\alpha \in \Omega$ (cf. Hackbusch [193]). Choose $I_\kappa := \{\alpha \in I : x_\alpha \in \Omega_\kappa\}$. For a suitable choice of the subdomain Ω_κ , this definition coincides with the index set $I_\kappa := \{\alpha \in I : \text{supp}(\phi_\alpha) \subset \overline{\Omega_\kappa}\}$. Then we have

$$V_{h,\kappa} = \{P_h x : x \in \text{range}(p_\kappa)\} = \{P_h p_\kappa x^\kappa : x^\kappa \in X_\kappa\}, \quad X_\kappa := \mathbb{R}^{I_\kappa}.$$

Since the matrix A is defined by $\langle Ax, y \rangle = a(P_h x, P_h y)$ ($x, y \in X$), the

following representations are equivalent:

$$\begin{aligned} \langle Ap_\kappa x^\kappa, p_\lambda x^\lambda \rangle &= a(P_h p_\kappa x^\kappa, P_h p_\lambda x^\lambda) \quad \text{with } x^\kappa \in X_\kappa \\ &= a(P_h x^{(\kappa)}, P_h p_\lambda x^{(\lambda)}) \quad \text{with } x^{(\kappa)} := p_\kappa x^\kappa \in \text{range}(p_\kappa) \\ &= a(v_h^{(\kappa)}, v_h^{(\lambda)}) \quad \text{with } v_h^{(\kappa)} := P_h x^{(\kappa)} \in V_{h,\kappa}. \end{aligned} \quad (12.44)$$

$P_h p_\kappa : X_\kappa \rightarrow V_{h,\kappa}$ is the bijective mapping that allows us to transfer formulations in the vector spaces X_κ into those in the Galerkin subspaces $V_{h,\kappa}$ and vice versa. Relation (12.44) allows us, e.g., to formulate the strengthened Cauchy–Schwarz inequality (12.31b) in the following equivalent form:

$$|a(v_h, w_h)| \leq \delta \sqrt{a(v_h, v_h)a(w_h, w_h)} \quad \text{for all } \begin{cases} v_h \in V_{h,\kappa}, \\ w_h \in V_{h,\lambda}. \end{cases} \quad (12.45)$$

In general, $a(v, w)$ is an integral $\int_\Omega \dots dx$ over the domain Ω (cf. §11.6.3.2; occasionally, additional boundary integrals may also be involved). Let $a_\tau(v, w) = \int_\tau \dots dx$ be the respective integral over $\tau \subset \Omega$. For a disjoint decomposition of $\Omega = \cup \tau_i$ into subsets τ_i , the following identity is obvious:

$$\sum_i a_{\tau_i}(v, w) = a(v, w) \quad \text{for all } v, w \in V.$$

The following simple lemma is an important tool for many proofs, since it only requires proving (12.45) over the subsets τ_i (e.g., over the triangles of a finite element triangulation).

Lemma 12.38. *If for all i , inequality (12.45) holds with $a_{\tau_i}(\cdot, \cdot)$ instead of $a(\cdot, \cdot)$, then (12.45) is also satisfied for $a(\cdot, \cdot)$ with the same constant δ .*

Proof. The Cauchy–Schwarz inequality yields

$$\begin{aligned} |a(v_h, w_h)| &= \left| \sum_i a_{\tau_i}(v_h, w_h) \right| \leq \sum_i |a_{\tau_i}(v_h, w_h)| \\ &\leq \delta \sum_i \sqrt{a_{\tau_i}(v_h, v_h)a_{\tau_i}(w_h, w_h)} \\ &\leq \delta \left[\sum_i a_{\tau_i}(v_h, v_h) \right]^{\frac{1}{2}} \left[\sum_i a_{\tau_i}(w_h, w_h) \right]^{\frac{1}{2}} = \delta \sqrt{a(v_h, v_h)a(w_h, w_h)}. \end{aligned} \quad \square$$

12.9 Multigrid Iterations as Subspace Decomposition Method

The coarse-grid correction is a common feature of the multigrid methods and the domain decomposition variant in §12.8.2. On the other hand, multigrid iterations can be described and analysed with respect to convergence in such a way that they can immediately be interpreted as a multiplicative Schwarz iteration.

A related statement is that the multigrid iteration can be regarded as a Gauss–Seidel method applied to a semidefinite matrix corresponding to the representation of the finite element matrix by the standard finite element basis replaced with a generating system (frame) (cf. Griebel [169, 168]).

12.9.1 Braess' Analysis without Regularity

The first proof of monotone two-grid convergence $\|M_\ell^{\text{TGM}}\|_A \leq \zeta < 1$ without need of regularity assumptions is due to Braess [58] (cf. §E.5, §11.6.6). Approaches to this proof can be found in [28]. Different from the previous choice, we define the coarse grid of the Poisson model problem as a grid rotated by 45° with step size $h_{\ell-1} = \sqrt{2}h_\ell$. Figure 12.7 shows the grid for a more general domain than the unit square. View the grid Ω_ℓ as a triangular grid. It is well-known that finite-element discretisation with linear triangular elements leads again to the five-point star (1.4a). The coarse-grid equation is the finite-element discretisation corresponding to the thick-lined triangles in Figure 12.7. The canonical prolongation p of the multigrid method is linear interpolation along the hypotenuses of the coarse triangles.

As smoothing we choose the chequer-board Gauss–Seidel iteration \mathcal{S}_ℓ . The ‘red’ points in Ω_h coincide with the coarse-grid points: $\Omega_\ell^r = \Omega_{\ell-1}$, while $\Omega_\ell^b = \Omega_\ell \setminus \Omega_{\ell-1}$ contains the ‘black’ points. Correspondingly, \mathcal{S}_ℓ is the product $\mathcal{S}_\ell^b \circ \mathcal{S}_\ell^r$ of one Gauss–Seidel half-step on the red points followed by a half-step on the black ones. For following considerations, it is sufficient to reduce the smoothing step to the half Gauss–Seidel iteration \mathcal{S}_ℓ^b . The two-grid method then reads as $\Phi_\ell^{\text{TGM}} := \Phi_\ell^{\text{CGC}} \circ \mathcal{S}_\ell^b$, where Φ_ℓ^{CGC} denotes the coarse-grid correction.

For analysis, we need the subspaces

$$\begin{aligned} V_{\ell,1} &:= \{v \in V_\ell : v_\ell(\xi, \eta) = 0 \text{ for all } (\xi, \eta) \in \Omega_\ell^s = \Omega_{\ell-1}\}, \\ V_{\ell,2} &:= V_{\ell-1}, \end{aligned}$$

where V_ℓ is the (finite-element) space of the continuous functions being linear over all triangles of the grids Ω_ℓ . Analogously, functions from $V_{\ell-1} \subset V_\ell$ are linear on the larger triangles of $\Omega_{\ell-1}$. All $v \in V_\ell$ satisfy $v(\xi, \eta) = 0$ for boundary points $(\xi, \eta) \in \partial\Omega$.

Decompose the complete index set $I = \Omega_\ell$ into

$$I_1 := \Omega_\ell \setminus \Omega_{\ell-1}, \quad I_2 := \Omega_{\ell-1}.$$

The vector spaces

$$X_{\ell,1} := \mathbb{K}^{I_1}, \quad X_{\ell,2} := \mathbb{K}^{I_2}$$

correspond to I_1 and I_2 . The second one coincides with the vector space denoted in (11.7) by $X_{\ell-1}$. The prolongations (in the sense of the domain decomposition method) are chosen as

$$\begin{aligned} p_1 : X_{\ell,1} &\rightarrow X_\ell && \text{according to (12.16a),} \\ p_2 = p : X_{\ell,2} &= X_{\ell-1} \rightarrow X_\ell && \text{is the canonical prolongation (11.64).} \end{aligned} \tag{12.46}$$

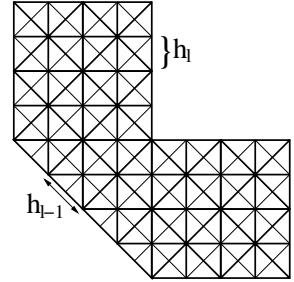


Fig. 12.7 Fine and coarse grids.

Exercise 12.39. (a) Prove for $P = P_\ell$ in (E.6) that

$$V_{\ell,k} = \text{range}(P p_\kappa) \quad \text{for } k = 1, 2.$$

(b) Let $A = A_\ell$ be an arbitrary five-point formula. Prove that the chequer-board Gauss–Seidel half-steps \mathcal{S}_ℓ^b and \mathcal{S}_ℓ^r are projections. If $A > 0$, \mathcal{S}_ℓ^b and \mathcal{S}_ℓ^r are symmetric iterations (cf. (6.19a,b)).

Lemma 12.40. Let $A > 0$. Φ_ℓ^{CGC} and \mathcal{S}_ℓ^b are A -orthogonal projections onto $\text{range}(p_\kappa) \subset X$. In addition, (12.9) holds: $\text{range}(p_1) + \text{range}(p_2) = X_\ell$ and (12.15a): $n_1 + n_2 = n$ with dimensions $n_\kappa := \dim(\text{range}(p_\kappa)) = \#\mathcal{I}_\kappa$, $n := \dim X_\ell$.

Proof. \mathcal{S}_ℓ^b and Φ_ℓ^{CGC} are projections, as can be concluded from Exercise 12.39b and Lemma 11.9 (the assumption (11.20) is satisfied for Galerkin discretisations by Proposition E.16). By Exercise 12.39b and Lemma 11.45 with $\nu = 0$, \mathcal{S}_ℓ^b and Φ_ℓ^{CGC} are symmetric. From Exercise 12.5c, we conclude that \mathcal{S}_ℓ^b and Φ_ℓ^{CGC} are A -orthogonal projections. \square

The results of Lemma 12.40 and the identity $\Phi_\ell^{\text{TGM}} = \Phi_\ell^{\text{CGC}} \circ \mathcal{S}_\ell^b$ prove the following.

Remark 12.41. The two-grid method Φ_ℓ^{TGM} described above is the multiplicative Schwarz iteration characterised by the prolongations (12.46). It corresponds to the case (12.15a) of two disjoint domains.

For convergence analysis, Theorem 12.26 is applicable. The quantity δ of the strengthened Cauchy–Schwarz inequality (12.31b) can be determined by Lemma 12.38. For this purpose, the triangles of the grid $\Omega_{\ell-1}$ are used as subsets τ_i . $v \in V_{\ell,1}$ is a linear function on τ_i , whereas $w \in V_{\ell,2}$ is piecewise linear on the smaller triangles of the grid Ω_ℓ and vanishes at all corners of τ_i . The estimation of the bilinear form $a_{\tau_i}(v, w) = \int_{\tau_i} \langle v, w \rangle \, dx$ yields the bound $\delta \sqrt{a_{\tau_i}(v, v) a_{\tau_i}(w, w)}$ with the constant $\delta = 1/\sqrt{2}$. Lemma 12.38 proves the next theorem.

Theorem 12.42. The two-grid method Φ_ℓ^{TGM} described above converges monotonically with respect to the energy norm with the contraction number

$$\|M_\ell^{\text{TGM}}\|_A \leq \frac{1}{2}.$$

The given two-grid convergence proof requires no regularity assumption. Often, it is viewed as an advantage when convergence can be shown for multigrid-like methods without regularity requirements. On the other hand, one sacrifices a possible increase of efficiency that can be achieved by applying more smoothing steps. As explained in Braess [59] an improved form of the Cauchy–Schwarz inequality (12.45) (thanks to an implicit regularity assumption!) leads to quantitative convergence statements for $\Phi_\ell^{\text{CGC}} \circ (\mathcal{S}_\ell^b \circ \mathcal{S}_\ell^r)^\nu$, demonstrating that the half smoothing step \mathcal{S}_ℓ^b is not optimal with respect to efficiency.

12.9.2 V-Cycle Interpreted as Multiplicative Schwarz Iteration

Let $\ell = \ell_{\max}$ be the maximal level, for which $A = A_\ell$ and $X = X_\ell$ are identified. In the following, we study the V-cycle $\Phi_\ell^V(\nu, 0)$ with ν pre- and no post-smoothing (cf. §11.7.5). The spaces X_i ($0 \leq i \leq \ell$) of dimension n_i introduced in (11.7) for the multigrid method are also taken as subspaces for the domain decomposition. The index set $J = J_\ell$ is $J = \{0, 1, \dots, \ell\}$ so that $k = \ell + 1$ is the number of subspaces.

Let $p : X_{i-1} \rightarrow X_i$ be the multigrid prolongation (11.8). To indicate the involved levels, we call this mapping $p_{i,i-1}$. Their products define

$$\begin{aligned} p_{i,j} &:= p_{i,i-1} \circ \dots \circ p_{j+1,j} && \text{for } 0 \leq j < i \leq \ell, \\ p_{i,i} &= I && \text{for } i = j. \end{aligned} \quad (12.47a)$$

The prolongations needed for domain decomposition are defined as

$$p_i := p_{\ell,i} : X_i \rightarrow X = X_\ell \quad (0 \leq i \leq \ell). \quad (12.47b)$$

In contrast to the previous examples, the ranges of p_i are not disjoint or partially overlapping, but monotonically increasing:

$$\text{range}(p_0) \subset \text{range}(p_1) \subset \dots \subset \text{range}(p_\lambda) = X.$$

Let the coarse-grid matrices be defined by the Galerkin product (11.20). Multiple application of the identity (11.20) yields

$$A_i = p_i^\mathsf{H} A_\ell p_i \quad (0 \leq i \leq \ell; A = A_\ell) \quad (12.48)$$

according to (12.11).

We introduce the auxiliary iteration Ψ_i on $X = X_\ell$ that corresponds to the solution of the i -th subproblem by one V-cycle step:

$$\Psi_i(x, b) = x - p_i N_i p_i^\mathsf{H} (Ax - b).$$

Here, the matrix N_i corresponds to the V-cycle $\Phi_i^V(\nu, 0)$. Using $M_i^V = I - N_i A_i$, we obtain the representation

$$\Psi_i(x, b) = x - p_i (I - M_i^V) A_i^{-1} p_i^\mathsf{H} (A_\ell x - b). \quad (12.49a)$$

For $i = \ell$, we regain the V-cycle at level ℓ because of $p_i = I$:

$$\Psi_\ell = \Phi_\ell^V(\nu, 0). \quad (12.49b)$$

The following presentation simplifies if we do not solve exactly at level $i = 0$ but apply the ν -fold pre-smoothing: $M_0^V = S_0^\nu$. This lead us to

$$\Psi_0(x, b) = x - p_0 (I - S_0^\nu) A_0^{-1} p_0^\mathsf{H} (A_\ell x - b). \quad (12.49c)$$

The following lemma is essential for interpreting the V-cycle as a multiplicative Schwarz iteration.

Lemma 12.43. Assume (12.47a,b), (12.48), and $M_0^V = S_0^\nu$. Then

$$\begin{aligned}\Phi_\ell^V(\nu, 0) &= \tilde{\Phi}_0 \circ \tilde{\Phi}_1 \circ \dots \circ \tilde{\Phi}_\ell, \quad \text{where} \\ \tilde{\Phi}_i(x, b) &:= x - p_i(I - S_i^\nu)A_i^{-1}p_i^H(A_\ell x - b)\end{aligned}$$

is the V-cycle iteration for the i -th subproblem $A_i x^i = c^i := p_i^H(A_\ell x - b)$ using ν smoothing steps.

Proof. Because of (12.49b,c), it is sufficient to prove

$$\Psi_i = \Psi_{i-1} \circ \tilde{\Phi}_i \quad (1 \leq i \leq \ell). \quad (12.50)$$

The iteration matrix of Ψ_i is equal to

$$M_{\Psi,i} = I - p_i(I - M_i^V)A_i^{-1}p_i^H A_\ell = I - p_i A_i^{-1}p_i^H A_\ell + p_i M_i^V A_i^{-1}p_i^H A_\ell.$$

In recursion formula (11.42b):

$$M_i^V = [I - p(I - M_{i-1}^V)A_{i-1}^{-1}rA_i] S_i^\nu,$$

we now have to use $p = p_{i,i-1}$ and $r = p_{i,i-1}^H$. Its insertion into $M_{\Psi,i}^V$ yields

$$M_{\Psi,i} = I - p_i A_i^{-1}p_i^H A_\ell + p_i [I - p(I - M_{i-1}^V)A_{i-1}^{-1}rA_i] S_i^\nu A_i^{-1}p_i^H A_\ell.$$

Noting $p_i p = p_{\ell,i} p_{i,i-1} = p_{i-1}$ and

$$rA_i = p_{i,i-1}^H A_i = p_{i,i-1}^H p_i^H A_\ell p_i = p_{i-1}^H A_\ell p_i,$$

we may write the last term as

$$[I - p_{i-1}(I - M_{i-1}^V)A_{i-1}^{-1}p_{i-1}^H A_\ell] p_i S_i^\nu A_i^{-1}p_i^H A_\ell = M_{\Psi,i-1} p_i S_i^\nu A_i^{-1}p_i^H A_\ell.$$

The projection $I - P_i = I - p_i A_i^{-1}p_i^H A_\ell$ (cf. (12.14a)) satisfies $p_{i-1}^H A_\ell (I - P_i) = 0$, so that $I - P_i = M_{\Psi,i-1}(I - P_i)$. This enables us to formulate the representation

$$\begin{aligned}M_{\Psi,i}^V &= I - P_i + M_{\Psi,i-1} p_i S_i^\nu A_i^{-1}p_i^H A_\ell \\ &= M_{\Psi,i-1}(I - P_i + p_i S_i^\nu A_i^{-1}p_i^H A_\ell) \\ &= M_{\Psi,i-1}(I - p_i A_i^{-1}p_i^H A_\ell + p_i S_i^\nu A_i^{-1}p_i^H A_\ell) \\ &= M_{\Psi,i-1} [I - p_i(I - S_i^\nu)A_i^{-1}p_i^H A_\ell] = M_{\Psi,i-1} M_{\tilde{\Phi},i},\end{aligned}$$

where $M_{\tilde{\Phi},i} = I - p_i(I - S_i^\nu)A_i^{-1}p_i^H A_\ell$ is the iteration matrix of $\tilde{\Phi}_i$. Hence, the product form (12.50) is proved. \square

12.9.3 Proof of V-Cycle Convergence

The interpretation of the V-cycle as a Schwarz iteration is less interesting for the purpose of algorithmic performance than for convergence analysis. We are investigating convergence by using Theorem 12.30. First, we discuss estimates of W_i .

In the case of the model problem, we know that $\|A_i\|_2 \leq Ch_i^{-2}$ holds for a uniform grid size h_i at level i . Assuming the coarsest mesh to be fixed, we have

$$\|A_i\|_2 \leq C' 4^i \quad \text{provided that } h_i = h_0/2^i \quad (\text{cf. (11.5a)}).$$

The same bound holds for a general finite element discretisation provided that in the refinement process from level $i - 1$ to i the element size is halved at most. Without loss of generality, we may assume that only one step of the smoothing procedure is performed (otherwise, redefine \mathcal{S}_i^ν by \mathcal{S}_i); however, \mathcal{S}_i must be positive definite and convergent. The latter property implies (12.21): $A_i \leq \Delta W_i$ with $\Delta < 2$ for the matrix W_i of the third normal form of the smoothing iteration. The upper bound of W_i should be of the same order as the upper bound of A_i discussed above:

$$W_i \leq 4^i C_W p_i^H p_i. \quad (12.51)$$

To obtain the constants C_1 and C_2 in (4.2) and (4.3c), one has to select a suitable decomposition $x = \sum p_i x^i$ with $x^i \in Y_i$ for appropriate subspaces $Y_i \subset X_i$. The orthogonal projection (with respect to the Euclidean scalar product) onto $\text{range}(p_i)$ is

$$Q_i := p_i (p_i^H p_i)^{-1} p_i^H \quad (0 \leq i \leq \ell).$$

Since $p_\ell = I$, we also have $Q_\ell = I$. Following Bramble–Pasciak–Xu [75] we decompose x into

$$x = Q_\ell x = (Q_\ell - Q_{\ell-1})x + (Q_{\ell-1} - Q_{\ell-2})x + \dots + (Q_1 - Q_0)x + Q_0 x.$$

Note that $(Q_i - Q_{i-1})x \in \text{range}(p_i)$. Introducing $Q_{-1} := 0$, we write

$$x = \sum_{i=0}^{\ell} p_i x^i \quad \text{with } p_i x^i := (Q_i - Q_{i-1})x,$$

i.e., $x^i = (p_i^H p_i)^{-1} p_i^H (Q_i - Q_{i-1})x$. These x^i belong to the subspaces

$$Y_i := \text{range}\{(p_i^H p_i)^{-1} p_i^H (Q_i - Q_{i-1})\} = \{x^i \in X_i : Q_{i-1} p_i x^i = 0\}.$$

Oswald [303] proves that the energy norm $\|\cdot\|_A$ is equivalent to the norm defined by

$$\|\|x\|\|^2 := \|Q_0 x\|_A^2 + \sum_{i=0}^{\ell} 4^i \|(Q_i - Q_{i-1})x\|_2^2$$

(see also Dahmen–Kunoth [101]). A compact proof can be found in Bornemann–Yserentant [57]. Applying this statement in particular to $x = p_i x^i$ with $x^i \in Y_i$, we obtain

$$4^i \|p_i x^i\|_2^2 \leq C_E \|x\|_A^2 = C_E \|x^i\|_A^2 \quad (x = p_i x^i, x^i \in Y_i), \quad (12.52)$$

where C_E is the equivalence constant: $\|x\|^2 \leq C_E \|x\|_A^2$. Let $x = \sum p_i x^i$ with $x^i \in Y_i$. Then, inequality (12.51) implies that

$$\sum \|x^i\|_W^2 = \sum \langle W_i x^i, x^i \rangle \leq C_W \sum 4^i \|p_i x^i\|_2^2 \leq C_W \|x\|^2 \leq C_E C_W \|x^i\|_A^2.$$

Hence, condition (12.33) holds with $C_1 = C_E C_W$.

As in Bornemann–Yserentant [57, Lemma 3.3], one can prove the strengthened Cauchy–Schwarz inequality

$$|\langle x^i, x^j \rangle_A| \leq C 2^{\frac{i-j}{2}} \|x^i\|_A 2^j \|p_j x^j\|_2 \quad \text{for } j > i, x^i \in X_i, x^j \in X_j$$

with respect to the subspaces X_i and X_j . Thanks to (12.52), we conclude that

$$|\langle x^i, y^j \rangle_A| \leq C' 2^{(i-j)/2} \|x^i\|_A \|y^j\|_A \quad \text{for } j > i, x^i \in X_i, y^j \in X_j \quad (12.53)$$

with $C' := C C_E^{1/2}$. Since $\|\cdot\|_A^2 \leq \Delta \|\cdot\|_W^2$ (cf. (12.21)), the estimates (12.34a) hold with $\varepsilon_{ij}^{XY} \leq 2^{(i-j)/2}$ for $j > i$. Therefore, the matrix E_{XY} in (12.31a) has the row-sum norm $\|E_{XY}\|_\infty \leq \sum_{\nu=1}^\infty 2^{-\frac{\nu}{2}} = 2 + \sqrt{2}$. Since the same estimate holds for E_{XY}^T , the constant in (12.31b) is bounded by

$$C_2 \leq C'(2 + \sqrt{2}) \quad \text{with } C' \text{ in (12.53)}$$

(cf. Exercise B.21a). Both constants, C_1 and C_2 are independent of the dimension and the number of levels. Hence, Theorem 4.3 proves h -independent convergence of the V-cycle.

The V-cycle $\Phi_\ell^V(\nu, 0)$ ($\nu > 0$) is nonsymmetric. Positive definiteness, however, holds for $\Phi_\ell^V(\nu, \nu)$ (cf. Lemma 11.45). By Exercise 11.60a, this iteration can be interpreted as the product

$$\Phi_\ell^V(\nu, \nu) = \Phi_\ell^V(0, \nu) \circ \Phi_\ell^V(\nu, 0).$$

Since $\Phi_\ell^V(0, \nu)$ is the adjoint of $\Phi_\ell^V(\nu, 0)$ (cf. (11.77a)), we arrive at the product representation

$$\Phi_\ell^V(\nu, \nu) = \tilde{\Phi}_\ell \circ \dots \circ \tilde{\Phi}_1 \circ \tilde{\Phi}_0 \circ \tilde{\Phi}_0 \circ \tilde{\Phi}_1 \circ \dots \circ \tilde{\Phi}_\ell,$$

(here the symmetry properties (11.75a) are assumed). Hence, the symmetric V-cycle $\Phi_\ell^V(\nu, \nu)$ can also be interpreted within the framework of multiplicative Schwarz iterations.

12.9.4 Hierarchical Basis Method

In §12.9.2 we use subspaces which completely overlap with those of lower level: $p_\ell V_\ell \supset p_{\ell-1} V_{\ell-1}$. The method of hierarchical bases adds only those basis functions which are linearly independent of the lower level functions.

Figure 12.8 shows a sequence of refining triangulations for Galerkin discretisation with piecewise linear functions. Let V^h be the space of the piecewise linear functions on the grid Ω_h (right picture in Fig. 12.8). Similarly, V^{2h} corresponds to Ω_{2h} and V^{4h} to Ω_{4h} . The inclusions $V^{4h} \subset V^{2h} \subset V^h$ holds. The Galerkin subspace V^h can be written as the direct sum of V^{2h} and

$$V_2 := \{v \in V^h : v = 0 \text{ at all nodal points of } \Omega_{2h}\};$$

hence, $V^h = V_2 \oplus V^{2h}$ (the direct sum is defined in §A.5.3). The prescribed zeros of $v \in V_2$ are marked in Figure 12.8 by ‘o’. Correspondingly, $V^{2h} = V_1 \oplus V^{4h}$ holds with

$$V_1 := \{v \in V^{2h} : v = 0 \text{ at all nodal points of } \Omega_{4h}\}.$$

With the notation $V_0 := V^{4h}$, we obtain the decomposition

$$V^h = V_0 \oplus V_1 \oplus V_2. \quad (12.54a)$$

The decomposition (12.54a) is nonoverlapping because (12.54a) is a direct sum.

In all spaces V_j ($0 \leq j \leq 2$) we may choose the usual (nodal) basis functions corresponding to the grid size h_j . According to (12.54a), the union of these bases yields a basis of V^h , the *hierarchical basis* (cf. Yserentant [415]). Of course, more general (e.g., irregular) triangulations than in Figure 12.8 and a larger number of grid levels may be used. The largest grid level is denoted by ℓ . In the latter case, (12.54a) becomes $V^h = V_0 \oplus \dots \oplus V_\ell$ with $h = h_\ell$. We write for short $V^i := V^{h_i}$:

$$V^i = V_0 \oplus V_1 \oplus \dots \oplus V_i \quad \text{for } 0 \leq i \leq \ell. \quad (12.54b)$$

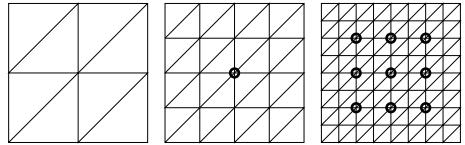


Fig. 12.8 Grid with step widths $4h$, $2h$, h .

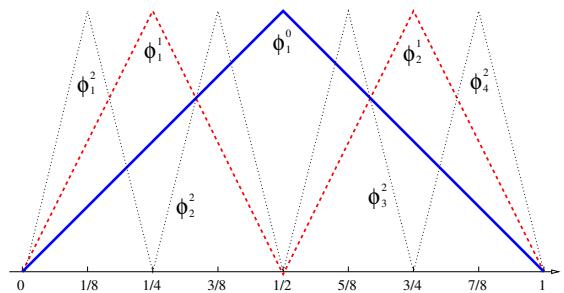


Fig. 12.9 Basis functions on $\Omega = (0, 1)$ of the levels 0, 1, 2.

The largest grid level is denoted by ℓ . In the latter case, (12.54a) becomes $V^h = V_0 \oplus \dots \oplus V_\ell$ with $h = h_\ell$. We write for short $V^i := V^{h_i}$:

The dimension of V_i is the number of nodal points in $I_i := \Omega_i \setminus \Omega_{i-1}$, where $\Omega_i := \Omega_{h_i}$ and $\Omega_{-1} := \emptyset$. These nodal points also serve as indices of the vector $x^i \in X_i = \mathbb{K}^{I_i}$. The coefficient vector x^i represents the finite element function $u \in V_i \subset V^i$ defined by

$$u = \sum_{Q \in I_i} x_Q^i \phi_Q^i,$$

where $\phi_Q^i \in V_i \subset V^i$ is the basic function of level i characterised by $\phi_Q^i(R) = \delta_{QR}$ for all $R \in \Omega_i$. Therefore, the coefficients of x^i are the nodal values of u in the subset I_i : $x_Q^i = u(Q)$ for $Q \in I_i$. Figure 12.9 shows the piecewise linear basis functions of the levels 0 to 2 in the one-dimension case. Here the dimensions are $\#I_0 = 1$, $\#I_1 = 2$, and $\#I_3 = 4$.

We have to distinguish between the representation of functions $u \in V^i$ and their coefficients.

(a) The finite element space V^i can be written as the direct sum of the subspaces V_j ($0 \leq j \leq i$, cf. (12.54b)). Accordingly, $u \in V^i$ has a unique decomposition

$$u = \sum_{j=0}^i u^{(j)} \in V^i \quad \text{with } u^{(j)} \in V_j. \quad (12.55)$$

(b) Using the decomposition $u = \sum_{j=0}^{\ell} u^{(j)}$, we represent each function $u^{(j)} = \sum_{Q \in I_j} x_Q^j \phi_Q^j \in V_j$ by the coefficient vector $x^j = (x_Q^j)_{Q \in I_j} \in X_j$ with $x_Q^j = u^{(j)}(Q)$. Hence, the function is represented by the coefficient vector $x = (x_0, \dots, x_\ell) \in X := X_0 \times X_1 \times \dots \times X_\ell$.

The prolongation $p_i : X_i \rightarrow X$ is defined by

$$(p_i x^i)_Q := \sum_{R \in I_i} x_R^i \phi_R^i(Q) \quad \text{for all } Q \in I_i.$$

The isomorphism $P_h : X^\ell \rightarrow V^h = V^\ell$ in §12.8.3 is given by

$$P_h x = \sum_{Q \in \Omega_\ell} x_Q^\ell \phi_Q^\ell,$$

where $b_Q^\ell \in V^\ell$ is the standard (piecewise linear) basic function of level ℓ characterised by $\phi_Q^\ell(R) = \delta_{QR}$ for all grid points $R \in \Omega_\ell$. Interpolation of $u = P_h x$ with $x = \sum p_j x^j$ ($x^j \in X_j$) at the points of Ω_i is the partial sum $u^i = \sum_{j=0}^i u^{(j)} = P_h \sum_{j=0}^i p_j x^j \in V^i$ (cf. (12.55)).

An important estimate of this interpolant is proved in Yserentant [416]:

$$\left\| \sum_{j=0}^i p_j x^j \right\|_A^2 \leq C_Y (\ell - i + 1) \|x\|_A^2 \quad \begin{cases} \text{for all } x = \sum_{j=0}^{\ell} p_j x^j, \\ x^j \in X_j, \ 0 \leq i \leq \ell. \end{cases} \quad (12.56)$$

Let $s^i := \sum_{j=0}^i p_j x^j$ be the partial sum. Inequality (12.56) implies

$$\|p_j x^j\|_A^2 = \|s^i - s^{i-1}\|_A^2 \leq 2\|s^i\|_A^2 + 2\|s^{i-1}\|_A^2 \leq 2C_Y (2\ell - 2i + 3) \|x\|_A^2$$

for $i > 0$, whereas $\|p_0 x^0\|_A^2 \leq C_Y (\ell + 1) \|x\|_A^2$ is identical to (12.56) for $i = 0$. Summing up all inequalities, we obtain

$$\sum_{j=0}^{\ell} \|p_j x^j\|_A^2 \leq C'_Y \ell^2 \|x\|_A^2 \quad \text{with } C'_Y := 5C_Y.$$

Hence, in the case of an exact solution of the subproblems (12.13), we have proved the inequality (12.29) with $C = C'_Y \ell^2$. The exact solution of $A_i y^i = c^i$ will be maintained only at level $i = 0$ (this corresponds to the exact solution on the coarsest grid in (11.33a)). For $i > 0$, we apply the Jacobi iteration: $W_i := D_i := \text{diag}\{A_i\}$. The matrices A_i and D_i turn out to be spectrally equivalent independently of h_i . In particular, $W_i \leq \text{const} \cdot A_i$ holds. Hence, (12.29) also implies inequality (12.30) with $C = \mathcal{O}(\ell^2)$ and proves $\gamma = \mathcal{O}(\ell^{-2})$ in (12.28) for the additive Schwarz variant. Concerning the estimate above, we can determine Γ in (12.28) by Theorem 12.20. The technique described in Lemma 12.38 leads us to

$$\langle p_i x^i, p_j x^j \rangle_A \leq \varepsilon_{ij} \|p_i x^i\|_A \|p_j x^j\|_A \quad \text{with } \varepsilon_{ij} \leq C_E 2^{-|i-j|/2}.$$

Hence, $\rho(E) \leq \|E\|_\infty$ is bounded by a constant and (12.27) proves $\Gamma = \mathcal{O}(1)$. In the end, we obtain convergence of the additive Schwarz iteration with the rate $1 - \mathcal{O}(1/\ell^2)$.

The additive iteration (cf. Yserentant [416]) described above can be viewed as the blockwise Jacobi iteration for the system associated with the hierarchical basis, where the block diagonal is $D = \text{blockdiag}\{A_0, D_1, \dots, D_\ell\}$. The respective multiplicative variant corresponds to the Gauss–Seidel method for the hierarchical basis system. The multiplicative Schwarz iteration (cf. Bank–Dupont–Yserentant [30]) also converges with the rate $1 - \mathcal{O}(1/\ell^2)$, as one easily derives from (12.35) with $C_1 = \mathcal{O}(\ell^2)$ and $C_2 = \mathcal{O}(1)$ using $Y_j = X_j$.

Concerning algorithmic implementation, in particular, of the fast transformation between the hierarchical basis representation $(x^0, x^1, \dots, x^\ell) \in X_0 \times X_1 \times \dots \times X_\ell$ and the nodal basis representation $x = \sum p_i x^i \in X$, we refer the interested reader to Yserentant [416].

Remark 12.44. The multiplicative Schwarz iteration defined by the hierarchical bases can be viewed as a special multigrid method (V-cycle). The solution of the i -th subproblem $A_i y^i = c^i$ by a secondary iteration step with $W_i := D_i := \text{diag}\{A_i\}$ describes smoothing at level i by a (pointwise) Jacobi step. However, there is a remarkable difference. The smoothing is not performed at all fine grid points of Ω_i , but only at points $(x, y) \in \Omega_i \setminus \Omega_{i-1}$ which do not belong to the coarser grid.

The amount of work per iteration step equals $\mathcal{O}(n)$ even when condition (11.37), $n_{i-1} \leq n_i/C_h$, is violated. This allows local refinements inserting only a few additional fine-grid points $\Omega_i \setminus \Omega_{i-1}$.

The presented convergence results are restricted to the boundary value problem in less than three spatial variables. Otherwise, interpolation has to be replaced by the $L_2(\Omega)$ -orthogonal projection projections onto V^i . The latter method is due to Bramble–Pasciak–Xu [76]. The relation of both methods is discussed by Yserentant [417] (see also Dryja–Widlund [115]).

12.9.5 Multilevel Schwarz Iteration

A characteristic of the Schwarz iteration in §12.8.2 is the coarse-grid correction connected to $I_0 = \Omega_H$. The two-grid situation $\{h, H\}$ can be generalised to the multigrid case $\{h = h_\ell < h_{\ell-1} < \dots < h_0 = H\}$. For this purpose, we rewrite the previous decomposition as $\{I_{0,\ell}, I_{1,\ell}, \dots, I_{k_\ell,\ell}\}$, where $I_{\kappa,\ell}$ ($1 \leq \kappa \leq \ell$) corresponds to the overlapping subdomains Ω'_κ and $I_{0,\ell}$ is related to the coarse grid $\Omega_{h_{\ell-1}}$. The analogous domain decomposition can be repeated for solving the coarse-grid equation: $I_{0,\ell}$ is replaced by $\{I_{0,\ell-1}, I_{1,\ell-1}, \dots, I_{k_{\ell-1},\ell-1}\}$, where $I_{\kappa,\ell-1}$ ($1 \leq \kappa \leq k_{\ell-1}$) represents the overlapping subdomains in the coarse grid and $I_{0,\ell-1} = \Omega_{h_{\ell-2}}$ is related to the next coarser grid. Recursive replacement of $I_{0,\ell-1}, \dots, I_{0,1}$ leads to $\{I_{0,0}, I_{\kappa,\lambda} : 1 \leq \kappa \leq k_\lambda, 1 \leq \lambda \leq \ell\}$ with corresponding prolongations $p_{\kappa,\lambda}$.

In contrast to usual multigrid methods, the multilevel additive Schwarz iteration works in parallel at all levels. For this variant, Dryja–Widlund [115, Theorem 3.2] prove the spectral condition number $\Gamma/\gamma = \mathcal{O}(\ell^2)$. Since usually $\ell = \log(h_\ell)$, this deterioration is rather weak.

12.9.6 Further Approaches

Above the theory of subspace iterations is used to prove convergence of the multigrid iteration (at least in the positive definite case). Since this theory yields convergence results not only for the multiplicative but also for the additive subspace iteration, one can define an additive multigrid version. It is usually termed *BPX method* according to Bramble–Pasciak–Xu [75]. In the symmetric positive case, it has better properties for locally refined grids, but it behaves worse with respect to smoothing. Increasing the number of smoothing steps hardly improves the convergence speed in contrast to the multiplicative version (cf. Theorem 11.14 and Bastian–Hackbusch–Wittum [34]).

Thus far, subspaces (subdomains) have been constructed as proper domain decompositions or decompositions with respect to different grid sizes. A further possibility is the decomposition of a function space by using symmetries (cf. Allgower–Böhmer–Zhen [4]). The prolongations appearing in the frequency decomposition variant of the multigrid method (cf. Hackbusch [186, 190]) can also be used directly as prolongations (12.7) of a domain decomposition method.

Domain decomposition methods for more general (i.e., not positive definite) problems are discussed by Cai–Widlund [92]. A very simple but elegant approach is due to Xu [408]. He uses a product iteration, where the first factor is a coarse-grid correction (11.19) corresponding to the coarsest grid ($\ell = 0$), while the second factor has the iteration matrix $M = I - W^{-1}A$, where W is taken from a (fast) iteration applied to $A_0x = b_0$, where $A_0 > 0$ is the positive definite part of A .

Instead of constructing multigrid methods on the basis of domain decompositions, one can adapt the multigrid iteration to a given decomposition of the domain Ω into overlapping subdomains. The method described in Hackbusch [183, §15.3.3] uses an in parallel executable smoothing procedure, which can be regarded as an approximation of the additive version of the classical Schwarz method. The slowness of the convergence of the Schwarz method presented in §12.7.1 is harmless, since it corresponds to the smooth error components. For a combination of the Schwarz iteration concept with multigrid methods of the second kind, see Hackbusch [177].

Chapter 13

\mathcal{H} -LU Iteration

Abstract The \mathcal{H} -LU iteration is a fast iteration for discretisations of boundary value problems. It even applies to fully populated matrices obtained by the boundary element method. The \mathcal{H} -LU iteration has an almost optimal order of convergence. Section 13.1 describes computing the general LU decomposition by using hierarchical matrices. In the case of sparse matrices, in particular, finite element matrices, the cluster tree can be modified (cf. Section 13.2) so that the corresponding LU decomposition partially preserves sparsity. The \mathcal{H} -LU decomposition is not exact, but the error can be rather small. Correspondingly, the \mathcal{H} -LU iteration described in Section 13.4 is very fast. The variant discussed in §13.4.2 is purely algebraic, i.e., the data needed for the iteration are only based on the underlying matrix. Concerning details about the technique of hierarchical matrices, we refer to Appendix D.

13.1 Approximate LU Decomposition

The LU decomposition based on the technique of hierarchical matrices (cf. Appendix D) yields an approximation of the exact LU factors in $A = LU$ and is called \mathcal{H} -LU decomposition (cf. Hackbusch [198, §7.6 and §7.8] and Grasedyck–Kriemann–Le Borne [165, 166]). The accuracy can be controlled by an appropriate local rank. Therefore the \mathcal{H} -LU factorisation is quite different from the incomplete (ILU) decomposition described in §7.3.

In the positive definite case, LU decomposition can be replaced by Cholesky decomposition $A = LL^\top$. For symmetric but not necessarily positive definite matrices A , we may use the LDL decomposition $A = LDL^\top$. In the following we restrict ourselves to the general LU case. In this section we make no assumption about sparsity of the matrix. The algorithms explained below can also be applied to fully populated matrices, e.g., arising from discretising an integral equation.

In Section 13.2 we shall assume that A is a sparse finite element matrix. Then it is largely possible to preserve sparsity, i.e., the computed factors L and U contain many vanishing matrix blocks.

13.1.1 Triangular Matrices

Triangular matrices can only be defined with respect to a prescribed ordering. The appropriate ordering of the index set I is described in §D.2.1.3. The ordering is consistent with $T(I)$ since each cluster $\tau \in T(I)$ contains consecutive indices:

$$\tau = \{i_{\alpha(\tau)}, i_{\alpha(\tau)+1}, \dots, i_{\beta(\tau)}\}. \quad (13.1)$$

Correspondingly, disjoint clusters τ, σ are ordered: $\tau < \sigma$ holds if $i < j$ for all $i \in \tau$ and $j \in \sigma$. Let $M \in \mathcal{H}(r, P)$ be a hierarchical matrix (cf. Definition D.12). All blocks $b = \tau \times \sigma \in P$ with $\tau \neq \sigma$ are lying completely in the strictly upper (U) or lower triangular part (L). Diagonal blocks $\tau \times \tau \in P$ belong to P^- and the corresponding matrix blocks $M|_{\tau \times \tau}$ are represented as full matrices.

The definition of the format of the hierarchical triangular matrices L and U of the LU decomposition is that they be triangular and hierarchical:

$$L, U \in \mathcal{H}(r, P), \quad \begin{cases} L_{i_\alpha i_\beta} = 0 & \text{for } \alpha < \beta, \\ L_{i_\alpha i_\alpha} = 1 & \text{for } 1 \leq \alpha \leq \#I, \\ U_{i_\alpha i_\beta} = 0 & \text{for } \alpha > \beta. \end{cases} \quad (13.2)$$

Solvability of a system $LUX = b$ requires that $U_{i_\alpha i_\alpha} \neq 0$ for all i_α .

The triangular matrices can also be replaced by *block-triangular matrices*:

off-diagonal blocks: $L|_{\tau \times \sigma} = O$ for $\tau < \sigma$ and $U|_{\tau \times \sigma} = O$ for $\tau > \sigma$,

diagonal blocks: $L|_{\tau \times \tau} = I$ and $U|_{\tau \times \tau} \in \mathcal{F}(\tau \times \tau)$ for $\tau \times \tau \in P$.

Concerning the restriction $\cdot|_b$ to a block b see (A.8b) or Notation D.6. Note that $U|_{\tau \times \tau}$ is no longer triangular. The block-triangle decomposition has the advantage that it may be well defined even if the standard LU decomposition does not exist.

13.1.2 Solution of $LUX = b$

Given a factorisation $A = LU$, the system $Ax = b$ is solved in two stages: the equation $Ly = b$ is treated by the procedure *Forward_Substitution* and $Ux = y$ by *Backward_Substitution*. These steps can easily be formulated for hierarchical matrices and performed exactly. The procedure *Forward_Substitution*(L, τ, y, b) yields the (exact) solution $y|_\tau$ of $L|_{\tau \times \tau}y|_\tau = b|_\tau$. To solve $Ly = b$, one has to call *Forward_Substitution*(L, I, y, b) with $\tau = I$ (the input vector b is overwritten).

```

procedure Forward_Substitution( $L, \tau, y, b$ );
if  $\tau \times \tau \in P$  then for  $j := \alpha(\tau)$  to  $\beta(\tau)$  do (cf. (13.1))
    begin  $y_j := b_j$ ; for  $i := j+1$  to  $\beta(\tau)$  do  $b_i := b_i - L_{ij}y_j$  end
    else for  $j := 1$  to  $\#S(\tau)$  do
        begin Forward_Substitution( $L, \tau[j], y, b$ );
            for  $i := j + 1$  to  $\#S(\tau)$  do  $b|_{\tau[i]} := b|_{\tau[i]} - L|_{\tau[i] \times \tau[j]} \cdot y|_{\tau[j]}$ 
        end;

```

The requirements for the parameters are: $\tau \in T(I \times I, P)$, $y, b \in \mathbb{K}^I$, and L satisfies (13.2) with $P \subset T(I \times I)$. In line 6, $\tau[1], \dots, \tau[\#S(\tau)]$ is an enumeration of the sons of τ .

The procedure *Backward_Substitution* for solving $Ux = y$ is quite similar. U, τ, y are input parameters, while x is the output. The vector y is overwritten.

```
procedure Backward_Substitution( $U, \tau, x, y$ );
if  $\tau \times \tau \in P$  then for  $j := \beta(\tau)$  downto  $\alpha(\tau)$  do
    begin  $x_j := y_j/U_{jj}$ ;
        for  $i := \alpha(\tau)$  to  $j - 1$  do  $y_i := y_i - U_{ij}x_j$ 
    end
else for  $j := \#S(\tau)$  downto 1 do
    begin Backward_Substitution( $U, \tau[j], x, y$ );
        for  $i := 1$  to  $j - 1$  do  $y|_{\tau[i]} := y|_{\tau[i]} - U|_{\tau[i] \times \tau[j]} \cdot x|_{\tau[j]}$ 
    end;
```

The complete solution of $LUX = b$ uses

```
procedure Solve_LU( $L, U, I, x, b$ ); { $L, U, I, b$  input;  $x$  output}
begin  $x := b$ ;
    Forward_Substitution( $L, I, x, x$ );
    Backward_Substitution( $U, I, x, x$ )
end;
```

Formulating the block version is left to the reader as an exercise. Since then the diagonal matrix blocks $U|_{\tau \times \tau}$ ($\tau \times \tau \in P$) must be inverted during the solution of $UX = y$, the best approach is to invert $U|_{\tau \times \tau}$ immediately after constructing U . Then the backward substitution procedure can multiply by the precomputed inverse stored in $U|_{\tau \times \tau}$.

Finally, we need an algorithm for solving $x^T U = y^T$. This equation is identical to $Lx = y$ with $L := U^T$; however, in this case, the lower triangular matrix L is not normed. The corresponding procedure is left to the reader:

```
procedure Forward_SubstitutionT( $U, \tau, x, y$ ); {solving  $x^T U = y^T$ }.
```

13.1.3 Matrix-Valued Solutions of $LX = Z$ and $XU = Z$

The matrix $L \in \mathcal{H}(r, P)$ with $P \subset T(I \times I)$ is a lower triangular matrix (cf. (13.2)). Let $X, Z \in \mathcal{H}(r, P')$ be rectangular hierarchical matrices corresponding to a partition $P' \subset T(I \times J)$. The index set I is the same as for $L \in \mathbb{K}^{I \times I}$. We want to solve the matrix equation

$$LX = Z$$

in $\mathbb{K}^{I \times J}$, which represents $\#J$ simultaneous equations of the form $Lx = z$. The following procedure *Forward_M* solves $L|_{\tau \times \tau} X|_{\tau \times \sigma} = Z|_{\tau \times \sigma}$ for the blocks $\tau \times \tau \in T(I \times I, P)$ and $\tau \times \sigma \in T(I \times J, P')$. The complete system $LX = Z$ in $I \times J$ is solved by *Forward_M*(L, X, Z, I, J). By $X_{\tau,j}$ we denote the j -th columns of $X \in \mathbb{K}^{\tau \times \sigma}$, i.e., $X_{\tau,j} = (X_{ij})_{i \in \tau}$.

```

procedure Forward_M( $L, X, Z, \tau, \sigma$ ) ;
  if  $\tau \times \sigma \in P^-$  then {column-wise forward substitution}
    for all  $j \in \sigma$  do Forward_Substitution( $L, \tau, X_{\tau,j}, Z_{\tau,j}$ )
  else if  $\tau \times \sigma \in P^+$  then
    begin {let  $Z|_{\tau \times \sigma} = AB^\top$  according to (D.1) with  $A \in \mathbb{K}^{\tau \times \{1, \dots, r\}}$ }
      for  $j = 1$  to  $r$  do Forward_Substitution( $L, \tau, A'_{\tau,j}, A_{\tau,j}$ );
       $X|_{\tau \times \sigma} :=$  rank- $r$  representation by  $A'B^\top$ 
    end else
    for  $i = 1$  to  $\#S(\tau)$  do for  $\sigma' \in S(\sigma)$  do
      begin Forward_M( $L, X, Z, \tau[i], \sigma'$ );
        for  $j = i + 1$  to  $\#S(\tau)$  do { $\ominus, \odot$ : operations with truncation}
         $Z|_{\tau[j] \times \sigma'} := Z|_{\tau[j] \times \sigma'} \ominus L|_{\tau[j] \times \tau[i]} \odot X|_{\tau[i] \times \sigma'}$ 
    end;

```

In the standard case of $\#S(\sigma) = 2$, the problem

$$L|_{\tau \times \tau} X|_{\tau \times \sigma} = Z|_{\tau \times \sigma}$$

has the block structure

$$\begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix}$$

with

$$L_{ij} = L|_{\tau[i] \times \tau[j]}, \quad X_{ij} = X|_{\tau[i] \times \sigma[j]}, \quad Z_{ij} = Z|_{\tau[i] \times \sigma[j]}.$$

The equations $L_{11}X_{11} = Z_{11}$ and $L_{11}X_{12} = Z_{12}$ of the first block row are solved for $i = 1$ by the call of *Forward_M* in line 10, whereas the remaining equations $L_{21}X_{11} + L_{22}X_{21} = Z_{21}$ of the first block row and $L_{21}X_{12} + L_{22}X_{22} = Z_{22}$ of the second one are reformulated as

$$L_{22}X_{21} = Z'_{21} := Z_{21} - L_{21}X_{11}, \quad L_{22}X_{22} = Z'_{22} := Z_{22} - L_{21}X_{12}$$

in line 12 and are solved for $i = 2$ in line 10 with respect to X_{21}, X_{22} .

For solving the equation $XU = Z$ with an upper triangular hierarchical matrix U and an unknown matrix X left of U , we use a similar procedure involving the procedure *Forward_SubstitutionT* defined above:

procedure *ForwardT_M*(U, X, Z, τ, σ) ; (13.3)

(details in [198, (7.33b)]).

13.1.4 Generation of the LU Decomposition

It remains to describe the generation of the hierarchical LU factors in

$$A = LU \in \mathbb{K}^{I \times I}.$$

To simplify the explanation we assume that $\#S(I) = 2$. Then the matrices in $A = LU$ have the structure

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} L_{11} & O \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} \\ O & U_{22} \end{bmatrix}. \quad (13.4)$$

This leads to the four subtasks

- (i) compute L_{11} and U_{11} as factors of the LU decomposition of A_{11} ,
- (ii) compute U_{12} from $L_{11}U_{12} = A_{12}$,
- (iii) compute L_{21} from $L_{21}U_{11} = A_{21}$,
- (iv) compute L_{22} and U_{22} as LU decomposition of $L_{22}U_{22} = A_{22} - L_{21}U_{12}$.

Problem (ii) is solved by the procedure *Forward_M*($L_{11}, U_{12}, A_{12}, \tau_1, \tau_2$), whereas for problem (iii) we use the procedure *ForwardT_M* in (13.3). The right-hand side in

$$L_{22}U_{22} = A_{22} - L_{21}U_{12}$$

can be computed by the usual formatted multiplication.

We still have to determine the LU factors of $L_{11}U_{11} = \dots$ and $L_{22}U_{22} = \dots$. This defines a recursion, which at the leaves is defined by the usual LU decomposition of full matrices.

The call of *LU-Decomposition*(L, U, A, I) yields the desired LU factors of A . More generally, the procedure *LU-Decomposition*(L, U, A, τ) solves the problem $L|_{\tau \times \tau} U|_{\tau \times \tau} = A|_{\tau \times \tau}$ for $\tau \in T(I \times I, P)$.

```
procedure LU-Decomposition( $L, U, A, \tau$ );
if  $\tau \times \tau \in P$  then compute  $L|_{\tau \times \tau}$  and  $U|_{\tau \times \tau}$  as LU factors of  $A|_{\tau \times \tau}$ 
else for  $i = 1$  to  $\#S(\tau)$  do
begin LU-Decomposition( $L, U, A, \tau[i]$ );
    for  $j = i + 1$  to  $\#S(\tau)$  do
        begin ForwardT_M( $U, L, A, \tau[j], \tau[i]$ );
            Forward_M( $L, U, A, \tau[i], \tau[j]$ );
            for  $r = i + 1$  to  $\#S(\tau)$  do
                 $A|_{\tau[j] \times \tau[r]} := A|_{\tau[j] \times \tau[r]} \ominus L|_{\tau[j] \times \tau[i]} \odot U|_{\tau[i] \times \tau[r]}$ 
    end end;
```

(13.5)

The sons of $S(\tau)$ are denoted by $\tau[1], \dots, \tau[\#S(\tau)]$.

13.1.5 Cost of the \mathcal{H} -LU Decomposition

Because of the triangular structure, the two matrices L and U need not more storage than a usual hierarchical matrix:

$$S_{LU}(r, P) = S_{\mathcal{H}}(r, P),$$

where $S_{\mathcal{H}}(r, P)$ is given in Lemma D.17.

As in Lemma D.18, one verifies that the cost of $Forward_Substitution(L, I, y, b)$ can be estimated by the double storage cost of L . An analogous result holds for $Backward_Substitution(U, \tau, x, y)$. Together we obtain

$$N_{LU}(r, P) \leq 2S_{\mathcal{H}}(r, P).$$

Comparing the costs for solving both systems $LX = Z$ and $XU = Z$ with a standard multiplication of hierarchical matrices, we obtain

$$N_{Forward_M}(r, P) + N_{ForwardT_M}(r, P) \leq N_{MM}(P, r, r)$$

with $N_{MM}(P, r, r)$ in (D.15). Generating the LU decomposition by the procedure in (13.5) also does not require more operations than matrix-matrix multiplication:

$$N_{LU \text{ decomposition}}(r, P) \leq N_{MM}(P, r, r).$$

13.2 \mathcal{H} -LU Decomposition for Sparse Matrices

13.2.1 Finite Element Matrices

Finite element matrices are *sparse* in the classical sense. They can exactly be transferred into the $\mathcal{H}(r, P)$ format. This transfer is required if we want to apply hierarchical matrix operations other than matrix-vector multiplication.

Lemma 13.1. *Let $\mathcal{H}(r, P) \subset \mathbb{K}^{I \times I}$ be an arbitrary hierarchical format, and P an admissible partition. Let $\text{dist}(\tau, \sigma)$ be defined by (D.8) and (D.9b). Then any finite element matrix belongs to $\mathcal{H}(r, P)$ for all $r \in \mathbb{N}_0$.*

Proof. For an admissible block $b = \tau \times \sigma \in P^+$, the indices $i \in \tau$ and $j \in \sigma$ belong to basis functions with disjoint supports X_i and X_j . Hence the finite element matrix restricted to b is a zero block and, therefore, belongs to $\mathcal{R}_r(b)$. \square

Modern direct solvers for sparse systems apply sophisticated algorithms to minimise the fill-in during the LU decomposition. Formally, this means finding a permutation P , so that the LU decomposition of PAP^\top (without pivoting) is sparser than for A . For instance, one may try to minimise the band width since fill-in of the LU factors occurs only within the band (cf. [314, §3.9.1]). Similarly, we may try to optimise the \mathcal{H} -LU decomposition for sparse matrices.

The precise conditions concerning the sparsity pattern will be discussed in §13.2.2. Let I be the index set in $A \in \mathbb{K}^{I \times I}$. The ordering of the index set, determining the LU decomposition, is derived from the cluster tree $T(I)$ (cf. (13.1)). Therefore, alternative permutations require alternative cluster trees. Such a cluster tree will be introduced in §13.2.3. The following LU variants are based on the articles Le Borne–Grasedyck–Kriemann [259] and Grasedyck–Kriemann–Le Borne [166].

The inverse of a sparse finite element matrix is a fully populated matrix. It is shown in Bebendorf–Hackbusch [39], Faustmann–Melenk–Praetorius [129], and Faustmann [128] that the inverse matrix can be well approximated by the format $\mathcal{H}(r, P)$. The involved truncation error decreases exponentially with r . These results can be transferred to the LU decomposition; i.e., the factors L and U are also well approximated by hierarchical triangular matrices in $\mathcal{H}(r, P)$ (cf. [198, §9.2.8] or Grasedyck–Kriemann–Le Borne [166], Faustmann [128, §6]). A similar result holds for the inverse and the LU factors of matrices arising from the boundary element method (cf. Faustmann–Melenk–Praetorius [130]).

13.2.2 Separability of the Matrix

Sparsity alone is not sufficient for our purpose. In addition, we need the following condition. The index set I can be decomposed disjointly:

$$I = I_1 \dot{\cup} I_2 \dot{\cup} I_s \quad \text{with} \quad \#I_1 \approx \#I_2, \quad \#I_s \ll \#I, \quad (13.6a)$$

so that the matrix A , which we want to partition, has the following block structure:

$$A = \begin{array}{c|c|c} & \overbrace{I_1}^{} & \overbrace{I_2}^{} & I_s \\ \hline & \left\{ \begin{array}{|c|c|} \hline & A_{11} & O \\ \hline A_{11} & O & \\ \hline \end{array} \right\} & \left\{ \begin{array}{|c|c|} \hline & A_{12} & \\ \hline A_{21} & O & \\ \hline \end{array} \right\} & A_{1s} \\ \hline & \left\{ \begin{array}{|c|c|} \hline & O & A_{22} \\ \hline O & A_{22} & \\ \hline \end{array} \right\} & & A_{2s} \\ \hline & & & A_{ss} \\ \hline & A_{s1} & A_{s2} & \\ \hline \end{array} \quad (13.6b)$$

The index set I_s is called the *separator* since $A|_{(I \setminus I_s) \times (I \setminus I_s)}$ is decomposed into the matrix blocks A_{11} and A_{22} ; the off-diagonal blocks A_{12} and A_{21} contain only zero entries.

Condition $\#I_1 \approx \#I_2$ in (13.6a) ensures that (i) A_{11} and A_{22} be similar in size, (ii) the zero blocks are large.

Condition $\#I_s \ll \#I$ requires the separator to be comparably small. More quantitative statements will follow.

The requirements (13.6a,b) can easily be formulated by the matrix graph $G(A)$ (cf. §C.2). I is the vertex set. There must be a (small) subset I_s so that the graph without the I_s -vertices and corresponding edges disaggregates into two unconnected subgraphs with the vertex sets I_1 and I_2 (cf. Fig. 13.1).

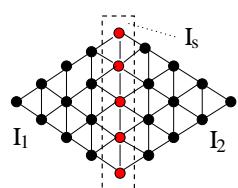


Fig. 13.1 Matrix graph separated by I_s .

The last formulation yields a sufficient condition for (13.6a,b). If $G(A)$ is a planar graph, a linear subgraph—as in Figure 13.1—is a sufficient choice of the separator. Planar graphs are, e.g., obtained by discretising two-dimensional boundary value problems by a standard difference method or by piecewise linear finite elements. If $n = \#I$ is the problem size, one expects a separator of the cardinality $\#I_s = \mathcal{O}(\sqrt{n})$, while $\#I_1, \#I_2 \approx n/2$. In the case of finite elements in a domain $\Omega \subset \mathbb{R}^2$, one determines a curve $\gamma \subset \bar{\Omega}$ with endpoints on $\Gamma = \partial\Omega$, consisting of edges belonging to the finite element triangulation (cf. Fig. 13.2). The indices $i \in I_s$ are associated with the nodal points in γ . The vertices left or right of γ form the respective sets I_1 or I_2 . If $i_1 \in I_1$ and $i_2 \in I_2$, supports of the basis functions ϕ_{i_1} and ϕ_{i_2} lie on different sides of γ and can overlap at most by their boundaries. This implies that $A_{i_1 i_2} = 0$, as required in (13.6b).

The example of a boundary value problem in Ω shows that the method can be iterated: γ divides Ω into subdomains Ω_1 and Ω_2 , and the submatrices A_{11} and A_{22} in (13.6b) belong to boundary value problems in these subdomains; hence, they are of the same kind as the original matrix.

The latter observation leads to the final assumption:

$$\text{The submatrices } A_{ii} := A|_{I_i \times I_i} \text{ (} i = 1, 2 \text{)} \text{ must again satisfy (13.6a–c) or be sufficiently small.} \quad (13.6c)$$

This requirement ensures that the partition can be continued recursively (Fig. 13.3 shows the result after two partitions). Obviously, the condition $\#I_s \ll \#I$ is vague. In particular, the symbol \ll is meaningless if $\#I$ is not large. In this case, the recursion terminates since ‘sufficiently small’ submatrices occur (cf. (13.6c)).

The partition (13.6a,b) is well known as the *dissection* method introduced by George [149]. It also corresponds to the (iterated form of the) domain decomposition method.

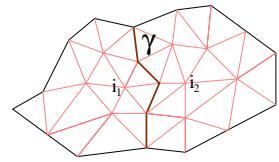


Fig. 13.2 Domain decomposition by γ .

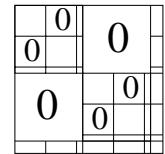


Fig. 13.3 Twofold decomposition.

13.2.3 Construction of the Cluster Tree

The partition of the index set I into the three subsets in (13.6a) can easily be performed. A variant of the partition in §D.2.1.2 works as follows. Assume that the indices $i \in I$ are again associated with nodal points $\xi_i \in \mathbb{K}^d$. Let the partition of the cuboid (minimal box) yield the binary decomposition of I into \hat{I}_1 and \hat{I}_2 . The first set $I_1 := \hat{I}_1$ remains unchanged, while the second is split again:

$$I_s := \{i \in \hat{I}_2 : \text{there are } A_{ij} \neq 0 \text{ or } A_{ji} \neq 0 \text{ for some } j \in I_1\}, \quad I_2 := \hat{I}_2 \setminus I_s.$$

Obviously, the partition into I_1, I_2, I_s satisfies condition (13.6a).

In principle, this decomposition algorithm could be continued recursively. The result would be a ternary tree $T(I)$. However, this procedure is not optimal. The reason are the different characters of the three subsets I_1 , I_2 , and I_s . For an illustration, assume the two-dimensional case $\Omega \subset \mathbb{R}^2$. The first two sets I_1 and I_2 correspond to the (two-dimensional) subdomains Ω_1 and Ω_2 (cf. Fig. 13.2), whereas the indices of I_s are vertices of the (one-dimensional) curve γ . We recall the bisection of the bounding box in §D.2.1.2. d bisection steps of a d -dimensional cuboid lead to 2^d subcuboids of half the size. This means that the diameter of an index set belonging to subdomains of Ω is reduced by about $1/\sqrt{2}$, whereas the diameter of an index set belonging to the (one-dimensional) separator γ is reduced by $1/2$. Therefore, with increasing level ℓ , the subset $T^{(\ell)}(I)$ defined in (D.7) contains index sets exhibiting increasingly different sizes. Therefore the block cluster tree contains rather flat blocks $\sigma \times \tau$.

The following modification (here explained and illustrated for $d = 2$) avoids a systematic distortion of the cluster sizes in $T^{(\ell)}(I)$. The cluster set $T(I)$ is divided into ‘two-dimensional’ clusters $T_d(I)$ and ‘one-dimensional’ clusters $T_{d-1}(I)$. Their definition is given by

- (a) $I \in T_d(I)$,
- (b) if $\tau \in T_d(I)$, the sons τ_1, τ_2 belong to $T_d(I)$, whereas τ_s belongs to $T_{d-1}(I)$,
- (c) all successors of $\tau \in T_{d-1}(I)$ belong to $T_{d-1}(I)$.

In Figure 13.4, the rectangles with dashed sides correspond to clusters in $T_{d-1}(I)$, the other rectangles correspond to $T_d(I)$.

The decomposition rules are as follows:

- (a) A cluster $\tau \in T_d(I)$ is always decomposed into three parts. Since, in the case of an LU decomposition, an ordering of the sons of τ is required, we define the order as follows: First, the sons $\tau_1, \tau_2 \in S(\tau) \cap T_d(I)$ are arranged in arbitrary order (edges depicted by solid lines in Fig. 13.4), then the son $\tau_s \in S(\tau) \cap T_{d-1}(I)$ follows (dashed line).
- (b) The treatment of a cluster $\tau \in T_{d-1}(I)$ depends on its graph distance to the next $T_d(I)$ -predecessor. For this purpose, we introduce

$$\varkappa(\tau) := \min\{\text{level}(\tau) - \text{level}(\tau') : \tau' \in T_d(I) \text{ predecessor of } \tau\}.$$

- (ba) If $\varkappa(\tau)$ is odd, τ remains unchanged (dotted edge in Fig. 13.4).
- (bb) If $\varkappa(\tau)$ is even, τ is decomposed in a binary¹ way according to §D.2.1.2 (broken-dotted edges in Fig. 13.4).

These rules guarantee that all clusters in $T^{(\ell)}(I)$ have successors at level $\ell + 2$ with a diameter of about half the size. For $d = 3$, one has to modify these rules suitably.

¹ Here a ternary splitting does not make sense.

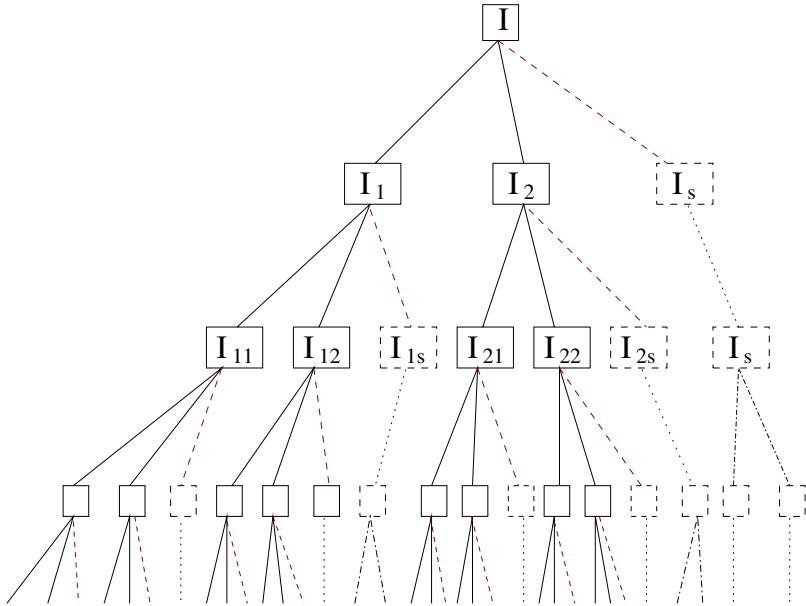


Fig. 13.4 Cluster tree $T(I)$.

The corresponding block cluster tree $T(I \times I)$ is obtained as in Definition D.8. A block partition of depth $L = 2$ is shown in Figure 13.3.

13.2.4 Application to Inversion

The inversion algorithm in §D.3.6 has an intrinsic disadvantage concerning its parallel treatment. The inversion of $M|_{\tau \times \tau}$ has to wait until the inversions in the blocks $\tau' \times \tau'$ ($\tau' \in S(\tau)$) are performed. This requires a sequential computing². Also in the case of partition (13.6b), one has first to invert the diagonal blocks A_{11} and A_{22} before the Schur complement in $I_s \times I_s$ can be formed and inverted, but (i) the inverses of A_{11} and A_{22} can be computed in parallel and (ii) the computations in $I_s \times I_s$ are significantly cheaper than the inversions of A_{11} and A_{22} because of $\#I_s \ll \#A_{11}$.

The algorithm is still sequential in the level-number: The inversion of $M|_{\tau \times \tau}$ can take place as soon as the inversions in $\tau' \times \tau'$ ($\tau' \in S(\tau)$) are performed.

More details about this method can be found in Hackbusch [192] and Hackbusch–Khoromskij–Kriemann [202]. Parallel \mathcal{H} -matrix implementations are discussed by Kriemann [245].

² Of course, the arising matrix-matrix multiplications and additions can be parallelised.

13.2.5 Admissibility Condition

The zero blocks in (13.6b) are characterised by

$$\tau' \times \tau'' \text{ with } \tau' \neq \tau'' \text{ and } \tau', \tau'' \in S(\tau) \cap T_d(I) \text{ for some } \tau \in T_d(I). \quad (13.7)$$

The blocks $b = \tau' \times \tau''$ are not admissible in the sense of Definition D.11, since the support sets $X_{\tau'}$ and $X_{\tau''}$ touch at the separating line γ , and therefore $\text{dist}(\tau', \tau'')$ vanishes. Nevertheless, it does not make sense to decompose b again. Therefore the admissibility condition adm^* in (D.11) is modified as follows:

$$\text{adm}^{**}(\tau' \times \tau'') := [\text{adm}^*(\tau' \times \tau'') \text{ or } \tau' \times \tau'' \text{ satisfies (13.7)}].$$

The minimal admissible partition $P \subset T(I \times I)$ is now defined in (D.12) with adm^* replaced by adm^{**} . So far, we divided P into the near- and far-field: $P = P^- \dot{\cup} P^+$. Now a ternary partition is appropriate: $P = P^0 \dot{\cup} P^- \dot{\cup} P^+$ with $P^0 := \{b \in P \text{ satisfies (13.7)}\}$, while $P \setminus P^0$ is split into $P^- \dot{\cup} P^+$ as before.

13.2.6 LU Decomposition

The algorithm in §13.1 can be applied without changes. The advantage of the new cluster tree $T(I)$ can be seen from the following statement.

Remark 13.2. Let the matrix $A \in \mathcal{H}(r, P)$ satisfy $A|_b = 0$ for all $b \in P^0$. Then the approximate LU decomposition according to (13.5) yields factors $L, U \in \mathcal{H}(r, P)$ satisfying again $L|_b = U|_b = 0$ for $b \in P^0$ (cf. Fig. 13.5).

Detailed numerical results and comparisons with other algorithms can be found in Grasedyck–Hackbusch–Kriemann [164].

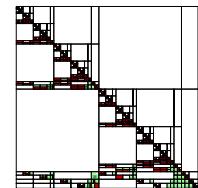


Fig. 13.5 Factor U ; white blocks are zero.

13.3 UL Decomposition of the Inverse Matrix

If a regular matrix A possesses an LU decomposition $A = LU$, A^{-1} can also be decomposed into $U'L'$ with $L' := L^{-1}$ and $U' := U^{-1}$ and vice versa. Here we use that the inverse of a (normed) triangular matrix is again (normed) triangular. Note the different ordering of the factors in $A^{-1} = U'L'$: the first matrix is the upper triangular, while the second one is the normed lower matrix.

Remark 13.3. The standard forward and backward substitution in $x \mapsto U^{-1}L^{-1}x$ avoids inversion, but is mainly sequential. In contrast to this, matrix-vector multiplications in $x \mapsto U'L'x$ can be parallelised much better (see Kriemann–Le Borne [246, Tables 3 and 4]).

Similar to the discussion of (13.4), the factors in $A^{-1} = U'L'$ can be determined from $L'AU' = I$ (cf. [198, §7.6.5] and Kriemann–Le Borne [246]).

13.4 \mathcal{H} -LU Iteration

13.4.1 General Construction

The \mathcal{H} -matrix technique may be considered as a direct method with the difference that the error is not characterised by the machine precision, but by the accuracy of the \mathcal{H} -matrix computation. Note that the \mathcal{H} -matrix accuracy can be adjusted to the discretisation error.

In fact, there is a smooth transition from a direct method to an iterative one. We recall that even the Gauss elimination becomes an iteration when it is re-iterated (cf. Skeel [341], Björck [48, §1.4.6]).

The \mathcal{H} -LU decomposition $A \approx LU$ induces the iteration $\Phi_{\mathcal{H}\text{-LU}}$:

$$x^{m+1} = x^m - W^{-1}(Ax^m - b) \quad \text{with } W = LU. \quad (13.8)$$

The properties of the method are collected in the next remark.

Remark 13.4. (a) Since an LU decomposition does not exist for any regular matrix, the existence of the \mathcal{H} -LU decomposition is not guaranteed in general. If the hierarchical LU decomposition is successful, the involved rank controls the error $I - W^{-1}A$.

(b) The inversion of $W = LU$ uses the procedures in §13.1.2, which are very fast.

(c) The data required to determine W are the matrix A including the geometric information about the nodal points ξ_i ($i \in I$). In the case of a sparse matrix, the geometric data can be replaced by the graph $G(A)$. In that case, the method is algebraic (cf. Definition 2.2b).

(d) If $A > 0$, also $W > 0$ is expected (here Cholesky decomposition should be used). There are strategies to ensure $W > 0$ in spite of truncation errors (cf. [198, §6.8.2]). Then the iteration is positive definite: $\Phi_{\mathcal{H}\text{-LU}} \in \mathcal{L}_{\text{pos}}$.

The statements $W \approx A$ or $N = W^{-1} \approx A^{-1}$ can be made more precise by the error estimate

$$\|I - NA\|_2 \leq \varepsilon < 1. \quad (13.9a)$$

Inequality (13.9a) implies the corresponding estimate with respect to the spectral radius:

$$\rho(I - NA) \leq \varepsilon < 1. \quad (13.9b)$$

If, e.g., $\varepsilon = \frac{1}{10}$ in (13.9a), each step of the iteration (13.8) improves the result by one decimal. $\varepsilon = \frac{1}{10}$ is already considered as fast convergence, whereas N with $\varepsilon = \frac{1}{10}$ in (13.9a) may be still regarded as a rough approximation of the inverse.

An alternative to (13.9a) is

$$\|I - NA\|_A = \|I - A^{1/2}NA^{1/2}\|_2 \leq \varepsilon < 1 \quad (13.9c)$$

for positive definite A . (13.9c) also implies (13.9b). The contraction properties (13.9a) or (13.9c) are very important if only a few iteration steps are performed.

For determining the approximate inverse $N = W^{-1}$, we have to weight up the following properties.

- *Relatively rough approximation (moderate $\varepsilon < 1$)*: In this case, a smaller local rank of the \mathcal{H} -matrix representation is sufficient; hence, the storage cost and computational cost is reduced. As a consequence, we have to perform several steps of the iterative method (13.8). However, the latter fact is of lesser importance since the matrix-vector multiplications Ax^m and Nd for $d := Ax^m - b$ are significantly faster than inversion or LU decomposition required for N .
- *Relatively accurate approximation (small $\varepsilon \ll 1$)*: The local rank of the \mathcal{H} -matrix representation will increase logarithmically with $1/\varepsilon$. On the other hand, the iterative method requires only one or two steps.

The effective amount of work is

$$\text{Eff}(\Phi_{\mathcal{H}\text{-LU}}) = \mathcal{O}(r^\alpha \log^\beta n) \quad (n = \#I)$$

with $\alpha, \beta > 0$ (cf. (2.31a)). Therefore smaller local ranks r may be preferred. Usually, the maximal r is $\mathcal{O}(\log^* n)$, since then the discretisation error is reached. Hence the effective amount of work is always bounded by $\mathcal{O}(\log^* n)$; i.e., the \mathcal{H} -LU iteration is almost optimal.

Let A and W be a positive definite matrix. We recall the spectral equivalence of A and W defined by

$$\frac{1}{c} \langle Ax, x \rangle \leq \langle Wx, x \rangle \leq c \langle Ax, x \rangle \quad \text{for all } x \in \mathbb{K}^I \quad (13.10)$$

with a constant $c > 0$ (cf. Definition 7.56). According to (7.51e), (13.10) is equivalent to $\frac{1}{c}I \leq A^{-1/2}WA^{-1/2} \leq cI$. Inversion yields $\frac{1}{c}I \leq A^{1/2}NA^{1/2} \leq cI$. Applying (13.9c) yields the next statement.

Remark 13.5. Inequality (13.9c) implies the spectral equivalence (13.10) with

$$c := \frac{1}{1 - \varepsilon} \approx 1 + \varepsilon.$$

The spectral equivalence may come into play by other means. The solution of nonlinear problems or parabolic differential equations can lead to the situation³ that many systems $A^{(\nu)}x^{(\nu)} = b^{(\nu)}$ are to be solved, involving different matrices $A^{(\nu)}$ which are still spectrally equivalent. Then it is sufficient to approximate the inverse $N = (A^{(0)})^{-1}$ of the first matrix and to use this approximation as preconditioner for all $A^{(1)}, A^{(2)}, \dots$ (cf. page 321).

³ For instance, in the nonlinear case the Newton method leads to different linearisations $A^{(\nu)}$, where ν is the index of the Newton iteration. In the parabolic case, the matrices $A(t)$ depend on the time t . The time steps $t = 0, \Delta t, 2\Delta t, \dots$ yield $A^{(\nu)} = A(\nu\Delta t)$.

13.4.2 Algebraic LU Decomposition

The LU decomposition described in §13.1 is still dependent on geometric data (coordinates of the nodal points). The following construction removes this dependence and uses only data contained in the matrix A , provided that A is a sparse matrix.

Given A , we obtain the graph $G(A)$ (cf. §C.2). More precisely, we use the undirected graph $G = G_{\text{sym}}(A) = G(A) \cup G(A^T)$. We assume that G is connected, since otherwise the system decomposes into at least two separated systems. For a connected graph, any two $\alpha, \beta \in I$ are connected by at least one path. The path length is defined by the number of edges between α and β . The minimal length of all paths between α and β defines the distance $\delta(\alpha, \beta)$.

The distance δ yields the necessary topology. It allows defining the diameter of a cluster and the distance of two clusters. Therefore, admissibility condition (D.10) can be formulated.

The construction of the cluster tree $T(I)$ and, in particular, of the special cluster tree corresponding to §13.2.3 is explained in [198, §9.2] and Grasedyck–Kriemann–Le Borne [165]. The latter paper contains numerical examples which show that this approach is rather robust.

13.5 Further Applications of Hierarchical Matrices

The ‘commonly used matrix formulation’ $Ax = b$ in (1.5) is not the only representation. The linear equation may take the form of a *matrix equation*. For this purpose, let $A, C \in \mathbb{K}^{n \times n}$ be given matrices, while $X \in \mathbb{K}^{n \times n}$ is an unknown matrix. Then

$$AX + XA^H = C \quad (13.11)$$

represents the Lyapunov equation. This is a system of linear equations for all entries of X . In principle, we can form vectors $x, c \in \mathbb{K}^{n^2}$ and a matrix $\mathcal{A} \in \mathbb{K}^{n^2 \times n^2}$ such that (13.11) is equivalent to $\mathcal{A}x = c$. However, if A is a large-scale matrix, the size of x is equal to n^2 which may be too large for practical computations. The remedy is to use a format for the matrix X which involves only $\mathcal{O}(n)$ or $\mathcal{O}(n \log^* n)$ data. Hierarchical matrices are a possible choice. Possibly, even global low-rank matrices can be used. As shown by Penzl [310], $\text{rank}(C) = r$ implies that the singular values of X decrease exponentially. This property ensures approximability by global low-rank matrices.

A slight generalisation is the Sylvester equation $AX + XB = C$ with given matrices A, B , and C . Corresponding statements and approximations by global low-rank and hierarchical matrices are discussed by Grasedyck [159], Baur [35], Baur–Benner [36], and Benner–Breiten [40].

Even the (nonlinear) Riccati equation $AX + XA^H - XBX = C$ can be solved (cf. Hackbusch [198, §15.2] and Grasedyck–Hackbusch–Khoromskij [163]).

Chapter 14

Tensor-based Methods

Abstract There are situations in which the size of the system $Ax = b$ is so large that it is impossible to store the vectors x, b in full format. The data size may take values like 1000^{1000} . Instead one needs sparse representations for all quantities A, x, b . The numerical tensor calculus offers very efficient tools for this purpose. In Section 14.1 we introduce tensor spaces and show typical examples. The key for the efficient numerical treatment is a suitable sparse tensor representation. In Section 14.2 we briefly define the r -term format, the subspace format as well as the hierarchical format. The inverse matrix approximated in §14.2.2.3 will play an important role as preconditioner. In Section 14.3 two different types of huge linear systems are described together with the definition of the truncated iteration for their solution. Finally, in Section 14.4, the variational approach and the alternating least squares method are mentioned.

14.1 Tensors

14.1.1 Introductory Example: Lyapunov Equation

In the standard case, we assume that the size of the linear systems $Ax = b$ is such that the vectors $x, b \in \mathbb{K}^n$ can be treated directly, whereas the regular matrix $A \in \mathbb{K}^{n \times n}$ needs special care (e.g., sparse matrix format or hierarchical matrix format). The situation changes if n becomes much larger. An example is the Lyapunov matrix equation $AX + XA^H = C$ mentioned in (13.11) and related matrix equations (cf. Benner–Breiten [40]). We can rewrite the Lyapunov equation (13.11) as a traditional system

$$Ax = b \quad \text{with } x, b \in \mathbb{K}^{n^2}, A \in \mathbb{K}^{n^2 \times n^2}$$

of n^2 linear equations for n^2 unknowns. The matrix A can be defined by using the tensor product:

$$A = I \otimes A + A \otimes I.$$

Vectors $\mathbf{x}, \mathbf{b} \in \mathbb{K}^{n^2}$ are also viewed as elements of the tensor space $\mathbb{K}^n \otimes \mathbb{K}^n$.

If A is a sparse $n \times n$ matrix, the representation $\mathbf{A} = I \otimes A + A \otimes I$ shows that the $n^2 \times n^2$ matrix \mathbf{A} can be represented by $\mathcal{O}(n)$ data. However, the solution $\mathbf{x} \in \mathbb{K}^{n^2}$ is not sparse. Here, we need new concepts to avoid the data size $\mathcal{O}(n^2)$. This is in particular necessary when we replace n^2 by n^d with $d > 2$.

One origin of Lyapunov equations are control problems for partial differential equations. For instance, A may be the discretisation of the Laplace equation, say in three spatial dimensions $x = (x_1, x_2, x_3)$. Then the matrix \mathbf{A} of the Lyapunov equation is an discretisation of the six-dimensional partial differential equation $\Delta_x u(x, y) + \Delta_y u(x, y) = \dots$

14.1.2 Nature of the Underlying Problems

In §1.5 we started with

- (P1) direct methods for solving $Ax = b$.

Ignoring floating-point effects, we are able to obtain the exact solution, whenever the matrix A is regular. However, the exactness of the solution is illusory because of

- (P2) discretisation errors.

Usually, the origin of the large-scale systems $Ax = b$ is the discretisation of a partial differential equation, e.g., by finite elements. Hence, the true interest is in the solution of an infinite-dimensional linear problem $Lu = f$. The solution of $Ax = b$ may define a finite element function $u_n := \sum_{\alpha} x_{\alpha} \phi_{\alpha}$ (cf. (E.6)) approximating the true solution u . Computing x more accurately than $u - u^h$ is wasted effort. In this book, we do not discuss the discretisation error since this is the subject of other monographs (e.g., [193, 201]). Nevertheless, some statements about the Galerkin discretisation error are given in §E since the multigrid convergence proof in §11.6.3 is based on these properties.

The central theme of this book is the

- (P3) iterative solution of $Ax = b$.

Since we perform only a finite number of iteration steps, we have to accept an inexact solution x^m . According to (P2), this is a reasonable approach, provided that the iteration error is below the discretisation error. A minor modification is the

- (P4) iterative solution of $\tilde{A}x = \tilde{b}$ with a perturbed matrix $\tilde{A} \approx A$ and a perturbed right-hand side $\tilde{b} \approx b$.

For instance, the true finite element discretisation $Ax = b$ may require the evaluation of integrals of the type $\int_{\Omega} a(x)(\nabla \varphi_i)^T (\nabla \varphi_j) dx$ with a variable coefficient $a(\cdot)$. Since the exact evaluation may be cumbersome, such integrals are approximated by numerical quadrature leading to $\tilde{A} \approx A$. Similarly, \tilde{b} may involve quadrature errors. Although this approach is called a ‘variational crime’, the additional

quadrature error is integrated into the overall finite element error. Its influence is analysed by Strang's first lemma (cf. Strang [358]).

A similar situation happens in the case of hierarchical matrices. If A is a fully populated matrix, e.g., the discretisation of an integral operator, A is replaced with a hierarchical matrix \tilde{A} . In this case, the difference $A - \tilde{A}$ is called the truncation error. Under suitable conditions, error estimates of $A - \tilde{A}$ can be provided.

In the case of tensor-structured problems, the linear systems $Ax = b$ becomes

$$(P5) \quad \tilde{A}\tilde{x} \approx \tilde{b} \text{ with } \tilde{A} \approx A, \tilde{b} \approx b, \text{ and } \tilde{x} \approx x.$$

Because of the huge data size, the true matrix A and vectors x, b cannot be used. They have to be approximated by $\tilde{A}, \tilde{x}, \tilde{b}$, which belong to certain tensor formats. Note that, in general, the exact solution x of $\tilde{A}x = \tilde{b}$ also has a too large data size. Therefore it is necessary to look for \tilde{x} of a feasible data size and approximating x .

Whether A, b and the true solution $x = A^{-1}b$ can be well approximated by $\tilde{A}, \tilde{b}, \tilde{x}$ is a difficult question and will not be discussed here (see Hackbusch [195] for more details). The discussion of the discretisation error in (P2) is mainly based on the smoothness of the coefficients of the partial differential equation. In the tensor case, smoothness is helpful for estimating approximation errors, but also nonsmooth data may allow a good approximation.

In the following discussion, we treat matrices and vectors in the same way. Matrices are considered as elements (vectors) of the vector space $\mathbb{K}^{I \times I}$. The basic assumption is that the huge-sized vector spaces are organised as tensor spaces. As in the case of hierarchical matrices, we consider subsets of low-rank tensors using different versions of ranks. These low-rank tensors define the ‘tensor formats’ discussed in §§14.2.1, 14.2.3, 14.2.4.

14.1.3 Definition of Tensor Spaces

Given d vector spaces V_j , $1 \leq j \leq d$, one can define the algebraic tensor space

$$\mathbf{V} := \bigotimes_{j=1}^d V_j.$$

\mathbf{V} is a vector space of the dimension $\dim(\mathbf{V}) = \prod_{j=1}^d \dim(V_j)$. Starting from bases $\{b_i^{(j)} : i \in I_j\}$ of V_j with $\#I_j = \dim(V_j)$, the tensor products $\{\bigotimes_{j=1}^d b_{i_j}^{(j)} : i_j \in I_j, 1 \leq j \leq d\}$ form a basis in \mathbf{V} . This also holds in the infinite dimensional case. The tensor product is multilinear, i.e.,

$$\left(\bigotimes_{j=1}^{k-1} v^{(j)} \right) \otimes \left(\alpha v^{(k)} + w^{(k)} \right) \otimes \bigotimes_{j=k+1}^d v^{(j)}$$

$$= \alpha \bigotimes_{j=1}^d v^{(j)} + \left(\bigotimes_{j=1}^{k-1} v^{(j)} \right) \otimes w^{(k)} \otimes \bigotimes_{j=k+1}^d v^{(j)}$$

holds for all $v^{(j)} \in V_j$, $w^{(k)} \in V_k$ for $k \in \{1, \dots, d\}$, and $\alpha \in \mathbb{K}$. The products $\bigotimes_{j=1}^d v^{(j)}$ are called *elementary tensors*. The algebraic tensor space is spanned by all elementary tensors. For more details see Hackbusch [195, 196].

14.1.4 Case of Grid Functions

Choosing the vector spaces $V_j = \mathbb{K}^{I_j}$ with $I_j = \{1, \dots, n_j\}$, $n_j \in \mathbb{N}$, we obtain the tensor space $\mathbf{V} = \mathbb{K}^{\mathbf{I}}$ with the index set

$$\mathbf{I} = I_1 \times I_2 \times \dots \times I_d.$$

To avoid secondary indices, we write the components of $v^{(j)} \in \mathbb{K}^{I_j}$ as $v^{(j)}[i]$ with $i \in v^{(j)}$. The elementary tensor $\mathbf{v} = \bigotimes_{j=1}^d v^{(j)}$ is indexed by the tuple $\mathbf{i} = (i_1, i_2, \dots, i_d) \in \mathbf{I}$:

$$\mathbf{v}[\mathbf{i}] = \prod_{j=1}^d v^{(j)}[i_j]. \quad (14.1)$$

The tensor \mathbf{v} can be interpreted as a d -dimensional grid function. The underlying grid is

$$\Omega_h = \bigtimes_{j=1}^d \{h, \dots, n_j h\}$$

corresponding to the inner grid points $x_{\mathbf{i}} = \mathbf{i}h = (i_1 h, \dots, i_d h)$ of the cuboid $[0, a_1] \times \dots \times [0, a_d]$ with $a_j = (n_j + 1)h$ (the boundary grid points with $i_j = 0$ and $i_j = n_j + 1$ may correspond to given Dirichlet boundary values; see Figure 1.1 for $d = 2$). Denote the value of a grid function at the grid point $x_{\mathbf{i}}$ by $\mathbf{v}[\mathbf{i}]$. Then this grid function belongs to the tensor space $\mathbf{V} = \mathbb{K}^{\mathbf{I}}$.

The spatial dimension d may be large, e.g., $d = 1000$. However, even for $d = 3$ an $n \times n \times n$ grid with $n = 10^6$ points per direction causes severe problems. The general assumption of this chapter is that $\#\mathbf{I}$ is beyond the storage capacity of the computer.

Remark 14.1. Concerning discretisation of elliptic boundary value problems, the usual strategy is to reach a certain discretisation error with as few as possible degrees of freedom. The tools may be a local finite element refinement and the hp discretisation. The underlying hypothesis is that the work for solving the linear system is proportional to the size of the system. This idea is completely misleading in cases where the problem can be discretised in a tensor-structured way.

The discretisation by a regular three-dimensional grid with n grid points per direction yields a system of size $N = n^3$. However, if the methods discussed below work, the overall cost is $\mathcal{O}(n)$. Under optimal conditions even $\mathcal{O}(\log(n))$ may be possible (cf. [195, §14]).

14.1.5 Kronecker Products of Matrices

Now we consider the vector spaces $V_j = \mathbb{K}^{I_j \times J_j}$ of $I_j \times J_j$ matrices. $A_j \in V_j$ can be viewed as a linear map $A_j : X_j \rightarrow Y_j$ with $X_j = \mathbb{K}^{I_j}$ and $Y_j = \mathbb{K}^{J_j}$. We form the tensor spaces

$$\mathbf{X} = X_1 \otimes X_2 \otimes \dots \otimes X_d \quad \text{and} \quad \mathbf{Y} = Y_1 \otimes Y_2 \otimes \dots \otimes Y_d.$$

Then the tensor product

$$\mathbf{A} = A_1 \otimes A_2 \otimes \dots \otimes A_d$$

can be interpreted as the linear map $\mathbf{A} : \mathbf{X} \rightarrow \mathbf{Y}$ defined by

$$\mathbf{A} : \bigotimes_{j=1}^d v^{(j)} \mapsto \bigotimes_{j=1}^d (A_j v^{(j)}) \quad (v^{(j)} \in V_j). \quad (14.2)$$

Here we use that a linear map is completely determined by its action on elementary tensors. Since $\mathbf{X} = \mathbb{K}^{\mathbf{I}}$ and $\mathbf{Y} = \mathbb{K}^{\mathbf{J}}$ with $\mathbf{I} = \times_{j=1}^d I_j$ and $\mathbf{J} = \times_{j=1}^d J_j$, the linear map can be considered as a matrix $\mathbf{A} \in \mathbb{K}^{\mathbf{I} \times \mathbf{J}}$.

Tensor products of matrices are also called Kronecker products. Often definitions

as $A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \cdots \\ a_{21}B & a_{22}B & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$ are used which require a particular ordering of the indices. The definition (14.2) is independent of any index ordering.

Note that we treat matrices in $\mathbb{K}^{\mathbf{I} \times \mathbf{I}}$ and vectors in $\mathbb{K}^{\mathbf{I}}$ in the same way. Even the usual sparsity does not help: Also a diagonal matrix has the data size $\#\mathbf{I}$.

14.1.6 Functions on Cartesian Products

Let f_j be functions defined on Ω_j (we may think of $\Omega_j \subset \mathbb{R}$, but any set Ω_j is possible). Use the variables $x_j \in \Omega_j$. The tensor product $\mathbf{f} := \bigotimes_{j=1}^d f_j$ is a d -variate function defined on the Cartesian product $\Omega := \times_{j=1}^d \Omega_j$ by the product

$$\mathbf{f}(x_1, x_2, \dots, x_d) = f_1(x_1) \cdot f_2(x_2) \cdot \dots \cdot f_d(x_d).$$

Note that (14.1) is the particular case of $\Omega_j = \{ih : i \in I_j\}$.

Specifying function spaces V_j , e.g., $V_j = L^2(\Omega_j)$, the product \mathbf{f} belongs to the algebraic tensor space $\mathbf{V}_{\text{alg}} := \bigotimes_{j=1}^d V_j$. Using the norm of $L^2(\Omega)$ in \mathbf{V}_{alg} , we can complete the normed vector space \mathbf{V}_{alg} and obtain the *topological tensor space* \mathbf{V}_{top} , which in this case coincides with $L^2(\Omega)$. In the finite dimensional case, the algebraic and topological tensor spaces coincide.

14.2 Sparse Tensor Representation

Although, in general, $\dim(\mathbf{V})$ data are needed to describe a tensor in \mathbf{V} , there are subsets of \mathbf{V} which can be represented by a smaller number of parameters. We try to approximate the desired element of \mathbf{V} by such a tensor. In the following, we describe three formats. All are based on the idea of tensors of low rank (different from the hierarchical matrices and their local low-rank blocks, this kind of rank is of global nature).

14.2.1 r -Term Format (Canonical Format)

14.2.1.1 Definition and Tensor Rank

Analogously to the rank- r matrices in (D.2), we define the set \mathcal{R}_r of tensors of the *representation rank* $r \in \mathbb{N}_0$ by

$$\mathbf{v} = \sum_{\nu=1}^r v_\nu^{(1)} \otimes v_\nu^{(2)} \otimes \dots \otimes v_\nu^{(d)} \quad \text{with } v_\nu^{(j)} \in V_j. \quad (14.3)$$

The storage size for any $\mathbf{v} \in \mathcal{R}_r$ is bounded by rdn , where $n = \max_j \dim V_j$. Note that d is now a factor and not an exponent. If r is of moderate size, this format is advantageous. Often, a tensor \mathbf{v} requires a high rank, but there may be some approximation $\mathbf{v}_\varepsilon \in \mathcal{R}_r$ with moderate rank $r = r(\varepsilon)$. An example will follow in §14.3.1.

Since each element of an algebraic tensor space is a finite linear combination of elementary tensors, each $\mathbf{v} \in \mathbf{V}$ must belong to some \mathcal{R}_r , i.e., $\mathbf{V} = \bigcup_{r \in \mathbb{N}_0} \mathcal{R}_r$. We may define the tensor rank

$$\text{rank}(\mathbf{v}) := \min\{r \in \mathbb{N}_0 : \mathbf{v} \in \mathcal{R}_r\}.$$

In principle, we would like to use the representation rank $r = \text{rank}(\mathbf{v})$ in (14.3). However, since the determination of $\text{rank}(\mathbf{v})$ is in general NP hard (cf. Håstad [214]), the representation rank used in practice will be larger than $\text{rank}(\mathbf{v})$.

In the case of $n \times m$ matrices, we know that $\min\{n, m\}$ is the maximal possible matrix rank. In the case of tensor spaces, the exact maximal tensor rank is not well known. A rough upper bound is $\dim(\mathbf{V}) / \max_j \dim(V_j)$.

The statements above include the case that $v_\nu^{(j)} \in V_j$ are matrices and \mathbf{v} is the matrix $\mathbf{A} \in \mathbb{K}^{\mathbf{I} \times \mathbf{J}}$ from §14.1.5.

14.2.1.2 Operations

For later computational issues it is important to mention that the standard tensor operations can be performed using the representation (14.3):

1. The sum $\mathbf{v} + \mathbf{w}$ of tensors $\mathbf{v} \in \mathcal{R}_r$ and $\mathbf{w} \in \mathcal{R}_s$ yields a tensor in \mathcal{R}_{r+s} .
2. The matrix-vector multiplication follows the definition (14.2).
3. The Euclidean scalar product of elementary tensors $\mathbf{v} = \bigotimes_{j=1}^d v^{(j)}$ and $\mathbf{w} = \bigotimes_{j=1}^d w^{(j)}$ is

$$\langle \mathbf{v}, \mathbf{w} \rangle = \prod_{j=1}^d \langle v^{(j)}, w^{(j)} \rangle. \quad (14.4)$$

In the last two cases, only operations appear involving ‘small’ quantities from \mathbb{K}^{I_j} and $\mathbb{K}^{I_j \times I_j}$. The summation in item 1 is without any cost. Since the first two operations increase the representation rank, we need a truncation to a smaller rank like in §D.3.2. This, however, is the weak point of the r -term format. Unlike the situation in Proposition A.45, the attempt to determine a best approximation may lead to an instability since \mathcal{R}_r is not closed (cf. [195, §9.4]).

Therefore, in general, one prefers the other formats discussed below. If, however, one succeeds in finding a good approximation of the form (14.3) with moderate r , this is the method of choice. Such an example will be discussed next.

14.2.2 A Particular Example

14.2.2.1 Definition of the Matrix

We consider positive definite matrices $A_j \in \mathbb{K}^{I_j \times I_j}$ and the tensor space $\mathbb{K}^{\mathbf{I} \times \mathbf{I}} = \bigotimes_{j=1}^d \mathbb{K}^{I_j \times I_j}$. By I we denote the identity matrices in $\mathbb{K}^{I_j \times I_j}$ and define

$$\mathbf{A} := \sum_{j=1}^d \mathbf{A}_j \quad \text{with} \quad \mathbf{A}_j = I \otimes \dots \otimes I \otimes \underbrace{A_j}_{\text{j-th position}} \otimes I \otimes \dots \otimes I. \quad (14.5)$$

(14.5) is a d -term representation¹ for \mathbf{A} of the form (14.3).

Lemma 14.2. *The following rules are valid.*

- (a) $\left(\bigotimes_{j=1}^d A_j \right) \left(\bigotimes_{j=1}^d B_j \right) = \bigotimes_{j=1}^d (A_j B_j)$.
- (b) $\left(\bigotimes_{j=1}^d A_j \right)^\top = \bigotimes_{j=1}^d A_j^\top$.
- (c) \mathbf{A} in (14.5) has the spectrum $\sigma(\mathbf{A}) = \left\{ \sum_{j=1}^d \lambda_j : \lambda_j \in \sigma(A_j) \right\}$.
- (d) If all A_j are positive definite, then also \mathbf{A} is positive definite.
- (e) The condition of the matrix \mathbf{A} in (14.5) satisfies

$$\min_{1 \leq j \leq d} \operatorname{cond}_2(A_j) \leq \operatorname{cond}_2(\mathbf{A}) \leq \max_{1 \leq j \leq d} \operatorname{cond}_2(A_j).$$

¹ In fact, even the (minimal) tensor rank of \mathbf{A} is d , provided that A_j are not multiples of I (cf. Buczyński–Landsberg [83]).

Proof. (a) follows from (14.2). The proof of (b) uses the property (14.4) of the scalar product. For (c) use the eigenvector $\bigotimes_{j=1}^d v_{\nu_j}^{(j)}$ of \mathbf{A} composed of the eigenvectors $v_{\nu_j}^{(j)}$ of A_j .

(d) Combining (b) and (c), we obtain that $\mathbf{A} = \mathbf{A}^\top$ has only positive eigenvalues.

(e) Let λ_j be the minimal and Λ_j the maximal eigenvalue of A_j . As a consequence, $\text{cond}_2(A_j) = \Lambda_j/\lambda_j$ holds. The extreme eigenvalues of \mathbf{A} are $\sum_{j=1}^d \lambda_j$ and $\sum_{j=1}^d \Lambda_j$. The inequalities $\min_{1 \leq j \leq d} \frac{\Lambda_j}{\lambda_j} \leq \frac{\Lambda_1 + \dots + \Lambda_d}{\lambda_1 + \dots + \lambda_d} \leq \max_{1 \leq j \leq d} \frac{\Lambda_j}{\lambda_j}$ can be proved by induction. \square

We conclude from Lemma 14.2e that the large size of \mathbf{A} does not effect the size of the matrix condition. Furthermore, the minimal eigenvalue of \mathbf{A} in (14.5) increases with d . To simplify the following notation, we scale the matrix \mathbf{A} so that $\lambda_{\min}(\mathbf{A}) = 1$ or, at least,

$$\sigma(\mathbf{A}) \subset [1, \infty).$$

In the following we want to approximate the inverse matrix \mathbf{A}^{-1} . In Remark D.1 matrix functions are mentioned. The inverse \mathbf{A}^{-1} can be regarded as the matrix function $f(\mathbf{A})$ for the function $f(x) = 1/x$. Accordingly, approximations of \mathbf{A}^{-1} can be obtained as $g(\mathbf{A})$, where $g(x) \approx 1/x$.

14.2.2.2 Exponential Sums

Functions of the form $E_r(x) = \sum_{\nu=1}^r \alpha_\nu \exp(-\beta_\nu x)$ are called exponential sums. For the general theory of approximation by exponential sums we refer to Braess [60, Chap. VI]. The best approximation of $1/x$ in $[1, \infty)$ is defined by the minimiser of

$$\varepsilon_r = \min_{\alpha_\nu, \beta_\nu} \max_{1 \leq x < \infty} \left| \frac{1}{x} - \sum_{\nu=1}^r \alpha_\nu \exp(-\beta_\nu x) \right|. \quad (14.6)$$

Lemma 14.3. *The minimum in (14.6) is taken for positive parameters α_ν, β_ν ($1 \leq \nu \leq r$). The approximation error ε_r decays exponentially:*

$$\varepsilon_r \leq \mathcal{O}(\exp(-c\sqrt{r})) \quad \text{with } c > 0.$$

Explicit bounds are described in Braess–Hackbusch [66, 67]).

As an illustration we show some values of ε_r :

r	5	10	15	20	30	40	50
ε_r	$6.428_{10^{-4}}$	$1.312_{10^{-5}}$	$6.311_{10^{-7}}$	$4.794_{10^{-8}}$	$6.188_{10^{-10}}$	$1.554_{10^{-11}}$	$5.992_{10^{-13}}$

If we replace the interval $[1, \infty)$ in (14.6) by a bounded interval $[1, R]$, ε_r behaves as $\mathcal{O}(\exp(-cr))$.

14.2.2.3 Approximation of the Inverse

Lemma 14.4. *Let A be a symmetric matrix. If f and g are two functions defined on the spectrum $\sigma(A)$, then*

$$\|f(A) - g(A)\|_2 \leq \max_{\lambda \in \sigma(A)} |f(\lambda) - g(\lambda)|$$

holds with respect to the spectral norm.

Proof. $A = UD U^{-1}$ holds a unitary matrix U . Therefore

$$\begin{aligned} \|f(A) - g(A)\|_2 &= \|Uf(D)U^{-1} - Ug(D)U^{-1}\|_2 \\ &= \|f(D) - g(D)\|_2 = \max_{\lambda \in \sigma(A)} |f(\lambda) - g(\lambda)| \end{aligned}$$

proves the inequality. \square

We apply this lemma with $A = \mathbf{A}$, $f(x) = 1/x$ and $g(x) = E_r(x)$. Since the maximum over $\lambda \in \sigma(\mathbf{A})$ is bounded by the maximum over the larger set $[0, \infty)$, the error bound ε_r defined in (14.6) applies again:

$$\|\mathbf{A}^{-1} - \mathbf{B}_r\|_2 \leq \varepsilon_r \quad \text{for } \mathbf{B}_r := E_r(\mathbf{A}). \quad (14.7)$$

Lemma 14.5. *The matrix \mathbf{B}_r in (14.7) has the r -term representation*

$$\mathbf{B}_r = \sum_{\nu=1}^r \alpha_\nu \bigotimes_{j=1}^d \exp(-\beta_\nu A_j) \in \mathcal{R}_r. \quad (14.8)$$

Proof. The matrices $\mathbf{A}_1, \dots, \mathbf{A}_d$ defined in (14.5) are commutative; therefore $\exp(-\beta_\nu \mathbf{A}) = \prod_{j=1}^d \exp(-\beta_\nu \mathbf{A}_j)$ holds. Using Lemma 14.2a, we obtain

$$\begin{aligned} \exp(-\beta_\nu \mathbf{A}_j) &= \sum_{k=0}^{\infty} \frac{1}{k!} (-\beta_\nu \mathbf{A}_j)^k = \sum_{k=0}^{\infty} \frac{1}{k!} (I \otimes \dots \otimes (-\beta_\nu A_j) \otimes \dots \otimes I)^k \\ &= \sum_{k=0}^{\infty} \frac{1}{k!} (I^k \otimes \dots \otimes (-\beta_\nu A_j)^k \otimes \dots \otimes I^k) \\ &= I \otimes \dots \otimes \sum_{k=0}^{\infty} \frac{1}{k!} (-\beta_\nu A_j)^k \otimes \dots \otimes I \\ &= I \otimes \dots \otimes \exp(-\beta_\nu A_j) \otimes \dots \otimes I \end{aligned}$$

and $\prod_{j=1}^d \exp(-\beta_\nu \mathbf{A}_j) = \bigotimes_{j=1}^d \exp(-\beta_\nu A_j)$. Summation over ν proves the lemma. \square

If we use the hierarchical matrix format for approximating the matrix exponential $\exp(-\beta_\nu A_j)$ (cf. Remark D.1), the overall storage cost is $\mathcal{O}(rdN \log^* N)$, while the data size of \mathbf{A}^{-1} is N^d , where $N := \max_j \#I_j$.

We conclude with two generalisations.

Remark 14.6. (a) If the matrices A_j are not normal, but diagonalisable: $A_j = S_j D_j S_j^{-1}$, the estimate of Lemma 14.4 holds in the modified form

$$\|f(A) - g(A)\|_2 \leq \text{cond}_2(S_j) \max_{\lambda \in \sigma(A)} |f(\lambda) - g(\lambda)|.$$

(b) The estimate (14.6) measures the error on the real interval $[1, \infty)$. If the matrices A_j have a spectrum in the complex set $\Sigma = \{z = x + iy \in \mathbb{C} : x \geq 1, |y| \leq \delta\}$, there are exponential sums approximating $1/x$ in Σ with an exponential rate.

14.2.3 Subspace Format (Tucker Format)

Given $\mathbf{v} \in \mathbf{V} = \bigotimes_{j=1}^d V_j$, there may be subspaces $U_j \subset V_j$ of a smaller dimension r_j so that $\mathbf{v} \in \mathbf{U} = \bigotimes_{j=1}^d U_j$. Choosing a basis $\{b_i^{(j)} : 1 \leq i \leq r_j\}$ of U_j , the tensor has the representation

$$\mathbf{v} = \sum_{i_1=1}^{r_1} \sum_{i_2=1}^{r_2} \cdots \sum_{i_d=1}^{r_d} \mathbf{c}[i_1, i_2, \dots, i_d] \bigotimes_{j=1}^d b_{i_j}^{(j)}. \quad (14.9)$$

Instead of one rank parameter, we now have a tuple $\mathbf{r} := (r_1, \dots, r_d) \in \mathbb{N}_0^d$. The minimal value of r_j in all possible representations (14.9) of \mathbf{v} is called the j -th (Tucker) rank and denoted by $\text{rank}_j(\mathbf{v})$.

The storage cost of the basis vectors is bounded by rdn with

$$r := \max_j r_j \quad \text{and} \quad n := \max_j \dim(V_j).$$

The coefficients $\mathbf{c}[i_1, i_2, \dots, i_d]$ form the *core tensor* $\mathbf{c} \in \bigotimes_{j=1}^d \mathbb{K}^{r_j}$ and require a storage of $\prod_{j=1}^d r_j$. The latter number may become too large for increasing d .

Note that the subspace $\mathbf{U} = \bigotimes_{j=1}^d U_j$ is not fixed, but adapted to the tensor \mathbf{v} . If two tensor \mathbf{v} and \mathbf{w} are to be added, the associated subspaces $\mathbf{U}(\mathbf{v})$ and $\mathbf{U}(\mathbf{w})$ yield the subspace $\mathbf{U} = \bigotimes_{j=1}^d U_j$ containing $\mathbf{v} + \mathbf{w}$ with $U_j := U_j(\mathbf{v}) + U_j(\mathbf{w})$. Similarly, the matrix-vector multiplication and the scalar product can be performed within the subspace format. Different from the r -term format, the truncation to smaller ranks r_j can be computed in a stable way based on the so-called HOSVD (higher order singular value decomposition; cf. [195, §8.3 and §10.1] and De Lathauwer et al. [103]).

14.2.4 Hierarchical Tensor Format

Figure 14.1 shows a partition of the set

$$D = \{1, \dots, d\}$$

by the binary tree T_D . Its leaves are the singletons $\{1\}, \dots, \{d\}$. A non-leaf vertex $\alpha \in T_D$ has two sons α', α'' . All vertices are subsets of D with the property $\alpha = \alpha' \cup \alpha''$ (disjoint union). For all $\alpha \subset D$, the tensor spaces $V_\alpha = \bigotimes_{j \in \alpha} V_j$ can be defined. Similar to the previous format, we associate a subspace $U_\alpha \subset V_\alpha$ for each $\alpha \in T_D \setminus \mathcal{L}(T_D)$ with the characteristic property²

$$\mathbf{U}_\alpha \subset \mathbf{U}_{\alpha'} \otimes \mathbf{U}_{\alpha''} \quad (\alpha', \alpha'': \text{sons of } \alpha \in T_D). \quad (14.10)$$

As in §14.2.3, the subspace \mathbf{U}_α is described by a basis $\{\mathbf{b}_i^{(\alpha)} : 1 \leq i \leq r_\alpha\}$ with $r_\alpha = \dim(\mathbf{U}_\alpha)$. However, only for the leaves $\alpha = \{j\} \in \mathcal{L}(T_D)$ the basis vectors $\mathbf{b}_i^{(\alpha)} = b_i^{(j)}$ are stored explicitly. For $\alpha \in T_D \setminus \mathcal{L}(T_D)$, the basis vectors $\mathbf{b}_\ell^{(\alpha)}$ are already true tensors requiring a large storage. Instead we conclude from (14.10) the representation

$$\mathbf{b}_\ell^{(\alpha)} = \sum_{i=1}^{r_{\alpha'}} \sum_{j=1}^{r_{\alpha''}} c_{ij}^{(\alpha, \ell)} \mathbf{b}_i^{(\alpha')} \otimes \mathbf{b}_j^{(\alpha'')} \quad (\alpha', \alpha'': \text{sons of } \alpha \in T_D).$$

For an indirect definition of $\mathbf{b}_\ell^{(\alpha)}$ it is sufficient to store the coefficient matrix $C^{(\alpha, \ell)} = (c_{ij}^{(\alpha, \ell)}) \in \mathbb{K}^{r_{\alpha'} \times r_{\alpha''}}$.

The subspace \mathbf{U}_D at the root of T_D must contain the tensor which we want to represent: $\mathbf{v} \in \mathbf{U}_D$. Since \mathbf{U}_D can be chosen with $\dim(\mathbf{U}_D) = 1$, the final representation $\mathbf{v} = c_1 \mathbf{b}_1^{(D)}$ requires only one real number.

The rank tuple is now $\mathbf{r} = (r_\alpha)_{\alpha \in T_D}$. Using the bounds $r := \max_\alpha r_\alpha$, $n := \max_j \dim(\mathbf{U}_{\{j\}})$, the total storage cost is bounded by

$$dr^3 + dnr.$$

Different from the Tucker format in §14.2.3, the storage is linear in d . In spite of the indirect definition of the basis vectors $\mathbf{b}_\ell^{(\alpha)}$, all tensor operations can be performed. Also the HOSVD truncation of the ranks applies (cf. [195, §11.3]).

The so-called TT or matrix product format corresponds to the hierarchical format with a linear tree T_D (cf. [195, §12] and Oseledets–Tyrtyshnikov [300]).

² This property has a close relation to the \mathcal{H}^2 -matrices mentioned in §D.2.7.

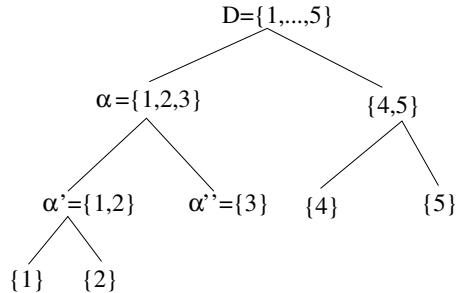


Fig. 14.1 Dimension partition tree.

14.3 Linear Systems

14.3.1 Poisson Model Problem

We generalise the model problem in §1.2 to d spatial dimensions,

$$-\Delta u = -\sum_{j=1}^d \frac{\partial^2 u}{\partial x_j^2} \quad \text{in } \Omega = (0, 1)^d \subset \mathbb{R}^d,$$

with Dirichlet condition $u = 0$ on $\partial\Omega$, and discretise by a finite difference scheme with the grid size $h = 1/N$ in a regular product grid. Let A_j be the (positive definite) tridiagonal matrix built by the (one-dimensional) negative second divided differences. Then the system matrix \mathbf{A} is of the form (14.5). Therefore the constructions of §14.2.2 can be used to build the rather accurate inverse of \mathbf{A} .

A similar result can be obtained for finite elements using a regular grid, but then the identity I is replaced with the mass matrix, i.e.,

$$\mathbf{A}_j = M_1 \otimes \dots \otimes M_{j-1} \otimes A_j \otimes M_{j+1} \otimes \dots \otimes M_d.$$

Define $\mathbf{M} := \bigotimes_{j=1}^d M_j$. Then $\mathbf{A} = \mathbf{M}\hat{\mathbf{A}}$ holds with

$$\hat{\mathbf{A}} = \sum_{j=1}^d \hat{\mathbf{A}}_j, \quad \hat{\mathbf{A}}_j = \mathbf{M}^{-1} \mathbf{A}_j = I \otimes \dots \otimes M_j^{-1} A_j \otimes \dots \otimes I, \quad \hat{A}_j = M_j^{-1} A_j.$$

The inverse M_j^{-1} and the product $M_j^{-1} A_j$ is easy to approximate by hierarchical matrices. Applying the technique of §14.2.2 together with the generalisation of Remark 14.6a, $\hat{\mathbf{A}}^{-1}$ can be approximated by $\hat{\mathbf{B}}_r$. Then $\hat{\mathbf{B}}_r \mathbf{M}^{-1}$ approximates the inverse of \mathbf{A} . Note that $\mathbf{M}^{-1} = \bigotimes_{j=1}^d M_j^{-1}$ by Lemma 14.2a.

Numerical examples involving matrices $\mathbf{A} \in \mathbb{R}^{n \times n}$ with $n = 1024^{256}$ ($N = 1024$ and $d = 256$) are described by Grasedyck [158].

14.3.2 A Parametrised Problem

Obviously the order of tensors is connected with the number of variables. However, in the case of boundary value problems, not all variables need to be involved in derivatives. A standard boundary value problem is $Lu = f$ in D with vanishing Dirichlet values and

$$L = \operatorname{div} a \operatorname{grad}. \tag{14.11}$$

D may be domain in \mathbb{R}^1 , \mathbb{R}^2 , or \mathbb{R}^3 . In the standard case, the coefficient function a depends on $x \in D$. Assume now that there are more variables $y_j \in Y_j$ ($1 \leq j \leq d$) which serve as parameters. In particular, we assume that the coefficient a in (14.11)

depends on $x \in \Omega$ and $y = (y_1, \dots, y_d) \in Y := \times_{j=1}^d Y_j$:

$$a = a(x, y) \quad (x \in D, y \in Y). \quad (14.12)$$

Such parametrised systems are a simplified version of boundary value problems with *stochastic* coefficients

$$a = a(x, \omega) \quad (x \in D, \omega \in \Omega),$$

where ω is a random variable. The infinite singular value decomposition

$$a(x, \omega) = a_0(x) + \sum_{j=1}^{\infty} \sigma_j \phi_j(x) X_j(\omega) \quad (14.13)$$

is called the Karhunen–Loëve expansion (cf. Karhunen [234], Loëve [267, §37.5B]). The expansion (14.13) shows that ω corresponds to infinitely many variables. Note that the solution of $Lu = f$ with a as in (14.12) or (14.13) is $u(x, y)$ or $u(x, \omega)$ depending on y or ω .

The standard approach is a truncation of the infinite sum (14.13) (cf. Matthies–Zander [277]). Concerning the decay of the singular values σ_j we refer, e.g., to Todor–Schwab [335]. A truncation to d terms leads to a deterministic problem with a parametrised coefficient (14.12). For another approach see Espig et al. [124].

Let $Z := \times_{j=1}^d Z_j$ be a (finite) grid contained in Y , i.e., the variable $y_j \in Y_j$ is restricted to finitely many grid points in Z_j . In principle, one has to solve the discretisation of $\operatorname{div} a(\cdot, y) \operatorname{grad}$ for all parameter combinations $y \in Z$. If $\#Z_j = n$, there are n^d systems $A(y)x = b$ to be solved.

Assume that a has an r -term representation

$$a = \sum_{\nu=1}^r a_{\nu} \otimes \bigotimes_{j=1}^d \psi_{\nu}^{(j)}, \quad (14.14)$$

i.e., $a(x, y) = \sum_{\nu=1}^r a_{\nu}(x) \psi_{\nu}^{(1)}(y_1) \psi_{\nu}^{(2)}(y_2) \cdots \psi_{\nu}^{(d)}(y_d)$. Let A_{ν} be the discretisation of $\operatorname{div} a_{\nu} \operatorname{grad}$. Then the parametrised problem takes the form

$$\mathbf{A}\mathbf{u} = \mathbf{f}, \quad \text{where } \mathbf{A} = \sum_{\nu=1}^r A_{\nu} \otimes \bigotimes_{j=1}^d \Psi_{\nu}^{(j)},$$

where $\Psi_{\nu}^{(j)}$ denotes the Hadamard product $\Psi_{\nu}^{(j)}(\varphi) = \psi_{\nu}^{(j)} \circ \varphi$ (pointwise product). Assuming that $0 < \underline{\alpha} \leq a(x, y) \leq \overline{\alpha}$ ensures ellipticity, we may choose the elementary tensor

$$\mathbf{B} := \Delta_h^{-1} \otimes \bigotimes_{j=1}^d \operatorname{id} \quad (14.15)$$

as a preconditioner, where Δ_h^{-1} is an approximate inverse of the discrete Laplace operator.

14.3.3 Solution of Linear Systems

Consider a linear system $\mathbf{Ax} = \mathbf{b}$, where the vectors $\mathbf{x}, \mathbf{b} \in \mathbf{V} = \bigotimes_{j=1}^d V_j$ and the matrix $\mathbf{A} \in \bigotimes_{j=1}^d \mathcal{L}(V_j, V_j) \subset \mathcal{L}(\mathbf{V}, \mathbf{V})$ are represented in one of the representation formats. The general linear iteration (2.10) is

$$\mathbf{x}^{m+1} = \mathbf{x}^m - \mathbf{N}(\mathbf{Ax}^m - \mathbf{b})$$

with some matrix \mathbf{N} given in a tensor representation. If this iteration is applied, e.g., with the starting value $\mathbf{x}^0 = 0$, the representation ranks of \mathbf{x}^m would blow up. Therefore truncation T must be applied. This yields the *truncated iteration*

$$\mathbf{x}^{m+1} = T(\mathbf{x}^m - \mathbf{N}(T(\mathbf{Ax}^m - \mathbf{b}))). \quad (14.16)$$

The cost per step is Nd times powers of the involved representation ranks.

A suitable choice of \mathbf{N} in (14.16) is \mathbf{B}_r (cf. (14.8)) in the case of the d -dimensional Poisson problem in §14.3.1, and \mathbf{B} (cf. (14.15)) in the case of the parametrised problem of §14.3.2.

\mathbf{B}_r in (14.8) is not only a preconditioner for the Laplace problem, but for any discretisation \mathbf{A} of an elliptic partial differential equation of second order as stated in Lemma 7.63 (cf. Khoromskij [238]).

14.3.4 CG-Type Methods

An acceleration of the iteration by the CG method is possible. As mentioned on page 238, floating-point errors cause a loss of the A -orthogonality of the search directions. This effect is even worse in the presence of tensor rank truncations. Articles about conjugate gradient methods in the context of tensor methods are by Tobler [362], Kressner–Tobler [242, 243, 244], Savas–Eldén [332], and Ballani–Grasedyck [24].

14.3.5 Multigrid Approach

Concerning adapting the multigrid iteration to the tensor case, we refer to Hackbusch [199]. The critical points are the coarse-grid, the cost, and the smoothing iteration.

The coarse-grid corresponds to an index set $\mathbf{I}^{(0)} = I_1^{(0)} \times \dots \times I_d^{(0)}$. If $\#I_j^{(0)}$ is of moderate size and d is large, the coarse-grid dimension $\#\mathbf{I}^{(0)} = \prod_{j=1}^d \#I_j^{(0)}$ is still rather large, rising the question how to solve the coarse-grid equation. This difficulty does not appear if $\#I_j^{(0)} = 1$ for all j .

Another aspect is the increase of the cost with increasing level-number ℓ . Assuming the standard halving of the grid size, the problem size in §11 is $\mathcal{O}(2^{\ell d})$. The effect for $d = 3$ is that the cost of the operations (prolongation, restriction, smoothing) at level $\ell - 1$ is only $1/8$ of the cost at level ℓ . Now the cost of the tensor operations is no more exponential in d , but proportional to 2^ℓ . In the case of the W-cycle, the result of Exercise 11.18 applies.

The practical choice of the smoothing iteration is limited. The Richardson iteration is a possible smoothing procedure. It requires only matrix-vector multiplication by the system matrix (followed by truncation). The Jacobi iteration is already difficult. If the diagonal parts of the matrices A_j in (14.5) are multiples of the identity matrix, the resulting Jacobi iteration coincides with the (particularly damped) Richardson iteration (cf. Remark 3.8). For more general diagonal parts, the inversion of \mathbf{D} causes severe difficulties (cf. [199, §4.4.2]).

The prolongations and restrictions are elementary tensor products (i.e., $r = 1$ in (14.3)). Therefore their applications do not increase the representation ranks. Constructing the coarse-grid matrices by the Galerkin product (11.20) does also not increase the representation rank of the matrix at the finest level.

14.3.6 Convergence

We obtain the standard convergence results, provided that no tensor truncation is applied. The additional effect of truncation is similar to the truncated iteration analysed by Hackbusch–Khoromskij–Tyrtyshnikov [203] (see also [198, §15.3.2]). The reached accuracy of the iterate x depends essentially on the choice of the ranks of x .

Numerical examples for the use of a tensor multigrid iteration can be seen in Ballani–Grasedyck [24, Example 7.5]. These examples with $N_{\text{fine}} := N^{(L)} = 1024$ and dimensions up to $d = 32$ demonstrate that the convergence behaviour does not depend on d .

A multigrid solution of the Sylvester matrix equation $AX - XB = C$ is described in Grasedyck–Hackbusch [162]. A nonlinear multigrid approach to the quadratic Riccati equation is presented by Grasedyck [160].

14.3.7 Parabolic Problems

Since the dimensions in the sense of number of coordinates are no limitation, space-time simultaneous discretisations with the additional time variable are not disadvantageous. In this respect, the results of Andreev–Tobler [5] are of interest. In that paper, an additive multigrid preconditioner is used.

14.4 Variational Approach

Finally, we mention a quite different approach for solving $\mathbf{Ax} = \mathbf{b}$ approximately. If \mathbf{A} is positive definite, we may minimise the quadratic cost function

$$\Phi(\mathbf{x}) := \frac{1}{2} \langle \mathbf{Ax}, \mathbf{x} \rangle - \langle \mathbf{b}, \mathbf{x} \rangle$$

(cf. (9.2)). For a more general, regular \mathbf{A} , define

$$\varPhi(\mathbf{x}) := \|\mathbf{Ax} - \mathbf{b}\|^2 \quad \text{or} \quad \varPhi(\mathbf{x}) := \|\mathbf{N}(\mathbf{Ax} - \mathbf{b})\|^2$$

with a suitable preconditioner \mathbf{N} and try to minimise $\varPhi(\mathbf{x})$. The minimisation over all tensors \mathbf{x} is not feasible because of the huge dimension. Instead one fixes a certain format for the representation of \mathbf{x} and minimises over all representation parameters of \mathbf{x} .

A popular technique is the alternating least squares (ALS) minimisation. Consider for instance the r -term format of §14.2.1 and use $\mathbf{x} = \sum_{\nu=1}^r \bigotimes_{j=1}^d v_\nu^{(j)}$ as an ansatz with fixed r . Then the functional \varPhi becomes $\phi(v^{(1)}, v^{(2)}, \dots, v^{(d)})$, where

$$v^{(j)} = (v_\nu^{(j)})_{\nu=1,\dots,r}.$$

In the first ALS step, we fix $v^{(j)}$ for all $j \neq 1$. Then, due to the multilinearity of tensors, $\phi(v^{(1)}, v^{(2)}, \dots, v^{(d)}) = \varphi_1(v^{(1)})$ is quadratic in $v^{(1)}$. The minimisation of $\varphi(v^{(1)})$ leads to a linear system $\partial \varphi_1(v^{(1)}) / \partial v^{(1)} = 0$ for $v^{(1)}$. Replacing the previous $v^{(1)}$ by the computed optimiser $v^{(1)}$, we fix all $v^{(j)}$ for $j \neq 2$ in $\phi(v^{(1)}, v^{(2)}, \dots, v^{(d)}) = \varphi_2(v^{(2)})$ and determine a new $v^{(2)}$, etc.

In practice, this method works quite well although the theoretical understanding of the convergence properties is rather involved (see, e.g., Espig–Hackbusch–Khachatryan [123]). Another difficulty is that \varPhi has many local minima and that the global minimisation problem is nonconvex.

Further references to variational approaches are Espig–Hackbusch–Rohwedder–Schneider [125], Falcó–Nouy [126], Holtz–Rohwedder–Schneider [224], Mohlenkamp [284], Oseledets [299] and others cited in these papers.

Appendix A

Facts from Linear Algebra

Abstract We introduce the notation of vector and matrices (cf. Section A.1), and recall the solvability of linear systems (cf. Section A.2). Section A.3 introduces the spectrum $\sigma(A)$, matrix polynomials $P(A)$ and their spectra, the spectral radius $\rho(A)$, and its properties. Block structures are introduced in Section A.4. Subjects of Section A.5 are orthogonal and orthonormal vectors, orthogonalisation, the QR method, and orthogonal projections. Section A.6 is devoted to the Schur normal form (§A.6.1) and the Jordan normal form (§A.6.2). Diagonalisability is discussed in §A.6.3. Finally, in §A.6.4, the singular value decomposition is explained.

A.1 Notation for Vectors and Matrices

We recall that the field \mathbb{K} denotes either \mathbb{R} or \mathbb{C} . Given a finite index set I , the linear space of all vectors $x = (x_i)_{i \in I}$ with $x_i \in \mathbb{K}$ is denoted by \mathbb{K}^I . The corresponding square matrices form the space $\mathbb{K}^{I \times I}$. $\mathbb{K}^{I \times J}$ with another index set J describes rectangular matrices mapping \mathbb{K}^J into \mathbb{K}^I .

The linear subspace of a vector space V spanned by the vectors $\{x^\alpha \in V : \alpha \in I\}$ is denoted and defined by

$$\text{span}\{x^\alpha : \alpha \in I\} := \left\{ \sum_{\alpha \in I} a_\alpha x^\alpha : a_\alpha \in \mathbb{K} \right\}.$$

Let $A = (a_{\alpha\beta})_{\alpha, \beta \in I} \in \mathbb{K}^{I \times I}$. Then $A^T = (a_{\beta\alpha})_{\alpha, \beta \in I}$ denotes the transposed matrix, while $A^H = (\overline{a_{\beta\alpha}})_{\alpha, \beta \in I}$ is the adjoint (or Hermitian transposed) matrix. Note that $A^T = A^H$ holds if $\mathbb{K} = \mathbb{R}$. Since (x_1, x_2, \dots) indicates a row vector, $(x_1, x_2, \dots)^T$ is used for a column vector.

Exercise A.1. Prove the following rules for T and H (where $\lambda \in \mathbb{K}$):

$$\begin{aligned} (A + B)^T &= A^T + B^T, & (AB)^T &= B^T A^T, & (\lambda A)^T &= \lambda A^T, \\ (A + B)^H &= A^H + B^H, & (AB)^H &= B^H A^H, & (\lambda A)^H &= \bar{\lambda} A^H, \\ (A^{-1})^T &= (A^T)^{-1}, & (A^{-1})^H &= (A^H)^{-1} = A^{-H}. \end{aligned}$$

The inverse of a transposed or adjoint matrix is shortly denoted by

$$A^{-\top} := (A^\top)^{-1}, \quad A^{-\text{H}} := (A^{\text{H}})^{-1}.$$

Definition A.2. A matrix $A \in \mathbb{K}^{I \times I}$ is called

- symmetric if $A = A^\top$,
- Hermitian if $A = A^{\text{H}}$,
- regular if A^{-1} exists,
- unitary if $A^{\text{H}}A = I$ (i.e., A regular and $A^{-1} = A^{\text{H}}$),
- normal if $AA^{\text{H}} = A^{\text{H}}A$.

Remark A.3. (a) Hermitian or unitary matrices are also normal.

- (b) All matrix properties of Definition A.2 carry over from A to the adjoint A^{H} .
(c) Products of regular (unitary) matrices are again regular (unitary).

A *diagonal matrix* D is completely described by its diagonal entries. We write

$$D = \text{diag}\{d_\alpha : a \in I\} \quad \text{for } D \text{ with } D_{\alpha\beta} = \begin{cases} d_\alpha & \text{for } \alpha = \beta, \\ 0 & \text{for } \alpha \neq \beta. \end{cases} \quad (\text{A.1})$$

If I is ordered, we may also write $D = \text{diag}\{d_1, d_2, \dots, d_n\}$. For an arbitrary matrix $A \in \mathbb{K}^{I \times I}$,

$$D = \text{diag}\{A\}$$

denotes the *diagonal part* $\text{diag}\{a_{\alpha\alpha} : \alpha \in I\}$ of A .

In the case of an ordered index set, a matrix T is called *tridiagonal* if $T_{ij} = 0$ for all $|i - j| > 1$; i.e., if T has the band width 1 (cf. Definition 1.6). The entries $\alpha_i = T_{i,i-1}$ define the lower side diagonal, $\beta_i = T_{ii}$ the (main) diagonal, and $\gamma_i = T_{i,i+1}$ the upper side diagonal, while all other entries of T vanish. Such a matrix is abbreviated as

$$T = \text{tridiag}\{(\alpha_i, \beta_i, \gamma_i) : i \in I\} \quad (\text{A.2})$$

(here the values α_1 and $\gamma_{\#I}$ are meaningless). By $\text{tridiag}\{A\}$ we denote the *tridiagonal part* of an arbitrary matrix A .

Assuming again an ordered index set, a matrix T is called a *lower triangular* matrix if $T_{ij} = 0$ for all $i < j$. Similarly, T is called *upper triangular* if $T_{ij} = 0$ for all $i > j$. T is a *strictly lower* or *upper triangular matrix* if, in addition, $T_{ii} = 0$ for all $i \in I$.

A.2 Systems of Linear Equations

Let $A \in \mathbb{K}^{I \times I}$ and $b \in \mathbb{K}^I$. The system of equations to be solved is

$$Ax = b, \quad \text{i.e.,} \quad \sum_{\beta \in I} a_{\alpha\beta} x_\beta = b_\alpha \quad \text{for all } \alpha \in I.$$

Since the right-hand side b may be perturbed (by rounding errors, etc.), the relevant question is: when is $Ax = b$ solvable for all $b \in \mathbb{K}^I$? The following theorem recalls that this property is equivalent to the regularity of A .

Theorem A.4. For $A \in \mathbb{K}^{I \times I}$, the following properties are equivalent:

- (a) A is regular;
- (b) $\text{rank}(A) = \#I$,
- (c) $\det(A) = 0$,
- (d) $Ax = 0$ has only the trivial solution $x = 0$,
- (e) $Ax = b$ is solvable for all $b \in \mathbb{K}^I$,
- (f) $Ax = b$ has at most one solution,
- (g) $Ax = b$ is uniquely solvable for all $b \in \mathbb{K}^I$.

A.3 Eigenvalues and Eigenvectors

The *spectrum* of a matrix $A \in \mathbb{K}^{I \times I}$ is defined by

$$\sigma(A) := \{\lambda \in \mathbb{C} : \det(A - \lambda I) = 0\}.$$

Each $\lambda \in \sigma(A)$ is called an *eigenvalue* of A . An eigenvalue has the *algebraic multiplicity* k if it is a k -fold root of the characteristic polynomial $\det(A - \lambda I)$. Since $\det(A - \lambda I)$ is a polynomial in λ of degree $n = \#I$, there exist exactly n eigenvalues when they are counted according to their algebraic multiplicity. The *geometric multiplicity* of λ is the dimension of $\ker(A - \lambda I)$.

The properties of the determinant prove the next properties.

Remark A.5. $\sigma(A^T) = \sigma(A)$ and $\sigma(A^H) = \sigma(\bar{A}) = \overline{\sigma(A)} := \{\bar{\lambda} : \lambda \in \sigma(A)\}$.

A vector $e \in \mathbb{C}^I$ is called an *eigenvector* of the matrix A , if $e \neq 0$ and

$$Ae = \lambda e. \tag{A.3}$$

By Theorem A.4c,d, we conclude from (A.3) that λ must be an eigenvalue. Vice versa, the same theorem proves the following lemma.

Lemma A.6. For each $\lambda \in \sigma(A)$, there exists an eigenvector e satisfying the eigenvalue problem (A.3). Hence the geometric multiplicity is at least one.

Exercise A.7. Let $A = (a_{ij})_{i,j \in I}$ be an upper or lower triangular matrix or a diagonal matrix. Prove that $\sigma(A) = \{a_{ii} : i \in I\}$.

Definition A.8. Two matrices $A, B \in \mathbb{K}^{I \times I}$ are called *similar* if there is a regular matrix T such that

$$A = T^{-1}BT. \tag{A.4}$$

If T is unitary, the matrices A and B are called *unitarily similar*.

Theorem A.9. (a) The eigenvalues of similar matrices A and B coincide: $\sigma(A) = \sigma(B)$. The algebraic multiplicities of the eigenvalues are also equal as well as the geometric multiplicities.

(b) If T is the similarity transformation in (A.4) and e is an eigenvector of A , then Te is an eigenvector of B .

Proof. The algebraic multiplicities are equal since

$$\begin{aligned}\det(A - \lambda I) &= \det(T^{-1}(B - \lambda I)T) = \det(T^{-1}) \det(B - \lambda I) \det(T) \\ &= \frac{1}{\det(T)} \det(B - \lambda I) \det(T) = \det(B - \lambda I).\end{aligned}$$

$\ker(A - \lambda I) = \ker(T^{-1}(B - \lambda I)T) = \ker(B - \lambda I)T$ proves identical dimensions of $\ker(A - \lambda I)$ and $\ker(B - \lambda I)$ and therefore of the geometric multiplicities.

Part (b) uses $B(Be) = TT^{-1}BT e = TAe = T(\lambda e) = \lambda(Be)$. \square

Theorem A.10. The products AB and BA have the same spectra with a possible exception of a zero eigenvalue:

$$\sigma(AB) \setminus \{0\} = \sigma(BA) \setminus \{0\}.$$

This statement is also true for rectangular matrices $A \in \mathbb{K}^{I \times J}$ and $B \in \mathbb{K}^{J \times I}$.

Proof. Let the eigenvector $e \neq 0$ belong to the eigenvalue $0 \neq \lambda \in \sigma(AB)$: $ABe = \lambda e$. Since $\lambda e \neq 0$, the vector $v := Be$ does not vanish. Multiplying by B yields $BABe = \lambda Be$, i.e., $BAv = \lambda v$ with $v \neq 0$. $\lambda \in \sigma(BA) \setminus \{0\}$ proves $\sigma(AB) \setminus \{0\} \subset \sigma(BA) \setminus \{0\}$. The reverse inclusion is analogous. \square

Given a polynomial $P(\xi) = \sum_{\nu} a_{\nu} \xi^{\nu}$ in $\xi \in \mathbb{C}$, we can extend the domain of definition of P by

$$P(A) := \sum_{\nu} a_{\nu} A^{\nu} \quad \text{for arbitrary } A \in \mathbb{K}^{I \times I}$$

to the set of square matrices. Here, A^0 is defined as the identity I . The proof of the following lemma is postponed to the end of §A.6.1.

Lemma A.11. (a) The spectra of A and $P(A)$ satisfy

$$\sigma(P(A)) = P(\sigma(A)) := \{P(\lambda) : \lambda \in \sigma(A)\}.$$

(b) The algebraic multiplicity of the eigenvalues $P(\lambda)$ of $P(A)$ is the sum of the multiplicities of all eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_k$ of A with $P(\lambda_j) = P(\lambda)$ ($1 \leq j \leq k$).

(c) Each eigenvector of A associated with the eigenvalue λ is also an eigenvector of $P(A)$ with the eigenvalue $P(\lambda)$.

Exercise A.12. Prove the following:

- (a) If $\sigma(A)$ contains no zeros of the polynomial $P(\xi)$, then the matrix $P(A)$ is regular.
- (b) The properties ‘diagonal’, ‘upper triangular matrix’, ‘lower triangular matrix’ carry over from A to $P(A)$. This statement is also true for the properties ‘symmetric’ and ‘Hermitian’, provided that P has real coefficients.
- (c) Let A be regular. All properties mentioned in (b) carry over from A to A^{-1} .

Lemma A.13. Let $A \in \mathbb{K}^{I \times I}$ be a strictly (upper or lower) triangular matrix. Then $A^m = 0$ holds for all $m > \#I$.

Proof. One proves by induction that A^m ($m \in \mathbb{N}$) has a vanishing main diagonal and $m - 1$ vanishing side diagonals: $(A^m)_{ij} = 0$ for $|i - j| < m$. For $m > \#I$, the inequality $|i - j| < m$ holds for all indices; hence $A^m = 0$. \square

Two matrices A and B are called *commutative* (or ‘ A and B commute’) if

$$AB = BA.$$

Exercise A.14. Prove: (a) If A and B are commutative, $P(A)$ and $Q(B)$ also commute for arbitrary polynomials P and Q . In particular, $P(A)$ and $Q(A)$ as well as $P(A)$ and A are commutative pairs.

- (b) For regular A , the matrices $P(A)$ and $P(A^{-1})$ commute.
- (c) Under the assumption of (a), $P(A)$, $P(A)^{-1}$, $Q(B)$, and $Q(B)^{-1}$ are pairwise commutative as long as the inverse matrices exist.

Two polynomials P and Q define the rational function $R(\xi) := P(\xi)/Q(\xi)$. By Exercise A.12a, the matrix

$$R(A) := P(A)(Q(A))^{-1} \tag{A.5}$$

is defined if and only if the spectrum $\sigma(A)$ contains no zero of the polynomial Q in the denominator. The foregoing results prove the next remark.

Remark A.15. Assume that $\sigma(A)$ contains no pole of the rational function R . Then the following statements hold:

- (a) (A.5) is equivalent to $R(A) = (Q(A))^{-1}P(A)$.
- (b) Lemma A.11 also holds for R instead of P . In particular, $\sigma(R(A)) = R(\sigma(A))$ is the spectrum of $R(A)$.
- (c) The properties ‘diagonal’, ‘upper triangular matrix’, ‘lower triangular matrix’ carry over from A to $R(A)$. The same statement holds for the terms ‘symmetric’ and ‘Hermitian’ if R has real coefficients.

Exercise A.16. Assume that $\sigma(A)$ contains no pole of the rational function R and prove:

- (a) The similarity $A = T^{-1}BT$ (cf. (A.4)) implies $R(A) = T^{-1}R(B)T$.
- (b) If $D = \text{diag}\{d_\alpha : \alpha \in I\}$, then $R(D) = \text{diag}\{R(d_\alpha) : \alpha \in I\}$.

A fundamental term for iterative methods is the following.

Definition A.17. The *spectral radius* $\rho(A)$ of a matrix A is the largest absolute value of the eigenvalues of A :

$$\rho(A) := \max\{|\lambda| : \lambda \in \sigma(A)\}.$$

Lemma A.18. *The spectral radius satisfies the following rules:*

$$\rho(\zeta A) = |\zeta| \rho(A) \quad \text{for all } \zeta \in \mathbb{K} \text{ and } A \in \mathbb{K}^{I \times I}, \quad (\text{A.6a})$$

$$\rho(A^k) = (\rho(A))^k \quad \text{for all } k \in \mathbb{N}_0 \text{ and } A \in \mathbb{K}^{I \times I}, \quad (\text{A.6b})$$

$$\rho(A) = \rho(B) \quad \text{for similar matrices } A, B \in \mathbb{K}^{I \times I}, \quad (\text{A.6c})$$

$$\rho(A) = \rho(A^H) = \rho(A^\top) \quad \text{for all } A \in \mathbb{K}^{I \times I}. \quad (\text{A.6d})$$

Proof. (i) Let the maximum of $\{|\lambda| : \lambda \in \sigma(A)\}$ be attained at $\lambda' \in \sigma(A)$. Then $|\zeta\lambda|$ and $|\lambda^k|$ ($\lambda \in \sigma(A)$) also take their maxima at $\lambda = \lambda'$, which proves (A.6a,b).

(ii) For similar matrices A and B , we have $\sigma(A) = \sigma(B)$ (cf. Theorem A.9a). This implies (A.6c).

(iii) (A.6d) is a consequence of Remark A.5. □

Exercise A.19. Prove the following: (a) A diagonal or triangular matrix has the spectral radius

$$\rho(A) = \max\{|a_{aa}| : a \in I\}.$$

(b) $\rho(A) = 0$ holds for strictly triangular matrices A .

(c) The spectral radius is not submultiplicative. Give an example of two 2×2 matrices A, B such that

$$\rho(AB) > \rho(A) \rho(B).$$

Lemma A.20. $\rho(AB) = \rho(BA)$ holds for all $A \in \mathbb{K}^{I \times J}$ and $B \in \mathbb{K}^{J \times I}$.

Proof. According to Theorem A.10, the spectra of AB and BA differ at most by the eigenvalue 0, which is irrelevant for the definition of the spectral radius. □

In the case of a multiple product, cyclic permutations satisfy

$$\rho(A_0 A_1 \dots A_m) = \rho(A_1 \dots A_m A_0).$$

A.4 Block Vectors and Block Matrices

As demonstrated by the matrices of the model problems in (1.8) and (1.9), vectors and matrices often have a special block structure. The exact definition of a block structure is based on a decomposition of the index set I into disjoint and nonempty subsets:

$$I = \bigcup_{\kappa \in B} I_\kappa, \quad I_\kappa, I_\lambda \ (\kappa, \lambda \in B) \text{ be pairwise disjoint.} \quad (\text{A.7})$$

B is the index set of the blocks. The vector $x \in \mathbb{K}^I$ decomposes into the vector blocks x^κ ($\kappa \in B$):

$$x^\kappa := x|_{I_\kappa} := (x_\alpha)_{\alpha \in I_\kappa}, \quad x = (x^\kappa)_{\kappa \in B}. \quad (\text{A.8a})$$

Example A.21. In the case of the model problem in §1.2, the grid Ω_h in (1.3) is the obvious index set. The grid consists of $N - 1$ ‘rows’

$$I_j = \{(ih, jh) : 1 \leq i \leq N - 1\}, \quad j = 1, \dots, N - 1.$$

In this case, $B = \{j : 1 \leq j \leq N - 1\}$ is the block index set.

Let $A \in \mathbb{K}^{I \times I}$. For each pair $\kappa, \lambda \in B$, the decomposition (A.7) of I defines a matrix block (a submatrix)

$$A^{\kappa\lambda} := A|_{I_\kappa \times I_\lambda} := (a_{\alpha\beta})_{\alpha \in I_\kappa, \beta \in I_\lambda} \quad \text{for all } \kappa, \lambda \in B. \quad (\text{A.8b})$$

In general, the blocks $A^{\kappa\lambda}$ are rectangular submatrices. The complete matrix can be built from these blocks:

$$A = (A^{\kappa\lambda})_{\kappa, \lambda \in B}. \quad (\text{A.8c})$$

Each submatrix $(a_{\alpha\beta})_{\alpha, \beta \in K}$ associated with a subset $K \subset I$ is called a *principal submatrix*. The diagonal blocks¹ $A^{\kappa\kappa}$ in (A.8b) are special principal submatrices.

The term ‘block’ is ambiguous. It is used for the index subset I_κ , for a vector block x^κ as well as for the submatrix $A^{\kappa\lambda}$.

Among all block structures there are two extreme cases: if B has only one element, A consists of one block coinciding with A ; if $B = I$, i.e., if all subsets $I_\kappa = \{\kappa\}$ have only one element, the terms ‘block’ and ‘matrix entry’ coincide.

If the block indices $B = \{1, \dots, k\}$ are ordered, a block matrix can be represented in the form

$$A = \begin{bmatrix} A^{11} & A^{12} & \cdots & A^{1k} \\ A^{21} & A^{22} & \cdots & A^{2k} \\ \vdots & \vdots & \ddots & \vdots \\ A^{k1} & A^{k2} & \cdots & A^{kk} \end{bmatrix}.$$

Note that, in general, only the diagonal blocks A^{ii} must be square submatrices.

¹ Note that a *diagonal block* $A^{\kappa\kappa}$ is not a diagonal matrix, but a block in the diagonal position of the block matrix (A.8c). If we want to express that $A^{\kappa\kappa}$ is diagonal, this is emphasised explicitly, e.g., by ‘a block $A^{\kappa\kappa}$ of diagonal structure’.

Example A.22. In the case of the model problem, the rows can be taken as blocks according to Example A.21. Then $A^{jj} = h^{-2}T$ (cf. (1.8)) are the diagonal blocks and $A^{j,j-1} = A^{j,j+1} = -h^{-2}I$ are the off-diagonal blocks. All further blocks are zero and therefore not represented in (1.8). Note that a visual representation by (1.8) is only possible if the indices are ordered.

Remark A.23. Block matrices can be interpreted twofold. First, they can be regarded as matrices that are structured by the index decomposition (A.7). Second, they can be represented as matrices of the index set B (not I) with matrix-valued (not \mathbb{K} -valued) entries. For example, the matrix multiplication $A \cdot B$ can be defined directly by the blocks: $(AB)^{\kappa\lambda} = \sum_{\gamma \in B} A^{\kappa\gamma} B^{\gamma\lambda}$

The second interpretation in Remark A.23 allows us to generalise the terms ‘diagonal, tridiagonal, and triangular matrix’ immediately to block matrices: A is called a *block-diagonal* matrix with respect to an index decomposition (A.7) if $A^{\kappa\lambda} = 0$ (zero block) for all $\kappa \neq \lambda$, $\kappa, \lambda \in B$.

Analogously to (A.1), we write

$$A = \text{blockdiag}\{D^\kappa : \kappa \in B\}$$

for a block-diagonal matrix with $A^{\kappa\kappa} = D^\kappa$. For an arbitrary matrix $C \in \mathbb{K}^{I \times I}$,

$$A = \text{blockdiag}\{C\} := \text{blockdiag}\{C^{\kappa\kappa} : \kappa \in B\}$$

denotes the *block-diagonal part* of C which we obtain after setting all off-diagonal blocks to zero. Since different block structures B may lead to different block-diagonal parts, the precise notation is $\text{blockdiag}_B\{C\}$.

Similarly, we write

$$A = \text{blocktridiag}\{(E^j, D^j, F^j) : j \in B\} \quad (\text{A.9})$$

for a *block-tridiagonal* matrix (cf. (A.2)) if B is ordered. A is an upper (lower) *block-triangular* matrix if $A^{ij} = 0$ for all $i, j \in B$ with $i > j$ ($i < j$).

Exercise A.24. Prove: (a) $(A^\top)^{\kappa\lambda} = (A^{\lambda\kappa})^\top$, $(A^H)^{\kappa\lambda} = (A^{\lambda\kappa})^H$.

(b) The diagonal blocks of Hermitian matrices are again Hermitian.

(c) Let A be a block diagonal or block-tridiagonal matrix with diagonal blocks $A^{\kappa\kappa}$ ($\kappa \in B$). The characteristic polynomial of A is the product of the characteristic polynomials of $A^{\kappa\kappa}$ ($\kappa \in B$). The spectrum and the spectral radius of A satisfy

$$\begin{aligned} \sigma(A) &= \bigcup \{\sigma(A^{\kappa\kappa}) : \kappa \in B\}, \\ \rho(A) &= \max\{|\lambda| : \lambda \text{ eigenvalue of } A^{\kappa\kappa}, \kappa \in B\} = \max_{\kappa \in B} \rho(A^{\kappa\kappa}). \end{aligned} \quad (\text{A.10})$$

(d) The diagonal blocks of block-triangular or block-diagonal matrices satisfy

$$(P(A))^{\kappa\kappa} = P(A^{\kappa\kappa}) \quad (\kappa \in B, P \text{ polynomial}).$$

(e) The block-diagonal structure is invariant with respect to the application of polynomials P :

$$P(\text{blockdiag}\{D^\kappa : \kappa \in B\}) = \text{blockdiag}\{P(D^\kappa) : \kappa \in B\}.$$

A.5 Orthogonality

A.5.1 Elementary Definitions

A *scalar product* of a vector space V is a positive, symmetric sesquilinear form² $\langle \cdot, \cdot \rangle : V \times V \rightarrow \mathbb{K}$, i.e., it satisfies

$$\langle x, x \rangle > 0 \quad \text{for all } 0 \neq x \in V, \quad (\text{A.11a})$$

$$\langle x + \lambda x', y \rangle = \langle x, y \rangle + \lambda \langle x', y \rangle \quad \text{for } x, x', y \in V, \lambda \in \mathbb{K}, \quad (\text{A.11b})$$

$$\langle x, y \rangle = \overline{\langle y, x \rangle} \quad \text{for } x, y \in V. \quad (\text{A.11c})$$

(A.11b, c) imply *semilinearity* with respect to the second argument:

$$\langle x, y + \lambda y' \rangle = \langle x, y \rangle + \bar{\lambda} \langle x, y' \rangle \quad \text{for } x, y, y' \in V, \lambda \in \mathbb{C}$$

In the real case $\mathbb{K} = \mathbb{R}$, one may write $\langle y, x \rangle$ and λ instead of $\overline{\langle y, x \rangle}$ and $\bar{\lambda}$.

The *Euclidean scalar product* on $V = \mathbb{K}^I$ is defined by

$$\langle x, y \rangle := \sum_{\alpha \in I} x_\alpha \overline{y_\alpha} \quad \text{for } x, y \in \mathbb{K}^I.$$

If not differently defined, $\langle \cdot, \cdot \rangle$ will always denote the Euclidean scalar product. Another representation of the Euclidean scalar product is $\langle x, y \rangle = y^H x$.

Two vectors $x, y \in V$ with $\langle x, y \rangle = 0$ are called *orthogonal* (with respect to $\langle \cdot, \cdot \rangle$) and symbolised by $x \perp y$. The vectors $x, y \in V$ are *orthonormal* if they are orthogonal and normalised, i.e., $\langle x, x \rangle = \langle y, y \rangle = 1$. A basis $\{b^\alpha : \alpha \in I\}$ is called an *orthonormal basis* if the vectors b^α are mutually orthonormal.

If W is a subspace of the vector space V , then $x \in V$ is called orthogonal to W (symbolised by $x \perp W$) if $x \perp w$ for all $w \in W$. W^\perp denotes the orthogonal complement of W :

$$W^\perp := \{x \in V : x \perp W\}.$$

Exercise A.25. Let $\dim(V) < \infty$. Prove that $(W^\perp)^\perp = W$.

Remark A.26 (orthogonalisation method). If b^1, b^2, \dots, b^m are m linearly independent vectors of V , then the procedure

$$w^i := b^i - \sum_{j=1}^{i-1} \langle v^j, b^i \rangle v^j \quad \text{and} \quad v^i := \frac{1}{\|w^i\|_2} w^i \quad (i = 1, \dots, m) \quad (\text{A.12})$$

produces m pairwise orthonormal vectors v^i spanning the same subspace:

$$\text{span}\{b^1, \dots, b^m\} = \text{span}\{v^1, \dots, v^m\}.$$

² ‘Sesquilinear’ reduces to ‘bilinear’ in the real case $\mathbb{K} = \mathbb{R}$.

A.5.2 Orthogonal and Unitary Matrices

Definition A.27. A matrix $A \in \mathbb{K}^{I \times J}$ is called orthogonal if $A^H A = I \in \mathbb{K}^{J \times J}$, i.e., if the columns of A are mutually orthonormal.

Note that a *square* matrix with $A^H A = I$ is unitary and also satisfies $AA^H = I$ (cf. Definition A.2).

Remark A.28 (QR). For $V = \mathbb{K}^I$, let $A := [b^1 \ b^2 \ \dots \ b^m] \in \mathbb{K}^{I \times m}$ be the matrix formed by the columns b^j , while $Q := [v^1 \ v^2 \ \dots \ v^m]$ with v^i defined in (A.12) is an orthogonal matrix. Then there is an upper triangular $m \times m$ matrix R with $A = QR$. This factorisation is called the *QR decomposition* of A (cf. Quarteroni–Sacco–Saleri [314, §3.4.3]).

Finally, we define projections and, in particular, orthogonal projections.

Definition A.29. (a) A linear map (matrix) $P : \mathbb{K}^I \rightarrow \mathbb{K}^I$ is called a *projection* onto $U \subset \mathbb{K}^I$ if $P^2 = P$ holds and $U := \{Px : x \in \mathbb{K}^I\}$ is the range of P .

(b) A projection P is called *orthogonal* if P is also Hermitian: $P = P^H$.

Exercise A.30. Let P be an orthogonal projection onto U . The vectors $u := Px$ and $u^\perp := x - u$ describe the unique decomposition of $x = u + u^\perp$ into $u \in U$ and $u^\perp \in U^\perp$.

A.5.3 Sums of Subspaces and Orthogonal Complements

Sums of subspaces are defined by $U + V = \{u + v : u \in U, v \in V\}$. The sum is called a *direct sum*, denoted by $U \oplus V$, if $U \cap V = \{0\}$. The simplest example of a direct sum is $U \oplus U^\perp$.

Exercise A.31. Prove that the following statements are equivalent: (i) $U \oplus V$ is a direct sum, (ii) any $x \in U + V$ has a unique decomposition $x = x' + x''$ with $x' \in U, x'' \in V$, (iii) $\dim(U + V) = \dim(U) + \dim(V)$.

Proposition A.32. Subspaces $U, V \subset X$ satisfy $(U \cap V)^\perp = U^\perp + V^\perp$.

Proof. Assume that $x \in U^\perp$. Then $x \perp U$ also implies $x \perp (U \cap V)$. This shows that $U^\perp \subset (U \cap V)^\perp$. Together with the analogous inclusion $V^\perp \subset (U \cap V)^\perp$, we obtain $U^\perp + V^\perp \subset (U \cap V)^\perp$ and the direct sum

$$Y := (U^\perp + V^\perp) \oplus (U \cap V).$$

The proposition is proved if $Y = X$ or, equivalently, $Y^\perp = \{0\}$. Let $x \in Y^\perp$. By definition of Y , $x \perp Y$ implies $x \perp U^\perp$, $x \perp V^\perp$, and $x \perp (U \cap V)$. The first two statements yield $x \in U$ and $x \in V$ (cf. Exercise A.25), so that $x \in U \cap V$. Together with $x \perp (U \cap V)$, we obtain $x = 0$; i.e., $Y^\perp = \{0\}$ and $Y = X$. \square

Conclusion A.33. Let $U_i \subset X$ ($1 \leq i \leq k$) be subspaces. Then

$$\left(\bigcap_{i=1}^k U_i \right)^\perp = \sum_{i=1}^k U_i^\perp.$$

Proof. The start of the induction proof is given by Proposition A.32 for $k = 2$. Let the statement hold for $k - 1$ and apply Proposition A.32 with $U := U_1$ and $V := \bigcap_{i=2}^k U_i$. Then $\left(\bigcap_{i=1}^k U_i \right)^\perp = (U \cap V)^\perp = U^\perp + V^\perp = U_1^\perp + \left(\bigcap_{i=2}^k U_i \right)^\perp = U_1^\perp + \sum_{i=2}^k U_i^\perp$ proves the statement. \square

A.6 Normal Forms

A.6.1 Schur Normal Form

The following theorem states that all matrices are unitarily similar to an upper triangular matrix (cf. Definition A.8). Evidently, the upper triangular matrix could also be replaced with a lower one. To be able to define a triangular matrix, the index set I must be ordered (different orderings yield different Schur normal forms).

Theorem A.34 (Schur normal form). For any matrix $A \in \mathbb{K}^{I \times I}$, there is a unitary matrix Q and an upper triangular matrix U such that

$$A = Q U Q^H. \quad (\text{A.13})$$

Q describes a unitary similarity transformation of A into upper triangular form (the normal form):

$$U = Q^H A Q. \quad (\text{A.14})$$

The term ‘normal form’ does not imply that Q and U are uniquely determined.

Proof. We proof Theorem A.34 by induction on $n := \#I$. For $n = 1$, Eq. (A.13) holds for $Q := I$ and $U := A$. Let the assertion be true for $n - 1$. Choose an eigenvalue $\lambda \in \sigma(A)$ and a corresponding eigenvector e (possible because of Lemma A.6). The normalised vector $x^1 := e/\sqrt{\langle e, e \rangle}$ can be extended to an orthonormal basis by suitable x^2, \dots, x^n (use Remark A.26). Let $X := [x^1, x^2, \dots, x^n]$ denote the matrix with the column vectors x^i . According to Definition A.2, X is a unitary matrix. Let \mathbf{e}^1 be the first unit vector: $\mathbf{e}_i^1 = \delta_{1,i}$. The first column of $A' := X^H A X$ is $A' \mathbf{e}^1 = X^H A X \mathbf{e}^1 = X^H A x^1 = \lambda X^H x^1 = \lambda X^H X \mathbf{e}^1 = \lambda \mathbf{e}^1$ because x^1 as well as e are eigenvectors corresponding to λ . The decomposition of the index set $I = \{1, \dots, n\}$ into $I_1 := \{1\}$ and $I_2 := \{2, \dots, n\}$ induces a block decomposition of A' into $A' = [\lambda e^1, \dots] = \begin{bmatrix} \lambda & a \\ 0 & A'' \end{bmatrix}$ with an $I_2 \times I_2$ matrix A'' and an I_2 row vector a . Since $\#I_2 = n - 1$, the induction hypothesis implies that there

is a unitary $I_2 \times I_2$ matrix Y such that $Y^H A'' Y = U'$ is an upper triangular $I_2 \times I_2$ matrix. Definition A.2 states that the $I \times I$ matrix $Y' := \begin{bmatrix} 1 & 0 \\ 0 & Y \end{bmatrix}$ augmented by one row and one column is again unitary. The product $U := Y'^H A' Y' = Y'^H X^H A X Y'$ results in

$$Y'^H \begin{bmatrix} \lambda & aY \\ 0 & A''Y \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & Y^H \end{bmatrix} \begin{bmatrix} \lambda & aY \\ 0 & A''Y \end{bmatrix} = \begin{bmatrix} \lambda & aY \\ 0 & Y^H A''Y \end{bmatrix} = \begin{bmatrix} \lambda & aY \\ 0 & U' \end{bmatrix}.$$

If U' is an upper triangular matrix, U is also. The product $Q := X Y'$ is unitary (cf. Remark A.3c). Hence, (A.14) and (A.13) are proved. \square

Exercise A.7 and Theorem A.9 yield the following result.

Corollary A.35. The diagonal of U in (A.13) contains the eigenvalues of A :

$$\sigma(A) = \{u_{ii} : i \in I\}.$$

Proof of Lemma A.11. Let P be a polynomial and represent A according to (A.13) by $Q U Q^H$. Because of $P(A) = Q P(U) Q^H$ (cf. Exercise A.16a), the characteristic polynomials of $P(A)$ and $P(U)$ coincide (cf. Theorem A.9a). $P(U)$ is again an upper triangular matrix (cf. Remark A.15c) with the diagonal entries $(P(U))_{ii} = P(U_{ii})$ (cf. Exercise A.16c). By Theorem A.9a, statements (a) and (b) of the lemma follow. Part (c) is evident. \square

A.6.2 Jordan Normal Form

Analysing the kernels of $(A - \lambda I)^k$ for $\lambda \in \sigma(A)$ and $k = n, n-1, \dots, 1$, one can construct a basis formed by principal vectors and eigenvectors generating a transformation T into the Jordan normal form (cf. Gantmacher [144, 145, VII.§7]). The Jordan normal form is an upper triangular matrix with a more restrictive structure than U in (A.13). The disadvantage is that, in general, T is not unitary.

The bidiagonal $k \times k$ matrix (*Jordan block*)

$$J(\lambda, k) := \left[\begin{array}{cccccc} \lambda & 1 & & & & & \\ & \lambda & 1 & & & & \\ & & \ddots & \ddots & & & \\ & & & \lambda & 1 & & \\ & & & & \ddots & & \\ & & & & & \ddots & \\ & & & & & & \lambda \end{array} \right] \quad \left. \right\} \quad k \text{ rows and columns}$$

has the eigenvalue λ with the algebraic multiplicity k . The geometrical multiplicity is equal to one since only one eigenvector exists.

Theorem A.36 (Jordan normal form). *For any matrix $A \in \mathbb{K}^{I \times I}$, there exists a regular matrix T transforming A into its Jordan normal form J :*

$$A = T J T^{-1} \quad \text{or equivalently} \quad J = T^{-1} A T. \quad (\text{A.15a})$$

Here, J is an upper triangular matrix with the block-diagonal structure

$$J = \text{blockdiag}_{i=1,\dots,K}\{J(\lambda_i, k_i)\} \quad \text{with } k_i \geq 1, \quad \sum_{i=1}^K k_i = n := \#I. \quad (\text{A.15b})$$

The numbers λ_i run over all eigenvalues in $\sigma(A)$. The k_i corresponding to equal eigenvalues λ_i sum up to the algebraic multiplicity of λ_i . K coincides with the maximum number of linearly independent eigenvectors.

Since A and J are similar, they have the same characteristic polynomial (cf. Theorem A.9a). By Exercise A.24c, their characteristic polynomial is

$$\chi(\xi) = \prod_{i=1}^K \det(J(\lambda_i, k_i) - \xi I) = \prod_{i=1}^K (\lambda_i - \xi)^{k_i}. \quad (\text{A.16a})$$

Since some of the λ_i in (A.16a) may coincide, k_i is not necessarily the multiplicity of λ_i . We define

$$\begin{aligned} \bar{k}(\lambda) &:= \text{algebraic multiplicity of } \lambda \in \sigma(A), \\ \underline{k}(\lambda) &:= \max\{k_i : \lambda_i = \lambda, 1 \leq i \leq K\} \quad \text{for } \lambda \in \sigma(A). \end{aligned} \quad (\text{A.16b})$$

Obviously, $\bar{k}(\lambda) \geq \underline{k}(\lambda)$ and $\chi(\xi) = \prod_{\lambda \in \sigma(A)} (\lambda - \xi)^{\bar{k}(\lambda)}$ hold, where the product has to be taken over all different eigenvalues in $\sigma(A)$. Hence, the polynomial

$$\mu(\xi) := \prod_{\lambda \in \sigma(A)} (\lambda - \xi)^{\underline{k}(\lambda)} \quad (\text{A.16c})$$

is a divisor of the characteristic polynomial $\chi(\xi)$. The polynomial $\mu(\xi)$ is called the *minimum function* of A , because it is the polynomial of smallest degree satisfying the following requirement (A.17).

Theorem A.37 (Cayley-Hamilton). *Let μ and χ be the minimum function and the characteristic polynomial of a matrix A , respectively. Then*

$$\mu(A) = \chi(A) = 0 \quad (0 : \text{zero matrix}). \quad (\text{A.17})$$

Proof. (i) To prove $p(B) = 0$ for a polynomial p , it suffices to show $q(B) = 0$ for a divisor polynomial q .

(ii) Define $q(x) := (\lambda - \xi)^{\underline{k}(\lambda)}$ with $\lambda = \lambda_i$ for some $i \in \{1, \dots, K\}$. Since $\lambda_i I - J(\lambda_i, k_i)$ is a strictly upper triangular matrix and $\bar{k}(\lambda) \geq \underline{k}(\lambda)$ according to definition (A.16b), $q(J(\lambda_i, k_i))$ is the zero matrix (cf. Lemma A.13). $q(\xi)$ is a divisor of $\mu(\xi)$; hence, $\mu(J(\lambda_i, k_i)) = 0$ follows from part (i) for all $i = 1, \dots, K$.

(iii) Exercise A.24e applied to the block-diagonal matrix J yields

$$\mu(J) = \text{blockdiag}_{i=1,\dots,K}\{\mu(J(\lambda_i, k_i))\} = \text{blockdiag}_{i=1,\dots,K}\{0\} = 0.$$

By Exercise A.16a, we conclude from (A.15a) that $\mu(A) = T\mu(J)T^{-1} = 0$. As μ is a divisor of χ , the remaining part of (A.17) follows from (i). \square

A.6.3 Diagonalisability

If $k_i = 1$ for all $i = 1, \dots, K$, the matrix J in (A.15b) becomes a diagonal matrix. In this case, (A.15a) describes a transformation into diagonal form.

Theorem A.38 (diagonalisability). Let $A \in \mathbb{K}^{I \times I}$. A regular matrix T transforming A into diagonal form,

$$A = TD T^{-1}, \quad D = \text{diag}\{\lambda_\alpha : \alpha \in I\}, \quad (\text{A.18})$$

exists if and only if there are $n := \#I$ linearly independent eigenvectors. In this case, A is termed ‘diagonalisable’. If, furthermore, all eigenvalues λ_α ($\alpha \in I$) are real, A is real diagonalisable.

Proof. Assuming (A.18), we conclude from $AT = TD$ that the α -th column vectors $e^\alpha := T e^\alpha$ (e^α : α -unit vector) of T are the (linearly independent) eigenvectors of A . Vice versa, given n linearly independent eigenvectors, from these column vectors we can build the matrix T satisfying $AT = TD$ and proving the statement (A.18). \square

Remark A.39. Let A and B be similar matrices. Then A is diagonalisable if and only if B is also.

In general, the transformation matrix T in (A.18) is not unitary. More precisely, the following theorem holds (we recall that unitary and normal matrices are defined in Definition A.2).

Theorem A.40. A unitary matrix Q transforming A into diagonal form,

$$A = QDQ^H, \quad Q \text{ unitary}, \quad D = \text{diag}\{\lambda_\alpha : \alpha \in I\}, \quad (\text{A.19})$$

exists if and only if A is normal.

Proof. (i) In the case of $A = QBQ^H$ with a unitary matrix Q , A is normal if and only if B is normal, since

$$A^H A = (QB^H Q^H)(QBQ^H) = QB^H B Q^H$$

and

$$A A^H = (QBQ^H)(QB^H Q^H) = QBB^H Q^H.$$

(ii) In the case of (A.19), we can apply part (i) with $B = D$. A diagonal matrix is always normal, hence A is also.

(iii) Let A be normal and assume that QUQ^H is its Schur normal form. Following part (i), U is normal. By induction on $n := \#I$ we are proving that a normal upper triangular matrix is diagonal. For $n = 1$, both terms are identical.

The $n \times n$ matrix U can be written in the block structure $U = \begin{bmatrix} \lambda & a^H \\ 0 & U' \end{bmatrix}$ with

an upper triangular $(n - 1) \times (n - 1)$ matrix U' and an $(n - 1)$ row vector a^H . Comparing

$$UU^H = \begin{bmatrix} \lambda & a^H \\ 0 & U' \end{bmatrix} \begin{bmatrix} \bar{\lambda} & 0 \\ a & U'^H \end{bmatrix} = \begin{bmatrix} |\lambda|^2 + a^H a & \dots \\ \dots & U' U'^H \end{bmatrix}$$

with

$$U^H U = \begin{bmatrix} \bar{\lambda} & 0 \\ a & U'^H \end{bmatrix} \begin{bmatrix} \lambda & a^H \\ 0 & U' \end{bmatrix} = \begin{bmatrix} |\lambda|^2 & \dots \\ \dots & U'^H U' \end{bmatrix},$$

we see that $a^H a = \langle a, a \rangle = 0$, hence $a = 0$. Furthermore, since U' is normal, the induction hypothesis implies that U' is diagonal. Therefore, U is also diagonal, i.e., $D := U$ satisfies (A.19). \square

Since Hermitian matrices A are special normal matrices (cf. Remark A.3a), they share the representation (A.19). $A = A^H$ is equivalent to $D = D^H$. On the other hand, $D = D^H$ characterises real diagonal matrices. This proves the next theorem.

Theorem A.41. *If and only if A is Hermitian, there is a unitary matrix Q transforming A into a real diagonal matrix D ,*

$$A = QDQ^H, \quad Q \text{ unitary}, \quad D = \text{diag}\{\lambda_\alpha : \alpha \in I\} \text{ real}. \quad (\text{A.20})$$

Not only polynomials but also general functions can be applied to diagonalisable matrices, as explained below.

Remark A.42. Let A be diagonalisable. For an arbitrary function $f : \sigma(A) \rightarrow \mathbb{C}$, the matrix $f(A)$ is defined by

$$f(A) := T \text{ diag}\{f(\lambda_\alpha) : \alpha \in I\} T^{-1}$$

with T and $D = \text{diag}\{\lambda_\alpha : \alpha \in I\}$ in (A.18). A and $f(A)$ commute. For a second function $g : \sigma(A) \rightarrow \mathbb{C}$, $f(A)$ and $g(A)$ also commute. Furthermore, we have for all regular matrices $S \in \mathbb{K}^{I \times I}$ that

$$f(SAS^{-1}) = S f(A) S^{-1}.$$

Theorem A.43. *Let A and B be normal. A and B commute if and only if there exists a simultaneous unitary transformation into diagonal form:*

$$Q^H A Q = \text{diag}\{\lambda_\alpha : \alpha \in I\}, \quad Q^H B Q = \text{diag}\{\mu_\alpha : \alpha \in I\}. \quad (\text{A.21})$$

The column vectors of Q are the common eigenvectors of A and B .

Proof. (i) Since diagonal matrices always commute, (A.21) implies that

$$Q^H A B Q = (Q^H A Q)(Q^H B Q) = (Q^H B Q)(Q^H A Q) = Q^H B A Q$$

and therefore $AB = BA$.

(ii) Let T be unitary with

$$T^H AT = D_A := \text{diag}\{\lambda_\alpha : \alpha \in I\}.$$

$AB = BA$ implies that $D_AX = XD_A$ with $X := T^HBT$. First, we suppose that $\lambda_\alpha \neq \lambda_\beta$ for $\alpha \neq \beta$. From

$$\lambda_\alpha X_{\alpha\beta} = (D_AX)_{\alpha\beta} = (XD_A)_{\alpha\beta} = \lambda_\beta X_{\alpha\beta},$$

it follows that $X_{\alpha\beta} = 0$ for $\alpha \neq \beta$. Hence, X is diagonal, i.e., $Q := T$ also transforms B into a diagonal matrix $X = T^HBT$. If, otherwise, there are multiple eigenvalues, X is a block-diagonal matrix and we can choose

$$S = \text{blockdiag}\{S^\kappa : \kappa \in B\}$$

such that S^κ is unitary and transforms the diagonal block $X^{\kappa\kappa}$ into diagonal form. $Q := TS$ has the desired properties. \square

A.6.4 Singular Value Decomposition

For any matrix $M \in \mathbb{K}^{I \times J}$ with $\text{rank}(M) = r$, there exist orthonormal vectors $(v_1, \dots, v_r) \in (\mathbb{K}^I)^r$, orthonormal vectors $(w_1, \dots, w_r) \in (\mathbb{K}^J)^r$, and *singular values* $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$ so that

$$M = \sum_{\nu=1}^r \sigma_\nu v_\nu w_\nu^\top. \quad (\text{A.22})$$

The vectors v_ν (w_ν) are called the left (right) *singular vectors*.

Remark A.44. According to Golub–Van Loan [157, §5.4.5], computing the singular value decomposition (SVD) of an $n \times n$ matrix costs about $21n^3$ operations.

One of the applications of SVD is optimal truncation to smaller rank. Let $s \in \mathbb{N}_0$ with $0 \leq s < r = \text{rank}(M)$ be the target rank of an approximation R to M . Set

$$M_s := \sum_{\nu=1}^s \sigma_\nu v_\nu w_\nu^\top.$$

The next statement uses the norms introduced in §B.1.

Proposition A.45. M_s is the best approximation of rank $\leq s$ to M with respect to the spectral norm $\|\cdot\| = \|\cdot\|_2$ and the Frobenius norm $\|\cdot\| = \|\cdot\|_F$, i.e.,

$$\|M - M_s\| \leq \|M - X\| \quad \text{for all } X \in \mathbb{C}^{I \times J} \text{ with } \text{rank}(X) \leq s.$$

The respective errors are

$$\|M - M_s\|_2 = \sigma_{s+1}, \quad \|M - M_s\|_F = \sqrt{\sum_{\nu=s+1}^r (\sigma_\nu)^2}.$$

Proof. See [198, §2.5 and §C.5.1]. \square

Appendix B

Facts from Normed Spaces

Abstract In Section B.1, the general norm as well as the Euclidean and maximum norm are introduced. Matrix norms corresponding to a vector norm and the condition of a matrix are discussed in §B.1.3. Section B.2 refers to Hilbert norms in general, and to the Euclidean and spectral norm in particular. The relation between norms and the spectral radius is investigated in Section B.3. The estimates and convergence results in §§B.3.2–B.3.3 are essential for analysing linear iterations. The numerical radius introduced in §B.3.4 is a matrix norm of practical interest.

B.1 Norms

B.1.1 Vector Norms

Let V be vector space over the field $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. A mapping $\|\cdot\| : V \rightarrow [0, \infty)$ is called a *norm* on V if

$$\|x\| = 0 \quad \text{only for } x = 0, \tag{B.1a}$$

$$\|x + y\| \leq \|x\| + \|y\| \quad \text{for all } x, y \in V \quad (\text{triangle inequality}), \tag{B.1b}$$

$$\|\lambda x\| = |\lambda| \|x\| \quad \text{for all } \lambda \in \mathbb{K}, x \in V. \tag{B.1c}$$

$\|\cdot\|$ is also used as a norm symbol. Special norms are indicated by subscripts.

In the following, let V be a finite-dimensional vector space. The standard finite-dimensional vector space is $V = \mathbb{K}^I$, where I is a finite index set.

Example B.1. The *maximum norm* $\|\cdot\|_\infty$ and the *Euclidean norm* $\|\cdot\|_2$ on $V = \mathbb{K}^I$ are defined as follows:

$$\|x\|_\infty := \max\{|x_\alpha| : \alpha \in I\}, \quad \|x\|_2 := \sqrt{\sum_{\alpha \in I} |x_\alpha|^2}. \tag{B.2}$$

Exercise B.2. (a) Check the properties (B.1a–c) for the norms in (B.2). Prove:

- (b) Let $c > 0$. If $\|\cdot\|$ is a norm on V , then $\|\cdot\| := c \|\cdot\|$ is a norm, too.
- (c) If $\|\cdot\|$ is a norm on $V = \mathbb{K}^I$ and if $A \in \mathbb{K}^{I \times I}$ is a regular matrix, then $\|\cdot\| := \|A\cdot\|$ is also a norm on V .

Lemma B.3. *The following inverse ‘triangle inequality’ holds:*

$$\|\|x\| - \|y\|\| \leq \|x - y\| \quad \text{for all } x, y \in V. \quad (\text{B.3})$$

Each norm defines a topology on V and therefore the continuity of mappings in the normed vector space $(V, \|\cdot\|)$. Inequality (B.3) leads to the following conclusion.

Conclusion B.4. *The norm $\|\cdot\|$ is a continuous (even Lipschitz continuous) mapping from $(V, \|\cdot\|)$ into \mathbb{R} .*

B.1.2 Equivalence of All Norms

Two norms $\|\cdot\|$ and $\|\cdot\|$ on V are called equivalent if there is a constant C such that

$$\frac{1}{C} \|x\| \leq \|\cdot\| \leq C \|x\| \quad \text{for all } x \in V. \quad (\text{B.4})$$

Exercise B.5. Prove: (a) Transitivity holds: If $(\|\cdot\|_a, \|\cdot\|_b)$ and $(\|\cdot\|_b, \|\cdot\|_c)$ are two pairs of equivalent norms, then $\|\cdot\|_a$ and $\|\cdot\|_c$ are also equivalent.

(b) The Euclidean norm and maximum norm satisfy the inequalities

$$\|x\|_\infty \leq \|x\|_2, \quad \|x\|_2 \leq \sqrt{\dim(V)} \|x\|_\infty \quad \text{for all } x \in V. \quad (\text{B.5})$$

Since in this book we always assume finite-dimensional vector spaces, the assumption of the following theorem is satisfied.

Theorem B.6. *If $\dim(V) < \infty$, all norms on V are equivalent.*

Proof. (i) Let $\{\mathbf{e}_\alpha : \alpha \in I\}$ be the basis of V . We define a reference norm by $\|x\| := \max\{|a_\alpha| : \alpha \in I\}$ with a_α from the representation $x = \sum_\alpha a_\alpha \mathbf{e}_\alpha$. Let $\|\cdot\|$ be an arbitrary norm on V . Because of the transitivity (cf. Exercise B.5a), it is sufficient to show the equivalence of $\|\cdot\|$ and $\|\cdot\|$:

(ii) The second part of (B.4), $\|\cdot\| \leq c \|\cdot\|$, follows from the triangle inequality with $c := \sum_\alpha \|\mathbf{e}_\alpha\|$:

$$\|\cdot\| = \left\| \sum_\alpha a_\alpha \mathbf{e}_\alpha \right\| \leq \sum_\alpha |a_\alpha| \|\mathbf{e}_\alpha\| \leq c \max_\alpha |a_\alpha| = c \|\cdot\|.$$

(iii) The set $S := \{x \in V : \|x\| = 1\}$ is bounded in $(V, \|\cdot\|)$ (the bound is 1) and closed because it is the inverse image of the value 1 under a continuous mapping

(cf. Conclusion B.4). Since $\dim(V) < \infty$, the set S is compact. Inequality (B.3) and part (ii) yield $|\|\cdot\|| - \|\cdot\| \leq \|\cdot\| \leq c\|\cdot\|$, i.e., $|\|\cdot\||$ is also continuous with respect to the normed space $(V, \|\cdot\|)$. Since a continuous function on a compact set attains its minimum, there is some $x_0 \in S$ with

$$\|\cdot\|_{x_0} \leq \|\cdot\| \quad \text{for all } x' \in S.$$

(iv) Since (B.4) is trivial for $x = 0$, we assume $x \neq 0$. By (B.1a) we have $\xi := \|x\| > 0$, hence $x' := x/\xi$ is well defined and satisfies $\|x'\| = 1$, i.e., $x' \in S$ holds. Part (iii) yields

$$\|x\| = \xi \leq \xi \|\cdot\| / \|\cdot\|_{x_0} = c_0 \xi \|\cdot\| = c_0 \|\cdot\|_{\xi x'} = c_0 \|\cdot\|_x$$

for $c_0 := 1 / \|\cdot\|_{x_0}$. Hence, (B.4) is proved with $C := \max\{c, c_0\}$. \square

Remark B.7. The constant C in (B.4) does not depend on $x \in V$ but does depend on V , more precisely on $\dim(V)$, as illustrated by the second inequality in (B.5).

B.1.3 Corresponding Matrix Norms

The space $\mathbb{K}^{I \times I}$ of the $I \times I$ matrices is also a linear vector space of dimension $(\#I)^2$; hence, one may define norms on $\mathbb{K}^{I \times I}$. These are called *matrix norms*, whereas norms on \mathbb{K}^I are called *vector norms*. Usually, we use matrix norms corresponding to some vector norm (cf. Definition B.9). Other well-known matrix norms are the Frobenius norm defined below and the numerical radius discussed in §B.3.4.

Example B.8. The *Frobenius norm* is $\|A\|_F := \sqrt{\sum_{\alpha, \beta \in I} |a_{\alpha\beta}|^2}$.

Since the matrices together with the matrix multiplication form an algebra, the following subclass of matrix norms is of special interest (cf. (B.9a)).

Definition B.9. Let $\|\cdot\|$ be a vector norm on \mathbb{K}^I . The *corresponding* (or *associated*) matrix norm¹ is

$$\|\cdot\| := \sup \{\|Ax\| / \|x\| : x \neq 0\}. \quad (\text{B.6})$$

Exercise B.10. Prove: (a) $\|\cdot\|$ in (B.6) is a norm on $\mathbb{K}^{I \times I}$.

(b) The supremum in (B.6) is attained by some x , so that ‘sup’ may be replaced with ‘max’.

(c) Two vector norms differing only by a factor (cf. Exercise B.2b) lead to identical corresponding matrix norms.

(d) $C := \|\cdot\|$ is the smallest possible bound in the inequality

$$\|Ax\| \leq C \|x\| \quad \text{for all } x \in \mathbb{K}^I. \quad (\text{B.7})$$

¹ Considering the matrix as a mapping (operator), this norm is also called the *operator norm*.

In the following we shall always denote the vector norm and the corresponding matrix norm by the same norm symbol. A confusion is impossible because of the disjoint domains of definition. In particular, $\|A\|_\infty$ and $\|A\|_2$ are the matrix norms corresponding to the maximum norm $\|x\|_\infty$ and the Euclidean norm $\|x\|_2$, respectively. Because of its affinity to the spectral radius (cf. §B.2.2), $\|A\|_2$ is called the *spectral norm*. $\|A\|_\infty$ is called the *row-sum norm* since the following characterisation (B.8) uses row sums.

Exercise B.11. Prove: (a) $\|A\|_\infty$ has the representation

$$\|A\|_\infty = \max \left\{ \sum_{\beta \in I} |a_{\alpha\beta}| : \alpha \in I \right\} \quad \text{for all } A \in \mathbb{K}^{I \times I}. \quad (\text{B.8})$$

- (b) $\|D\|_\infty = \|D\|_2 = \max_{\alpha \in I} |d_\alpha| = \rho(D)$ holds for a diagonal matrix.
- (c) The spectral norm $\|A\|_2$ of real matrices is independent of whether $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$ is chosen in (B.6). Hint: Use (B.21a).

Theorem B.12. Let $\|\cdot\|$ denote the vector norm on \mathbb{K}^I as well as the corresponding matrix norm (B.6) on $\mathbb{K}^{I \times I}$. Then

$$\|AB\| \leq \|A\| \|B\| \quad \text{for } A, B \in \mathbb{K}^{I \times I} \quad (\text{submultiplicativity}), \quad (\text{B.9a})$$

$$\|Ax\| \leq \|A\| \|x\| \quad \text{for } A \in \mathbb{K}^{I \times I}, x \in \mathbb{K}^I. \quad (\text{B.9b})$$

Proof. (b) By Exercise B.10, (B.7) holds with $C := \|\|A\|\|$.

(a) Apply (B.9b) to Bx instead of x : $\|ABx\| \leq \|A\| \|Bx\|$. Another application of (B.9b) to Bx yields $\|ABx\| \leq \|A\| \|B\| \|x\|$. Hence, (B.7) is satisfied by $C := \|A\| \|B\|$ and AB instead of B . Exercise B.10d shows that $\|AB\| \leq C$. \square

Exercise B.13. Prove: (a) $\|I\| = 1$ holds for all associated norms.

(b) The Frobenius norm in Example B.8 does not correspond to any vector norm, provided that $\#I > 1$.

(c) Given a vector norm $\|\cdot\|$ and a regular matrix T , we may define an additional vector norm $\|\cdot\|_T$ by

$$\|\|x\|\|_T := \|Tx\| \quad (\text{B.10a})$$

(cf. Exercise B.2c). The equally denoted corresponding matrix norm satisfies

$$\|\|A\|\|_T = \|TAT^{-1}\| \quad \text{for all } A \in \mathbb{K}^{I \times I}. \quad (\text{B.10b})$$

Definition B.9 associates each vector norm with a matrix norm. This mapping is not injective (cf. Exercise B.10d). However, for any corresponding matrix norm $\|\cdot\|_M$, one can reconstruct the underlying vector norm $\|\cdot\|_V$ up to a factor as follows. Choose any $0 \neq a \in \mathbb{K}^I$. The product $x a^\top$ ($x \in \mathbb{K}^I$) represents the matrix $(x_\alpha a_\beta)_{\alpha, \beta \in I}$. $\|x\|_a := \|x a^\top\|_M$ is a vector norm differing from $\|\cdot\|_V$ only by a factor. Let $\|\cdot\|_{a, M}$ be the matrix norm corresponding to $\|\cdot\|_a$. Then an arbitrary matrix norm $\|\cdot\|_M$ corresponds to some vector norm if and only if $\|\cdot\|_M = \|\cdot\|_{a, M}$.

Let X and Y be two normed spaces with the norms $\|\cdot\|_X$ and $\|\cdot\|_Y$, and let $A : X \rightarrow Y$ be a linear mapping. Then

$$\|A\|_{Y \leftarrow X} := \sup \{ \|Ax\|_Y / \|x\|_X : 0 \neq x \in X \} \quad (\text{B.11})$$

denotes the corresponding matrix norm.

B.1.4 Condition and Spectral Condition Number

The *condition* of a regular matrix with respect to the matrix norm $\|\cdot\|$ is defined by

$$\text{cond}(A) := \|A\| \|A^{-1}\|. \quad (\text{B.12})$$

In particular, $\text{cond}_2(A) = \|A\|_2 \|A^{-1}\|_2$ denotes the *Euclidean condition* (associated with the Euclidean norm). The condition describes the behaviour of the linear system $Ax = b$ under perturbation.

Proposition B.14. *Let the matrix A be regular and $b \neq 0$. Let x be the solution of $Ax = b$, while the perturbed right-hand side $b + \delta b$ yields the solution $x + \delta x$; i.e., $A(x + \delta x) = b + \delta b$. The relative errors satisfy*

$$\frac{\|\delta x\|}{\|x\|} \leq \text{cond}(A) \frac{\|\delta b\|}{\|b\|}$$

(the matrix norm in $\text{cond}(A)$ corresponds to the vector norm $\|\cdot\|$). For a similar statement involving a perturbation of the matrix A and for a proof, we refer to Quarteroni–Sacco–Saleri [314, Theorem 3.1] and Björck [48, §1.2.7].

By definition, the condition $\text{cond}(\cdot)$ depends on the underlying matrix norm. In order to be independent of a reference norm, we define the *spectral condition number*

$$\kappa(A) := \rho(A)\rho(A^{-1}). \quad (\text{B.13})$$

- Exercise B.15.** Prove:
- (a) $\text{cond}_2(A) = \kappa(A)$ holds for normal matrices A .
 - (b) $\kappa(A) = \max\{|\lambda| : \lambda \in \sigma(A)\} / \min\{|\lambda| : \lambda \in \sigma(A)\}$ holds for regular A .
 - (c) If A has a positive spectrum with the minimal eigenvalue $\lambda_{\min}(A)$ and the maximal eigenvalue $\lambda_{\max}(A)$, definition (B.13) reduces to

$$\kappa(A) = \lambda_{\max}(A)/\lambda_{\min}(A). \quad (\text{B.14})$$

- (d) Positive definite matrices fulfil $\text{cond}_2(A) = \kappa(A)$ with the representation (B.14), where the extreme eigenvalues are also given by

$$\lambda_{\max}(A) = \|A\|_2, \quad \lambda_{\min}(A) = 1/\|A^{-1}\|_2. \quad (\text{B.15})$$

Hint: For (a), use (B.21b).

B.2 Hilbert Norm

B.2.1 Elementary Properties

We recall the well-known properties of the scalar product (cf. §A.5).

Remark B.16. Each scalar product induces the norm

$$\|x\| := \sqrt{\langle x, x \rangle} \quad (\text{B.16})$$

on V . The Schwarz inequality holds:

$$|\langle x, y \rangle| \leq \|x\| \|y\| \quad \text{for } x, y \in \mathbb{K}^I.$$

An equal sign in the latter inequality implies that x and y are linearly dependent. Furthermore, a dual statement holds:

$$\|x\| = \max \{ |\langle x, y \rangle| / \|y\| : 0 \neq y \in V \}. \quad (\text{B.17})$$

Remark B.17. (a) The norm (B.16) induced by the Euclidean scalar product is the Euclidean norm $\|\cdot\|_2$ in (B.2).

(b) The Hermitian is the adjoint matrix with respect to $\langle \cdot, \cdot \rangle$:

$$\langle Ax, y \rangle = \langle x, A^H y \rangle \quad \text{for } x, y \in \mathbb{K}^I, A \in \mathbb{K}^{I \times I}. \quad (\text{B.18})$$

In the following, $\langle \cdot, \cdot \rangle$ is fixed as the Euclidean scalar product. Concerning general scalar products, we refer to Remark C.10.

B.2.2 Spectral Norm

In §B.1.3, the spectral norm $\|\cdot\|_2$ is defined as the matrix norm corresponding to the Euclidean vector norm.

Lemma B.18. *The Euclidean norm and spectral norm are invariant with respect to unitary transformations in the following sense. A unitary matrix $Q \in \mathbb{K}^{I \times I}$ satisfies*

$$\|Qx\|_2 = \|x\|_2 \quad \text{for all } x \in \mathbb{K}^I, \quad (\text{B.19a})$$

$$\|Q\|_2 = \|Q^H\|_2 = 1, \quad (\text{B.19b})$$

$$\begin{aligned} \|QA\|_2 &= \|AQ\|_2 = \|Q^H A\|_2 = \|AQ^H\|_2 = \|Q A Q^H\|_2 = \|Q^H A Q\|_2 \\ &= \|A\|_2 \end{aligned} \quad (\text{B.19c})$$

Proof. (a) We have $\|Qx\|_2^2 = \langle Qx, Qx \rangle = \langle x, Q^H Qx \rangle = \langle x, x \rangle = \|x\|_2^2$ thanks to (B.16), Definition A.2 and (B.18).

(b) As, by Remark A.3b, Q^H is unitary if Q is so, it is sufficient to prove the assertions for Q . (B.19b) follows from definition (B.6) by using (B.19a).

(c) (B.9a) and (B.19b) yield $\|QA\|_2 \leq \|Q\|_2 \|A\|_2 = \|A\|_2$. The same estimate with Q^H and QA instead of Q and A shows that $\|A\|_2 = \|Q^H QA\|_2 \leq \|QA\|_2$, and therefore $\|QA\|_2 = \|A\|_2$. All further statements in (B.19c) can be proved analogously or follow from the previous ones. \square

Lemma B.19. *An equivalent definition of the spectral norm is*

$$\|A\|_2 = \max \left\{ \frac{|\langle Ax, y \rangle|}{\|x\|_2 \|y\|_2} : 0 \neq x, y \in \mathbb{K}^I \right\}.$$

Proof. Express $\|Ax\|_2$ in (B.6) by using (B.17). \square

Lemma B.19, (B.18), and (A.11c) prove the next statement.

Corollary B.20. $\|A^H\|_2 = \|\bar{A}\|_2 = \|A^T\|_2 = \|A\|_2$.

Exercise B.21. (a) For all matrices $A \in \mathbb{K}^{I \times I}$, we have $\|A\|_2 \leq \sqrt{\|A\|_\infty \|A^H\|_\infty}$, $\|A\|_2 \leq \sqrt{n} \|A\|_\infty$, and $\|A\|_\infty \leq \sqrt{n} \|A\|_2$ with $n := \#I$.

(b) $\|A\|_2 \leq \|A\|_\infty$ for Hermitian A .

(c) $|a_{\alpha\beta}| \leq \|A\|_2$ for all matrix entries of A .

(d) $|a_{\alpha\beta}| \leq C$ for all $\alpha, \beta \in I$ implies $\|A\|_2 \leq nC$.

Exercise B.22. Prove that $\|Ax\|_2 = \|x\|_2$ holds for orthogonal matrices A (cf. Definition A.27).

An important problem is the *least squares problem* mentioned in §1.1 and Remark 5.17. Given a rectangular matrix $A \in \mathbb{K}^{n \times m}$ of full rank with $n > m$ and $b \in \mathbb{K}^n$, we want to

$$\text{minimise } \|Ax - b\|_2 \text{ over all } x \in \mathbb{K}^m.$$

The approach by Gauss uses the characterisation of the optimal least squares solution x by the *normal equations* $A^H Ax = A^H b$. The squared condition of $A^H A$ can be avoided by the following approach. The QR decomposition $A = QR$ and Exercise B.22 lead to $\|Ax - b\|_2 = \|QRx - b\|_2 = \|Q(Rx - Q^H b)\|_2 = \|Rx - Q^H b\|_2$ and prove the following.

Remark B.23. For A and b above, determine the QR decomposition $A = QR$. Then the minimiser is given by

$$x^* = R^{-1} Q^H b.$$

The minimised least squares are

$$\|Ax^* - b\|_2 = \|(I - QQ^H)b\|_2.$$

B.3 Correlation Between Norms and Spectral Radius

B.3.1 Spectral Norm and Spectral Radius

Lemma B.24. Let $\|\cdot\|$ be a corresponding matrix norm. Then

$$|\lambda| \leq \|A\| \quad \text{for all eigenvalues } \lambda \text{ of the matrix } A, \quad (\text{B.20a})$$

$$\rho(A) \leq \|A\| \quad \text{for all square matrices } A. \quad (\text{B.20b})$$

Proof. According to Lemma A.6, for each eigenvalue λ there is an eigenvector e with $Ae = \lambda e$. (B.1c) and (B.9b) yield $|\lambda| \|e\| = \|\lambda e\| = \|Ae\| \leq \|A\| \|e\|$; hence, we obtain (B.20a). Inequality (B.20b) follows from (B.20a). \square

In §B.1.3, the spectral norm $\|\cdot\|_2$ is defined as the matrix norm corresponding to the Euclidean vector norm. The term ‘spectral norm’ is due to the close relation to spectral properties of the matrix. For normal matrices this norm coincides with the spectral radius. Even in the general case, it can be expressed by the spectral radius as shown below.

Theorem B.25. The spectral norm satisfies

$$\|A\|_2 = \sqrt{\rho(A^H A)} = \sqrt{\rho(A A^H)} \quad \text{for all } A \in \mathbb{K}^{I \times I}, \quad (\text{B.21a})$$

$$\|A\|_2 = \rho(A) \quad \text{for all normal matrices } A \in \mathbb{K}^{I \times I}. \quad (\text{B.21b})$$

The statement (B.21a) also holds for rectangular matrices $A \in \mathbb{K}^{I \times J}$.

Proof. (i) By Definition B.9, the square $\|Ax\|_2^2$ is the maximum of

$$\|Ax\|_2^2 / \|x\|_2^2 = \langle Ax, Ax \rangle / \langle x, x \rangle = \langle A^H Ax, x \rangle / \langle x, x \rangle$$

over all $x \neq 0$. The Hermitian matrix $A^H A$ has the representation $Q D Q^H$ with the diagonal matrix $\text{diag}\{\lambda_\alpha\}$ constructed from the eigenvalues λ_α of $A^H A$. These eigenvalues are positive (cf. Exercise C.11a, Lemma C.3). There is some $\beta \in I$ with $\rho(A^H A) = \lambda_\beta$. Substitution $y = Q^H x$ and (B.19a) yield $\|Ax\|_2^2 / \|x\|_2^2 = \langle y, D y \rangle / \langle y, y \rangle$. The latter expression is maximal for the unit vector $y = \mathbf{e}_\beta$ and yields the value $\|A\|_2^2 = \rho(A^H A)$. The second equality in (B.21a) follows from Lemma A.20.

(ii) Given a normal matrix A , due to Theorem A.40, we find a unitary matrix Q and a diagonal matrix D such that $A = Q D Q^H$. (B.19c) shows that $\|A\|_2 = \|D\|_2$. According to Exercise B.11b, we have $\|D\|_2 = \rho(D)$. As A and D are similar, $\rho(A) = \rho(D)$ holds (cf. (A.6c)) and proves (B.21b). \square

$A^H A$ and AA^H are Hermitian matrices and therefore normal. Hence, we deduce from (B.21a,b) that

$$\|A\|_2^2 = \|A^H A\|_2 = \|A A^H\|_2 \quad \text{for all } A \in \mathbb{K}^{I \times J}.$$

B.3.2 Matrix Norms Approximating the Spectral Radius

Lemma B.26. For any matrix $A \in \mathbb{K}^{I \times I}$ and any $\varepsilon > 0$, there exists a vector norm and corresponding matrix norm $\|\cdot\|_{A,\varepsilon}$ with the property

$$\rho(A) \leq \|A\|_{A,\varepsilon} \leq \rho(A) + \varepsilon. \quad (\text{B.22})$$

Proof. Let $A = Q U Q^H$ be the Schur normal form (A.13). The eigenvalues of A are the diagonal elements $\lambda_i := u_{ii}$ of U (cf. Exercise A.7). Hence, the diagonal matrix $D := \text{diag}\{\lambda_1, \dots, \lambda_n\}$, $n := \#I$, satisfies (B.23) (cf. (B.21b)):

$$\rho(A) = \rho(D) = \|D\|_2. \quad (\text{B.23})$$

We define $\xi := \min\{1, \varepsilon / [n \|A\|_2]\}$ and apply a similarity transformation with the diagonal matrix $X := \text{diag}\{1, \xi, \xi^2, \dots, \xi^{n-1}\}$ to U :

$$V := X^{-1} U X = \begin{bmatrix} \lambda_1 & \xi u_{12} & \xi^2 u_{13} & \dots \\ 0 & \lambda_2 & \xi u_{23} & \dots \\ \vdots & & \ddots & \end{bmatrix} = D + R, \quad R_{ij} = \begin{cases} \xi^{j-i} u_{ij} & \text{for } i < j, \\ 0 & \text{for } i \geq j. \end{cases}$$

By Exercise B.21c and (B.19c),

$$|R_{ij}| \leq \xi^{j-i} |u_{ij}| \leq \xi \|U\|_2 = \xi \|A\|_2 \leq \varepsilon/n$$

holds for $i < j$ by the choice of ξ . Exercise B.21d yields $\|R\|_2 \leq \varepsilon$. We define the vector norm $\|x\|_{A,\varepsilon} := \|X^{-1} Q^H x\|_2$ (cf. Exercise B.2c). The corresponding matrix norm is $\|A\|_{A,\varepsilon} = \|X^{-1} Q^H A Q X\|_2$ (cf. (B.10b)). Hence, we arrive at

$$\|A\|_{A,\varepsilon} = \|X^{-1} Q^H A Q X\|_2 = \|X^{-1} U X\|_2 = \|V\|_2 \leq \|D\|_2 + \|R\|_2.$$

(B.23) and $\|R\|_2 \leq \varepsilon$ yield $\|A\|_{A,\varepsilon} \leq \rho(A) + \varepsilon$. The first part of inequality (B.22) is trivial because of (B.20b). \square

The following theorem demonstrates the asymptotic relationship between an arbitrary matrix norm and the spectral radius. Note that the matrix norm need not be corresponding to a vector norm.

Theorem B.27. For all $A \in \mathbb{K}^{I \times I}$ and any matrix norm, the spectral radius is the following limit:

$$\rho(A) = \lim_{m \rightarrow \infty} \|A^m\|^{1/m}. \quad (\text{B.24})$$

Proof. (i) Let $\|\cdot\|$ and $\|\cdot\|$ be two matrix norms. First, we prove that the limes superior does not depend on the choice of the underlying matrix norm, i.e.,

$$\overline{\lim} \|\cdot\| A^m \|\|^{1/m} = \overline{\lim} \|A^m\|^{1/m}.$$

By the equivalence of norms stated in Theorem B.6, there is a constant C with $\|\cdot\| \leq C\|\cdot\|$, so that $\|A^m\|^{\frac{1}{m}} \leq C^{\frac{1}{m}} \|A^m\|^{\frac{1}{m}}$. Since $C^{\frac{1}{m}} \rightarrow 1$ for $m \rightarrow \infty$, $\overline{\lim} \|A^m\|^{\frac{1}{m}} \leq \overline{\lim} \|A^m\|^{\frac{1}{m}}$. Interchanging the roles of the two norms, we obtain the reverse inequality. This proves the above statement.

(ii) Next, we consider the case $\rho(A) = 0$. The Schur normal form yields $U = Q^H A Q$ with $\rho(U) = 0$, i.e., U is strictly triangular (cf. Exercise A.19a). The assertion follows from Lemma A.13: $A^m = 0$ holds for all $m > \#I$.

(iii) Now assume $\rho := \rho(A) > 0$ and define $B := -\frac{1}{\rho}A$. Hence, the assertion is equivalent to

$$\lim \|B^m\|^{1/m} = 1$$

because of $\rho(B) = 1$. Using the norm $\|\cdot\| = \|\cdot\|_{B,\varepsilon}$ ($\varepsilon > 0$) in Lemma B.26, inequality (B.22) shows that

$$1 = \rho(B) = \rho(B^m)^{\frac{1}{m}} \leq \|B^m\|^{\frac{1}{m}} \leq (\|B\|^m)^{\frac{1}{m}} = \|B\| \leq \rho(B) + \varepsilon = 1 + \varepsilon$$

for all m . Hence, $1 \leq \underline{\lim} \|B^m\|^{1/m} \leq \overline{\lim} (\|B^m\|)^{1/m} \leq 1 + \varepsilon$. By part (i), $\overline{\lim} (\|B^m\|)^{1/m}$ is independent of the norm $\|\cdot\| = \|\cdot\|_{B,\varepsilon}$ and, therefore, holds for all $\varepsilon > 0$. The resulting inequality $1 \leq \underline{\lim} \|B^m\|^{1/m} \leq \overline{\lim} \|B^m\|^{1/m} \leq 1$ implies the existence of the limit $\lim \|B^m\|^{1/m} = 1$. \square

B.3.3 Geometrical Sum of Matrices

The finite geometrical series (Neumann's series) satisfies

$$\left[\sum_{\nu=0}^{m-1} A^\nu \right] [I - A] = I - A^m. \quad (\text{B.25})$$

If 1 is not an eigenvalue of A , i.e., if $I - A$ is regular, (B.25) can be rewritten as

$$\sum_{\nu=0}^{m-1} A^\nu = (I - A^m) (I - A)^{-1}.$$

Lemma B.28. Let $A \in \mathbb{K}^{I \times I}$. $\lim_{m \rightarrow \infty} \|A^m\| = 0$ holds if and only if $\rho(A) < 1$.

Proof. (i) Assume $\rho := \rho(A) < 1$ and choose any ρ' with $\rho < \rho' < 1$. We conclude from (B.24), i.e., $\|A^m\|^{\frac{1}{m}} \rightarrow \rho$, that $\|A^m\| < \rho'^m$ for $m \geq m_0$ with sufficiently large m_0 . Hence, $\rho(A) < 1$ is sufficient for $\|A^m\| \rightarrow 0$.

(ii) In the remaining case of $\rho := \rho(A) \geq 1$, inequality (B.20b) shows that $\|A^m\| \geq \rho(A^m) = \rho(A)^m \geq 1$ (cf. (A.6b)). Hence, $\rho(A) < 1$ is also a necessary condition for $\|A^m\| \rightarrow 0$. \square

Theorem B.29. *If and only if $\rho(A) < 1$, the geometrical sum converges and results in the value*

$$\sum_{\nu=0}^{\infty} A^{\nu} = (I - A)^{-1}. \quad (\text{B.26})$$

Proof. (i) Assume $\rho(A) < 1$. By Lemma B.28 we can go to the limit $m \rightarrow \infty$ in (B.25) and obtain

$$\left(\sum_{\nu=0}^{\infty} A^{\nu} \right) (I - A) = I.$$

On the other hand, $\rho(A) < 1$ implies that $I - A$ is regular, i.e., (B.26) follows.

(ii) For $\rho(A) \geq 1$, according to Lemma B.28, the terms A^{ν} do not converge to zero. Hence, the geometrical sum must diverge. \square

B.3.4 Numerical Radius of a Matrix

An intermediate position between the spectral radius and the spectral norm is taken by a matrix norm called the *numerical radius* of the matrix A :

$$r(A) := \max\{|\langle Ax, x \rangle| / \|x\|_2^2 : 0 \neq x \in \mathbb{C}^I\}. \quad (\text{B.27})$$

The set

$$\mathcal{F}(A) := \{\langle Ax, x \rangle / \|x\|_2^2 : 0 \neq x \in \mathbb{C}^I\} \subset \mathbb{C}$$

is called the *field of values* of A (cf. Greenbaum [167, §1.3.6]). Since $\mathcal{F}(A)$ is compact, the maximum in (B.27) exists.

The interesting property is that $r(A)$ can be estimated against the spectral norm with dimension-independent equivalent constants (cf. (B.28b,d)).

Lemma B.30. *The numerical radius $r(A)$ has the following properties:*

$$r(A^H) = r(A) \quad \text{for all } A \in \mathbb{C}^{I \times I}, \quad (\text{B.28a})$$

$$\rho(A) \leq r(A) \leq \|A\|_2 \quad \text{for all } A \in \mathbb{C}^{I \times I}, \quad (\text{B.28b})$$

$$r(A) = \|A\|_2 = \rho(A) \quad \text{for all normal } A \in \mathbb{C}^{I \times I}, \quad (\text{B.28c})$$

$$\|A\|_2 \leq 2r(A) \quad \text{for all } A \in \mathbb{C}^{I \times I}. \quad (\text{B.28d})$$

Proof. (i) (B.28a) follows from (B.18) and (A.11c): $|\langle Ax, x \rangle| = |\langle A^H x, x \rangle|$.

(ii) We conclude from $|\langle Ax, x \rangle| \leq \|Ax\|_2 \|x\|_2 \leq \|A\|_2 \|x\|_2^2$ that $r(A) \leq \|A\|_2$. Let x be the eigenvector of A associated with the eigenvalue λ of maximal modulus: $|\lambda| = \rho(A)$. $|\langle Ax, x \rangle| = |\lambda| \langle x, x \rangle$ leads to $r(A) \geq |\lambda| = \rho(A)$.

(iii) (B.28b) and (B.21b) prove (B.28c).

(iv) Any matrix A has the unique decomposition

$$A = A_0 + iA_1, \quad A_0 := \frac{1}{2}(A + A^H), \quad A_1 := \frac{1}{2i}(A - A^H) \quad (\text{B.29})$$

into the Hermitian part A_0 and the skew-Hermitian part iA_1 of A . Obviously, A_0 and A_1 are Hermitian: $A_0 = A_0^H$, $A_1 = A_1^H$. Hermitian matrices are normal; hence, $\|A_k\|_2 = r(A_k)$ ($k = 0, 1$) and

$$\|A\|_2 \leq \|A_0\|_2 + \|A_1\|_2 = r(A_0) + r(A_1). \quad (\text{B.30})$$

$\langle By, y \rangle$ has a real value for all Hermitian matrices B and all y . Hence, $\langle Ax, x \rangle$ consists of the real part $\langle A_0 x, x \rangle = \lambda \langle x, x \rangle$ and of the imaginary part $\langle A_1 x, x \rangle$. Setting $\zeta := \langle Ax, x \rangle / \|x\|_2^2$, we obtain that

$$|\langle A_k x, x \rangle| / \|x\|_2^2 \leq |\zeta| \leq r(A) \quad \text{for } k = 0, 1 \text{ and all } x \neq 0.$$

Maximisation over all $x \neq 0$ yields $r(A_k) \leq r(A)$. From (B.30), we conclude (B.28d). \square

Theorem B.31. *The spectral radius $r(\cdot)$ is a matrix norm, but submultiplicativity is restricted² to powers of A (cf. Pearcey [309]):*

$$r(A^n) \leq r(A)^n \quad \text{for } n \in \mathbb{N}_0, \quad A \in \mathbb{C}^{I \times I}. \quad (\text{B.31})$$

Proof. In (i)–(iii) we prove the norm properties.

(i) Because of (B.28d), $r(A) = 0$ implies $\|A\|_2 = 0$ and therefore $A = 0$. Hence (B.1a) holds.

(ii) Let $x \neq 0$ with $\|x\|_2 = 1$ be the maximiser in the definition of $r(A + B)$, i.e., $r(A + B) = |\langle (A + B)x, x \rangle|$. Then

$$\begin{aligned} r(A + B) &= |\langle (A + B)x, x \rangle| = |\langle Ax, x \rangle + \langle Bx, x \rangle| \\ &\leq |\langle Ax, x \rangle| + |\langle Bx, x \rangle| \leq r(A) + r(B) \end{aligned}$$

proves the triangle inequality (B.1b).

(iii) The property (B.1c) is trivial: $r(\lambda A) = |\lambda|r(A)$ ($\lambda \in \mathbb{C}$, $A \in \mathbb{C}^{I \times I}$).

(iv) Concerning (B.31), we repeat the proof of [309], which is rather untypical for matrix estimations. The roots of the polynomial $1 - z^n$ are given by the powers of $\zeta := \exp(2\pi i/n)$. This leads to the factorisation

$$1 - z^n = \prod_{k=1}^n (1 - \zeta^k z). \quad (\text{B.32a})$$

² A simple counterexample concerning submultiplicativity is given by Pearcey [309].

To prove the polynomial identity

$$1 = \frac{1}{n} \sum_{j=1}^n \prod_{k \neq j} (1 - \zeta^k z) \quad (\text{B.32b})$$

($\prod_{k \neq j}$ abbreviates $\prod_{k \in \{1, \dots, n\} \setminus \{j\}}$), we have to check this identity at n different points (uniqueness of interpolation by a polynomial of degree $n-1$). Denote the right-hand side in (B.32b) by $Q(z)$. For $z = 1$, all products $\prod_{k \neq j}$ vanish except for $j = n$. Hence

$$Q(1) = \frac{1}{n} \prod_{1 \leq k \leq n-1} (1 - \zeta^k)$$

follows. Since

$$\prod_{1 \leq k \leq n-1} (z - \zeta^k) = \frac{z^n - 1}{z - 1} = \sum_{\mu=0}^{n-1} z^\mu$$

has a removable singularity at $z = 1$ with the value n , we obtain $Q(1) = 1$. One easily checks that $Q(z) = Q(\zeta z)$ because of $\zeta^n = 1$. Therefore, $Q(1) = 1$ proves $Q(\zeta^k) = 1$ for all k .

(v) Because of $r(\lambda A) = |\lambda| r(A)$ (cf. part (iii)), it is sufficient to prove the desired inequality (B.31) for matrices A with $r(A) = 1$. Therefore (B.31) is proved if we are able to show $r(A^n) \leq 1$, i.e.,

$$|\langle A^n x, x \rangle| \leq 1 \quad \text{for all } x \in \mathbb{C}^I \text{ with } \|x\|_2 = 1. \quad (\text{B.32c})$$

Choose any $x \in \mathbb{C}^I$ with $\|x\|_2 = 1$. First, we have to check the case that x is an eigenvector with $Ax = \zeta^k x$ for some $k \in \mathbb{N}$. Obviously, $|\langle A^n x, x \rangle| \leq 1$ is satisfied for this x . Excluding this case, the following vectors

$$x^{(j)} := \prod_{k \neq j} (1 - \zeta^k A) x \quad \text{and} \quad \xi^{(j)} := \frac{1}{\|x^{(j)}\|_2} x^{(j)} \quad \text{for } 1 \leq j \leq n \quad (\text{B.32d})$$

are well defined and satisfy $(1 - \zeta^j A) x^{(j)} = (1 - A^n) x$ (cf. (B.32a)). Using (B.32b) with A instead of z , we obtain

$$\begin{aligned} & 1 - \langle A^n x, x \rangle \Big|_{\|x\|_2=1} = \langle [I - A^n] x, x \rangle \\ & \stackrel{(\text{B.32b})}{=} \left\langle [I - A^n] x, \frac{1}{n} \sum_{j=1}^n \prod_{k \neq j} (1 - \zeta^k A) x \right\rangle \\ & = \frac{1}{n} \sum_{j=1}^n \left\langle [I - A^n] x, x^{(j)} \right\rangle \stackrel{(\text{B.32a})}{=} \frac{1}{n} \sum_{j=1}^n \left\langle \left[\prod_{k=1}^n (1 - \zeta^k A) \right] x, x^{(j)} \right\rangle \\ & \stackrel{(\text{B.32d})}{=} \frac{1}{n} \sum_{j=1}^n \left\langle [1 - \zeta^j A] x^{(j)}, x^{(j)} \right\rangle = \frac{\|x^{(j)}\|_2^2}{n} \sum_{j=1}^n \left\langle [1 - \zeta^j A] \xi^{(j)}, \xi^{(j)} \right\rangle \\ & = \frac{\|x^{(j)}\|_2^2}{n} \sum_{j=1}^n \left(1 - \zeta^j \langle A \xi^{(j)}, \xi^{(j)} \rangle \right). \end{aligned}$$

Using $\|\xi^{(j)}\|_2 = 1$, we can bound the real part of $1 - \zeta^j \langle A\xi^{(j)}, \xi^{(j)} \rangle$ by

$$1 - |\zeta^j \langle A\xi^{(j)}, \xi^{(j)} \rangle| = 1 - |\langle A\xi^{(j)}, \xi^{(j)} \rangle| \geq 1 - r(A) \geq 0.$$

This proves $\Re(1 - \langle A^n x, x \rangle) \geq 0$ and $\Re \langle A^n x, x \rangle \leq 1$. The assumption $r(A) = 1$ is not changed when we replace A with

$$B := \vartheta A, \quad \text{where } \vartheta \in \mathbb{C}, |\vartheta| = 1.$$

Since

$$\langle B^n x, x \rangle = \vartheta^n \langle A^n x, x \rangle,$$

we can choose ϑ such that

$$\Re \langle B^n x, x \rangle = |\langle A^n x, x \rangle|.$$

Using the previous result for B , we obtain

$$|\langle A^n x, x \rangle| = \Re \langle B^n x, x \rangle \leq 1 = r(A)$$

proving (B.32c). \square

Exercise B.32. (a) Let A be decomposed as in (B.29). Prove that

$$r(A) \leq \sqrt{r(A_0)^2 + r(A_1)^2} = \sqrt{\|A_0\|_2^2 + \|A_1\|_2^2}. \quad (\text{B.33})$$

(b) Given $\vartheta \in \mathbb{C}$, decompose ϑA according to (B.29) into $\vartheta A = A_{\vartheta,0} + iA_{\vartheta,1}$. Prove that

$$r(A) = \max \{\rho(A_{\vartheta,0}) : \vartheta \in \mathbb{C}, |\vartheta| = 1\}.$$

Appendix C

Facts from Matrix Theory

Abstract The previous chapters are concerned with linear algebra aspects of matrices and with the properties of normed spaces. Here we introduce two types of partial order relations between matrices. Section C.1 is devoted to positive definite matrices. The graph of a matrix and irreducible matrices are introduced in Section C.2. Positive matrices and the Perron–Frobenius theory are discussed in Section C.3. The M-matrices introduced in Section C.4 are of interest in connection with classical iterative methods. The generalisation to H-matrices is given in Section C.5. Schur complements are defined in Section C.6.

C.1 Positive Definite Matrices

C.1.1 Definition and Notation

Definition C.1. Let $\langle \cdot, \cdot \rangle$ denote the Euclidean scalar product on \mathbb{K}^I . A matrix $A \in \mathbb{K}^{I \times I}$ is called

positive definite, if $A = A^H$ and $\langle Ax, x \rangle > 0$ for all $0 \neq x \in \mathbb{K}^I$, (C.1a)

positive semidefinite, if $A = A^H$ and $\langle Ax, x \rangle \geq 0$ for all $x \in \mathbb{K}^I$, (C.1b)

negative definite, if $-A$ is positive definite, (C.1c)

negative semidefinite, if $-A$ is positive semidefinite. (C.1d)

We write $A > 0$ for positive definite matrices and $A \geq 0$, $A < 0$, $A \leq 0$ in the cases (C.1b–d).

The terms ‘positive (semi-) definite’ and ‘negative (semi-) definite’ define a partial¹ ordering in the set of Hermitian matrices. For arbitrary Hermitian matrices A and B , we define:

$A > B$, if $A - B > 0$, i.e., if $A - B$ is positive definite.

$A \geq B$, $A < B$, and $A \leq B$ are defined analogously. Any inequality like $A > B$ implies tacitly that the involved matrices are Hermitian.

¹ Partial ordering means that there may be matrices A and B such that neither $A \geq B$ nor $A \leq B$.

The definitions (C.1a–d) depend on the choice of the scalar product (see Exercise C.11c). Furthermore, we emphasise that some authors use the term ‘symmetric positive definite’ (SPD) or ‘Hermitian positive definite’ (HPD) instead of ‘positive definite’, while ‘ A positive definite’ is used in another sense, denoting not necessarily Hermitian matrices satisfying the coercivity condition

$$\Re \langle Ax, x \rangle > 0 \quad \text{for all } 0 \neq x \in \mathbb{K}^I. \quad (\text{C.2})$$

C.1.2 Rules and Criteria for Positive Definite Matrices

Lemma C.2. *The following rules are valid:*

$$\begin{aligned}
A > 0 &\iff CAC^H > 0 && \text{for regular } C \in \mathbb{C}^{I \times I}, \\
A > B &\iff CAC^H > CBC^H && \text{for regular } C \in \mathbb{C}^{I \times I}, \\
A \geq 0 &\implies CAC^H \geq 0 && \text{for all } C \in \mathbb{C}^{I \times I}, \\
A \geq B &\implies CAC^H \geq CBC^H && \text{for all } C \in \mathbb{C}^{I \times I}, \\
A, B \geq 0 &\implies A + B \geq 0 && \text{for all } A, B \in \mathbb{C}^{I \times I}, \\
A, B \geq 0 &\implies A + B > 0 && \text{if } A > 0 \text{ or } B > 0, \\
A > 0 &\iff \xi A > 0 && \text{for all } \xi > 0, A \in \mathbb{C}^{I \times I}, \\
\zeta I \leq A \leq \xi I &\iff \sigma(A) \subset [\zeta, \xi] && \text{for Hermitian } A \in \mathbb{C}^{I \times I}, \\
-\xi I \leq A \leq \xi I &\iff \|A\|_2 \leq \xi && \text{for Hermitian } A \in \mathbb{C}^{I \times I}, \\
A \geq B > 0 &\iff 0 < A^{-1} \leq B^{-1} && \text{for all } A, B \in \mathbb{C}^{I \times I},
\end{aligned} \tag{C.3a-f}$$

Proof. (i) $x \neq 0$ implies $y := C^H x \neq 0$. Hence, the inequality $0 < \langle Ay, y \rangle = \langle AC^H x, C^H x \rangle = \langle CAC^H x, x \rangle$ shows that $CAC^H > 0$. A second application to C^{-1} instead of C yields the reverse implication.

(ii) The proof of (C.3b) is analogous to (C.3a). (C.3c) and (C.3d) follow immediately from the definition in (C.1a,b).

(iii) A can be diagonalised by a unitary Q : $A = QDQ^H$. (C.3b') with $C = Q^H$ brings (C.3e) into the form $\zeta I \leq D \leq \xi I$, where the diagonal matrix D contains the eigenvalues $\lambda \in \sigma(A)$ as diagonal entries. The equivalence of $\zeta I \leq D \leq \xi I$ and $\sigma(A) \subset [\zeta, \xi]$ is easy to see.

(iv) Choosing $\zeta = -\xi$ in (C.3e) and exploiting the equivalence of $\sigma(A) \subset [-\xi, \xi]$ with $\rho(A) = \|A\|_2 \leq \xi$, we obtain (C.3f).

(v) The proof of (C.3g) is postponed (after Remark C.6). \square

Lemma C.3. *$A > 0$ (respectively, $A \geq 0$) holds if and only if A is Hermitian and all eigenvalues are positive (nonnegative).*

Proof. The demonstration of this assertion is elementary for a diagonal matrix D . Let $A = QDQ^H$ be the diagonalisation of A (cf. (A.20)). By Lemma C.2, $A > 0$ is equivalent to $D > 0$. Since both matrices have the same eigenvalues, the assertion is proved. \square

C.1.3 Remarks Concerning Positive Definite Matrices

Lemma C.4. (a) Any positive definite matrix is regular.

(b) A is positive definite if and only if A^{-1} is positive definite:

$$A > 0 \iff A^{-1} > 0.$$

(c) Each principal submatrix $(a_{\alpha\beta})_{\alpha,\beta \in J}$ ($J \subset I$) of a positive (semi-)definite matrix is again positive (semi-)definite:

$$\begin{aligned} A > 0 &\implies (a_{\alpha\beta})_{\alpha,\beta \in J} > 0, \\ A \geq 0 &\implies (a_{\alpha\beta})_{\alpha,\beta \in J} \geq 0 \quad \text{for } J \subset I. \end{aligned} \tag{C.4}$$

(d) All diagonal elements of a positive (semi-)definite matrix are positive (non-negative):

$$A > 0 \Rightarrow a_{\alpha\alpha} > 0 \quad \text{and} \quad A \geq 0 \Rightarrow a_{\alpha\alpha} \geq 0 \quad \text{for all } \alpha \in I.$$

(e) Let A be positive (semi-)definite. The diagonal part $D = \text{diag}\{A\}$ and each block-diagonal part $D = \text{blockdiag}\{A\}$ of A are again positive (semi-)definite.

Proof. The parts (a,b) are a consequence of Lemma C.3 because A^{-1} has the inverse eigenvalues of A (cf. Remark A.15b).

Part (c) follows from definition (C.1a) if one restricts x to the subspace with $x_\alpha = 0$ for $\alpha \notin J$.

Part (d) is a special case of (c) for $J := \{\alpha\}$, and (e) follows from (d) and (c). \square

Lemma C.5. (a) $0 \leq A \leq B$ implies $\|A\|_2 \leq \|B\|_2$ and $\rho(A) \leq \rho(B)$.

(b) $0 \leq A < B$ implies $\|A\|_2 < \|B\|_2$ and $\rho(A) < \rho(B)$.

Proof. Because $\rho(A) = \|A\|_2$ and $\rho(B) = \|B\|_2$, it is sufficient to show that $\rho(A) \leq \rho(B)$. $A \geq 0$ has an eigenvalue $\lambda = \rho(A)$ and a corresponding eigenvector x with $\|x\| = 1$. $\rho(A) = \langle Ax, x \rangle \leq \langle Bx, x \rangle \leq r(B) = \rho(B)$ (cf. (B.28c)) proves part (a). Part (b) follows analogously. \square

Assume that $A > 0$. Applying Remark A.42 and Lemma C.3 to the nonnegative square root $f(x) = \sqrt{x}$ (well-defined in $[0, \infty)$) yields the matrix $A^{1/2} := f(A)$. More generally, A^α is well-defined for $\alpha > 0$.

Remark C.6. (a) If A is positive definite, then $A^{1/2}$ represents again a positive definite matrix. For its inverse, we use the notation $A^{-1/2}$. It is also equal to $A^{-1/2} = (A^{-1})^{1/2}$. For a positive semidefinite A , the matrix $A^{1/2}$ is well-defined as a positive semidefinite matrix.

(b) $A^{1/2}$ commutes with A and any polynomial (function) of A .

(c) $A^{1/2}$ is the unique positive semidefinite solution of the matrix equation $X^2 = A > 0$.

Proof of (C.3g) in Lemma C.2. (C.3') for $C = B^{-\frac{1}{2}}$ yields $X := B^{-\frac{1}{2}}AB^{-\frac{1}{2}} \geq I$. By (C.3e), all eigenvalues of X are larger or equal to 1. Therefore, the eigenvalues of X^{-1} are ≤ 1 . Using (C.3e), we conclude that $X^{-1} \leq I$, hence $B^{\frac{1}{2}}A^{-1}B^{\frac{1}{2}} \leq I$. A further application of (C.3b') shows that $A^{-1} \leq B^{-1/2}IB^{-1/2} = B^{-1}$. \square

Property (C.3c) implies that the positive (semi-)definite matrices form a semi-group with respect to matrix addition. This does not hold for multiplication. In general, AB is no longer positive (semi-)definite since it is not necessarily Hermitian.

However, we still have the following statement.

Remark C.7. If A and B are positive (semi-)definite, the product AB is real diagonalisable and has only positive (nonnegative) eigenvalues.

Proof. (i) The proof is simple if one of the factors, say A , is regular. Then we use the similarity transformation $AB \mapsto A^{-1/2}ABA^{1/2} = A^{1/2}BA^{1/2}$. Note that the positive (semi-)definite matrix $A^{1/2}BA^{1/2}$ has positive (nonnegative) eigenvalues. The general proof uses the following steps.

(ii) If $A \geq 0$, set $A_\varepsilon := A + \varepsilon I$. Since $A_\varepsilon > 0$, we can apply part (i). The limit $\varepsilon \rightarrow 0$ proves the desired result. \square

Let A be a positive definite matrix. As explained in Exercise B.2c,

$$\|x\|_A := \|A^{1/2}x\|_2 = \sqrt{\langle Ax, x \rangle} \quad \text{for } x \in \mathbb{C}^I \quad (\text{C.5a})$$

describes again a norm, often called the *energy norm* (with respect to $A > 0$).

Exercise C.8. If $A > 0$ has some decomposition $A = C^H C$ (e.g., Cholesky decomposition), then $\|x\|_A = \|Cx\|_2$ holds.

The notation in (C.5a) and (B.10a) is related via $\|\cdot\|_A = \|\cdot\|_{A^{1/2}}$. Using the definition in (C.5a) and Exercise B.13c, we prove the following statement.

Remark C.9. Let A be positive definite. The norm $\|\cdot\|_A$ in (C.5a) is generated by the (*energy*) scalar product

$$\langle x, y \rangle_A := \langle Ax, y \rangle. \quad (\text{C.5b})$$

An equivalent representation of $\|\cdot\|_A$ is

$$\|x\|_A := \sqrt{\langle Ax, x \rangle} \quad \text{for } x \in \mathbb{K}^I. \quad (\text{C.5c})$$

The corresponding matrix norm $\|\cdot\|_A$ is related to the spectral norm by

$$\|B\|_A = \|A^{1/2}BA^{-1/2}\|_2 \quad \text{for all } B \in \mathbb{K}^{I \times I}. \quad (\text{C.5d})$$

There is a one-to-one correspondence between positive definite matrices and scalar products. In (C.5b), each matrix $A > 0$ is associated with a scalar product. The reverse direction is described in the next remark.

Remark C.10. Let $\langle\langle \cdot, \cdot \rangle\rangle$ be an arbitrary scalar product in \mathbb{K}^I and denote the Euclidean scalar product by $\langle \cdot, \cdot \rangle$. Then there is a positive definite matrix A with

$$\langle\langle x, y \rangle\rangle = \langle Ax, y \rangle = \left\langle A^{1/2}x, A^{1/2}y \right\rangle \quad \text{for } x, y \in \mathbb{K}^I.$$

Proof. Using the unit vectors \mathbf{e}_α , define A by $a_{\alpha\beta} := \langle\langle \mathbf{e}_\alpha, \mathbf{e}_\beta \rangle\rangle$. □

- Exercise C.11.** Prove: (a) $CC^H \geq 0$ and $C^H C \geq 0$ are always valid.
 (b) If C is regular, $CC^H > 0$ and $C^H C > 0$ are even positive definite.
 (c) The adjoint C^* of a matrix C with respect to the scalar product $\langle\langle \cdot, \cdot \rangle\rangle$ is defined by

$$\langle\langle Cx, y \rangle\rangle = \langle\langle x, C^*y \rangle\rangle \quad \text{for all } x, y \in \mathbb{K}^I.$$

The identity $C^* = A^{-1}C^H A$ holds. C is selfadjoint with respect to $\langle\langle \cdot, \cdot \rangle\rangle$ if $C = C^*$. The latter condition is equivalent to

$$C^H A = A C.$$

C is positive definite with respect to $\langle\langle \cdot, \cdot \rangle\rangle$ if $C = C^*$ and $\langle\langle Cx, x \rangle\rangle > 0$ for all $x \neq 0$.

C.2 Graph of a Matrix and Irreducible Matrices

A graph is a pair (V, E) of V (the set of *vertices*) and $E \subset V \times V$ (the set of *edges*). In the following, $V = I$ is the index set associated with the matrix $A \in \mathbb{K}^{I \times I}$. Since, given A , the index set I is fixed, only the specification of E is of interest.

In this general definition, (V, E) is a directed graph. An undirected graph can be modelled by the condition that $(\alpha, \beta) \in E$ holds if and only if $(\beta, \alpha) \in E$.

Definition C.12 (matrix graph). Let $A \in \mathbb{K}^{I \times I}$ be a matrix corresponding to the index set I . The following subset of all pairs from $I \times I$ is denoted as graph $G(A)$ of the matrix A :

$$G(A) = \{(\alpha, \beta) \in I \times I : a_{\alpha\beta} \neq 0\}.$$

The set $G(A)$ can be visualised as follows. The indices $\alpha \in I$ are called vertices, while $(\alpha, \beta) \in G(A)$ is called a (directed) edge from vertex α to vertex β and is graphically represented by an arrow pointing from α to β (cf. Fig. C.1a).

In the case of $a_{\alpha\beta} \neq 0$ and $a_{\beta\alpha} \neq 0$, the vertices α and β are connected in both directions (e.g., $\alpha = 3$, $\beta = 4$ in Fig. C.1a). The matrix A has a symmetric structure if $a_{\alpha\beta} \neq 0$ holds if and only if $a_{\beta\alpha} \neq 0$. In this case, all edges are bidirected and one can omit the specification of the directions (cf. Fig. C.1b). The graph $G(A)$ is minimal for the zero matrix, $G(0) = \emptyset$, and maximal for fully populated matrices: $G(A) = I \times I$.

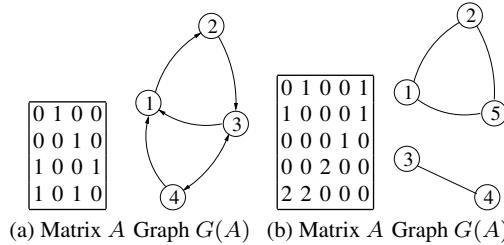


Fig. C.1 (a) Directed graph of a matrix. (b) Symmetric structure.

If there is an edge $(\alpha, \beta) \in G(A)$, we say that ‘ α is directly connected to β ’. The statement ‘ α is connected to β ’ means that there is at least one path consisting of direct connections:

$$\begin{aligned} \alpha = \alpha_0, \alpha_1, \alpha_2, \dots, \alpha_k = \beta \quad \text{with} \\ k \in \mathbb{N} \quad \text{and} \quad (\alpha_{i-1}, \alpha_i) \in G(A) \quad \text{for all } i = 1, \dots, k. \end{aligned} \tag{C.6}$$

The number k is called the *length* of the path (C.6).

Exercise C.13. Prove: (a) The relation ‘ α is connected with β ’ is transitive, but not necessarily symmetric.

(b) Let $n := \#I$. If α is connected with β , the path (C.6) can be chosen so that its length is $k \leq n - 1$.

Definition C.14 (irreducible matrix). A matrix $A \in \mathbb{K}^{I \times I}$ is called *irreducible* if any $\alpha \in I$ is connected to any $\beta \in I$ or if² $\#I = 1$. Otherwise, A is called *reducible*.

Theorem C.15. $A \in \mathbb{K}^{I \times I}$ is reducible if and only if there is an ordering of the indices such that A takes the block form

$$A = \underbrace{\begin{bmatrix} A^{11} & A^{12} \\ 0 & A^{22} \end{bmatrix}}_{\substack{I_1 \\ I_2}} \} I_1 \quad (A^{11}, A^{22} : \text{square blocks}) \tag{C.7}$$

with nonempty disjoint index subset $I_1, I_2 \subset I$ ($I_1 \cup I_2 = I$).

Proof. (i) Let A be as in (C.7). Choose any $\alpha \in I_2$, $\beta \in I_1$. Assume that a path (C.6) exists connecting α with β . Then there must be an edge $(\alpha_{\ell-1}, \alpha_\ell) \in G(A)$ with $\alpha_{\ell-1} \in I_2$ and $\alpha_\ell \in I_1$. The contradiction results from $a_{\alpha_{\ell-1}, \alpha_\ell} = 0$ because this entry belongs to the block $A^{21} = 0$. Hence, α is not connected to β and therefore A is reducible.

² The second condition is added to ensure that all 1×1 matrices $A = (a_{11})$ are irreducible, even if $a_{11} = 0$ and $G(A) = \emptyset$.

(ii) Let A be reducible. Then there must be indices $\alpha, \beta \in I$, such that α is not connected to β . Choose

$$I_1 := \{\gamma \in I : \gamma \text{ connected to } \beta\} \cup \{\beta\}, \quad I_2 := I \setminus I_1.$$

The sets are nonempty, since $\beta \in I_1$ and $\alpha \in I_2$. Enumerate first I_1 then I_2 . An entry $a_{\delta\gamma}$ from the block A^{21} has the indices $\delta \in I_2$, $\gamma \in I_1$. $a_{\delta\gamma} = 0$ must hold, since otherwise δ is connected to γ , which by definition of I_1 is connected to β ; hence, $\delta \in I_1$ would follow in contradiction to $\delta \in I_2$. \square

Consider the block matrix $\begin{bmatrix} A^{11} & A^{12} \\ 0 & A^{22} \end{bmatrix}$ in (C.7). The decomposition can be continued recursively. If, e.g., A^{22} is reducible, we apply Theorem C.15 to A^{22} and decompose into $A^{22} = \begin{bmatrix} A^{22,11} & A^{22,12} \\ 0 & A^{22,22} \end{bmatrix}$ and so on. The recursion stops as soon as the principal submatrices are irreducible. The resulting matrix has a block structure $\{I^1, \dots, I^k\}$ as indicated in

$$A = \begin{bmatrix} A^{11} & A^{12} & \cdots & A^{1k} \\ 0 & A^{22} & \cdots & A^{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & A^{kk} \end{bmatrix} \{I_1 \atop I_2 \atop \vdots \atop I_k}, \quad A^{ii} \text{ irreducible } (1 \leq i \leq k). \quad (\text{C.8})$$

Exercise C.16. Prove: (a) The matrix A in Figure C.1a is irreducible, whereas the matrix in Figure C.1b is reducible.

(b) If A has a symmetric structure (i.e., $G(A) = G(A^H)$) and $\#I > 1$, A is irreducible if and only if the graph $G(A)$ is connected.

(c) Triangular and diagonal matrices with $\#I > 1$ are reducible.

(d) The matrix of the Poisson model problem is irreducible.

If the matrix of a linear system of equations is reducible, the system allows a reduction. If A is reducible and $\{I_1, I_2\}$ describes the block structure in (C.7), $Ax = b$ can be solved in two stages using smaller systems of equations:

$$A^{22}x^2 = b^2, \quad A^{11}x^1 = b^1 - A^{12}x^2.$$

Remark C.17. Let $n := \#I > 1$. Define $G_k(A) := \{(\alpha, \beta) \in I \times I : \text{there is a path (C.6) of length } \leq k \text{ connecting } \alpha \text{ with } \beta\}$. Then the following statements are valid:

(a) $G(A) = G_1(A) \subset G_2(A) \subset \dots \subset G_{k-1}(A) \subset G_k(A) \subset \dots$

(b) $G_k(A) = G_{n-1}(A)$ for all $k \geq n - 1$.

(c) A is irreducible if and only if $G_{n-1}(A) = I \times I$.

Proof. (a) A chain (C.6) of length $k = 1$ is a direct connection, i.e., an edge from $G(A)$. Vice versa, any edge from $G(A)$ belongs to $G_1(A)$; hence, $G(A) = G_1(A)$.

(b) Statement (b) follows according to Exercise C.13b.

(c) If α is connected with β , then according to (a) and (b), the edge (α, β) must belong to $G_{n-1}(A)$. This proves the assertion (c). \square

C.3 Positive Matrices

C.3.1 Definition and Notation

The positive definiteness $A > 0$ establishes a partial order relation in the set of the Hermitian matrices. Another partial order relation in the algebra of all real matrices is defined by *elementwise inequalities*:

$$A > B \iff a_{\alpha\beta} > b_{\alpha\beta} \quad \text{for all } \alpha, \beta \in I, \quad (\text{C.9a})$$

$$A \geq B \iff a_{\alpha\beta} \geq b_{\alpha\beta} \quad \text{for all } \alpha, \beta \in I, \quad (\text{C.9b})$$

In this section, we use the signs ‘ $>$ ’ and ‘ \geq ’ only in the componentwise sense of (C.9a,b).

Remark C.18. The counterexample $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, $B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ shows that $A \geq B$ and $A \neq B$ do not imply $A > B$. Therefore, we introduce the notation

$$A \not\geq B \iff A \geq B \text{ and } A \neq B.$$

The positive matrices are characterised by

$$A > 0$$

(i.e., $a_{\alpha\beta} > 0$ for all $\alpha, \beta \in I$), the nonnegative matrices are defined by

$$A \geq 0.$$

The next remark states that the nonnegative matrices form a semigroup concerning addition and multiplication.

Remark C.19. $A, B \geq 0$ yields $A + B \geq 0$ and $AB \geq 0$. $AB > 0$ follows from $A, B > 0$, while $A + B > 0$ holds if one of the matrices $A, B \geq 0$ is positive.

Analogous order relations can be defined for vectors:

$$x \geq y \iff x_\alpha \geq y_\alpha \quad \text{for all } \alpha \in I.$$

Similarly, $x > y$, $x \leq y$, $x < y$, $x \not\geq y$, etc. are defined.

Exercise C.20. Prove: (a) $A \geq B \geq 0$ and $x \geq y \geq 0$ implies $Ax \geq By$. In particular, $Ax \geq 0$ holds for $A \geq 0$ and $x \geq 0$.

(b) A is positive if and only if $Ax > 0$ for all $x \not\geq 0$.

The *absolute value* of a matrix (or a vector) is again a matrix (vector) and is defined componentwise (do not confuse with a norm!):

$$|A| := (|a_{\alpha\beta}|)_{\alpha, \beta \in I} \in \mathbb{R}^{I \times I}, \quad |x| := (|x_\alpha|)_{\alpha \in I} \in \mathbb{R}^I.$$

The established order relations match particularly well with the maximum norm and the row-sum norm as corresponding matrix norm.

Exercise C.21. Prove that

$$\begin{aligned}\|x\|_\infty &= \| |x| \|_\infty, \quad x \geq y \geq 0 \Rightarrow \|x\|_\infty \geq \|y\|_\infty, \\ \|A\|_\infty &= \| |A| \|_\infty, \quad A \geq B \geq 0 \Rightarrow \|A\|_\infty \geq \|B\|_\infty.\end{aligned}$$

Some properties of positive matrices correspond to those of positive definite matrices. However, concerning (C.3g) we obtain the opposite statement.

Exercise C.22. Let $\#I > 1$. A and A^{-1} cannot simultaneously be positive.

The following results combine the positivity or nonnegativity property with the matrix graph (cf. §C.2).

Exercise C.23. Prove that $G(\alpha A + \beta B) = G(A) \cup G(B)$ for nonnegative matrices $A, B \geq 0$ and positive numbers α, β .

Remark C.24. (a) $G_k(A) \subset G([I + A]^k)$ holds for $A \geq 0$ and $G_k(A)$ defined in Remark C.17.

(b) If $A \geq 0$ is irreducible, then $(I + A)^{n-1} > 0$ and $\sum_{\nu=0}^{n-1} A^\nu > 0$ for $n := \#I$.

Proof. (i) Since the matrices in (a) and (b) are nonnegative, the condition $a_{\alpha\beta} \neq 0$ in Definition C.12 can be replaced with $a_{\alpha\beta} > 0$. Define the matrix $A' := I + A$ and let $(\alpha_0, \alpha_k) \in G_k(A)$, i.e., there is a path of direct connections $(\alpha_{\ell-1}, \alpha_\ell) \in G(A)$ for $1 \leq \ell \leq k$. Since $A'_{\alpha\beta} \geq 0$, the coefficient

$$(A'^k)_{\alpha_0, \alpha_k} = \sum_{\beta_1, \dots, \beta_{k-1} \in I} a'_{\alpha_0, \beta_1} a'_{\beta_1, \beta_2} \cdot \dots \cdot a'_{\beta_{k-1}, \alpha_k}$$

of the matrix A'^k can be bounded from below by

$$(A'^k)_{\alpha_0, \alpha_k} \geq a'_{\alpha_0, \alpha_1} a'_{\alpha_1, \alpha_2} \cdot \dots \cdot a'_{\alpha_{k-1}, \alpha_k}. \quad (\text{C.10})$$

For $\alpha_{\ell-1} \neq \alpha_\ell$, we conclude from $(\alpha_{\ell-1}, \alpha_\ell) \in G(A)$ that $a'_{\alpha_{\ell-1}, \alpha_\ell} > 0$, whereas for $\alpha_{\ell-1} = \alpha_\ell$ the diagonal entry

$$a'_{\alpha_{\ell-1}, \alpha_\ell} = 1 + a_{\alpha_{\ell-1}, \alpha_\ell} \geq 1 > 0$$

is also positive. Hence, all factors $a'_{\alpha_{\ell-1}, \alpha_\ell}$ appearing in (C.10) are positive. This proves $(A'^k)_{\alpha_0, \alpha_k} > 0$, $(\alpha_0, \alpha_k) \in G(A'_k)$, and finally the statement $G_k(A) \subset G(A'^k)$ of part (a).

(ii) Obviously, $B := (I + A)^{n-1} \geq 0$ holds (cf. Remark C.19). An irreducible matrix A satisfies

$$I \times I \underset{\text{Remark C.17c}}{=} G_{n-1}(A) \underset{\text{Remark C.17d}}{\subset} G((I + A)^{n-1}) = G(B).$$

Hence, $(\alpha, \beta) \in G(B)$ is always true, i.e., $B_{\alpha\beta} > 0$ holds. This proves $B > 0$. The case of $\sum A^\nu$ is analogous. \square

C.3.2 Perron–Frobenius Theory of Positive Matrices

The main result of this section is the theorem of Oskar Perron [311] and Ferdinand Georg Frobenius [141, 142].

Theorem C.25 (Perron–Frobenius). *Let $A \geq 0$ be an irreducible matrix in $\mathbb{R}^{I \times I}$, where $n := \#I > 1$. Then the following statements hold:*

$$\rho(A) > 0 \quad \text{is a simple eigenvalue of } A, \quad (\text{C.11a})$$

$$\lambda = \rho(A) \quad \text{is associated with a positive eigenvector } x > 0, \quad (\text{C.11b})$$

$$\rho(B) > \rho(A) \quad \text{for all } B \not\geq A. \quad (\text{C.11c})$$

The proof of the theorem is prepared by Lemmata C.26–C.30. We start with some auxiliary constructions. The set

$$E := \{x \in \mathbb{R}^I : \|x\|_\infty = 1, x \geq 0\}$$

consists of vectors with $0 \leq x_\beta \leq 1$ and at least one component $x_\beta = 1$.

Lemma C.26. *Assume that $A \geq 0$. The set*

$$K := \{(x, \rho) \in E \times \mathbb{R} : \rho > 0, Ax \geq \rho x\}$$

is compact. The maximum

$$r := \max\{\rho : (x, \rho) \in K \text{ for some } x \in E\} \quad (\text{C.12a})$$

exists. For any pair $(y, r) \in K$, we have

$$Ay \geq ry \quad \text{and not } Ay > ry. \quad (\text{C.12b})$$

Proof. (i) Let (x, ρ) be the limit of the sequence $(x_\nu, \rho_\nu) \in K$. We conclude from $Ax_\nu \geq \rho_\nu x_\nu$ that $Ax \geq \rho x$. Therefore, $(x, \rho) \in K$ proves that K is closed.

(ii) The boundedness of x is trivial because of $\|x\|_\infty = 1$. The component ρ of $(x, \rho) \in K$ is bounded by $0 \leq \rho \leq \|A\|_\infty$, because the index $\alpha \in I$ with $x_\alpha = 1$ satisfies $\rho = \rho x_\alpha \leq (Ax)_\alpha \leq \|Ax\|_\infty < \|A\|_\infty \|x\|_\infty \leq \|A\|_\infty$. This completes the proof that K is compact.

(iii) Let r be the supremum of $\{\rho : (x, \rho) \in K \text{ for some } x \in E\}$. There is a sequence $(x_\nu, \rho_\nu) \in K$ with $\rho_\nu \rightarrow r$. Since K is compact, a subsequence converges to $(y, r) \in K$. By definition of K , the inequality $Ay \geq ry$ must hold. If $Ay > ry$, r could be increased in contradiction to the maximality of r . \square

Lemma C.27. *Assume that $A \geq 0$ is irreducible with $n := \#I > 1$. Let r be defined according to (C.12a) and assume that $y \in E$ satisfies (C.12b). Then*

$$r > 0, \quad y > 0, \quad Ay = ry;$$

i.e., y is a positive eigenvector of A corresponding to the positive eigenvalue r .

Proof. (i) The residual vector $z := Ay - ry$ is nonnegative because of (C.12b). Under the assumption $z \neq 0$, Remark C.24b yields $(I+A)^{n-1}z > 0$ and therefore,

$$\begin{aligned} 0 < (I+A)^{n-1}z &= (I+A)^{n-1}(Ay - ry) = (I+A)^{n-1}(A - rI)y \\ &= (A - rI)(I+A)^{n-1}y = Ay' - ry' \quad \text{for } y' := (I+A)^{n-1}y. \end{aligned}$$

From $y > 0$, we conclude again that $y' = (I+A)^{n-1}y > 0$. The normalised vector $y'' := y'/\|y'\|_\infty$ belongs to E . $Ay' > ry'$ implies that $(y'', r) \in K$ and $Ay'' > ry''$, which contradicts (C.12b). Hence, the assumption $z \neq 0$ is not valid. Thus, $z = 0$ proves $Ay = ry$.

(ii) In part (i) we already used $(I+A)^{n-1}y > 0$. Therefore, the eigenvalue equation $Ay = ry$ yields $(1+r)^{n-1}y > 0$. By $1+r \geq 1 > 0$, $y > 0$ follows.

(iii) If $r = 0$, $Ay = ry = 0$ would follow. From $Ay = 0$ and $y > 0$, we conclude that $A = 0$. Because $n > 1$, A would be reducible. This contradiction proves that $r > 0$. \square

Lemma C.28. Assume that A is irreducible and $|B| \leq A$. Then we have

$$\rho(B) \leq r \quad (r \text{ according to (C.12a)}), \quad (\text{C.13a})$$

$$\rho(B) = r \iff (|B| = A, B = \omega DAD^{-1}, |D| = I, |\omega| = 1). \quad (\text{C.13b})$$

Proof. (i) Let y be the normalised eigenvector corresponding to $\beta \in \sigma(B)$, i.e., $By = \beta y$, $\|y\|_\infty = 1$. By

$$|\beta| |y| = |\beta y| = |By| \leq |B| |y| \leq A |y|, \quad (\text{C.13c})$$

$(|y|, |\beta|)$ belongs to K and proves that $|\beta| \leq r$. Since $\beta \in \sigma(B)$ is arbitrary, (C.13a) is shown: $\rho(B) \leq r$.

(ii) Let $|\beta| = r$. The vector y in part (i) satisfies $(|y|, r) \in K$. By Lemma C.27, $|y| > 0$ is an eigenvector of $A : A|y| = r|y|$. The inequality

$$r|y| = |\beta||y| \leq |B||y| \leq A|y| = r|y|$$

(cf. (C.13c)) implies that $|B||y| = A|y|$. Since $|y| > 0$ and $|B| \leq A$, the equality $|B| = A$ follows. The definition $D := \text{diag}\{y_\alpha/|y_\alpha| : \alpha \in I\}$ makes sense because of $|y| > 0$ and leads to $D|y| = y$. Define $\omega := \beta/r$ with $r > 0$ in Lemma C.27. The conditions $|D| = I$ and $|\omega| = 1$ are satisfied. The eigenvalue equation $By = \beta y$ becomes

$$\frac{1}{\omega} D^{-1} BD |y| = r |y|.$$

The matrix $C := \frac{1}{\omega} D^{-1} BD$ fulfills $|C| = |B| = A$ and $C|y| = r|y| = A|y| = |C||y|$. $|y| > 0$ implies that $C = |C| = A$. This proves the direction ' \Rightarrow ' in (C.13b).

(iii) Now let the right-hand part of (C.13b) be valid. Then B has an eigenvalue $\beta = wr$ proving $|\beta| = r$ and, by part (i), also $\rho(B) = r$. \square

Lemma C.29. $r = \rho(A)$ holds for any irreducible matrix $A > 0$.

Proof. The right-hand part of (C.13b) is satisfied for $B := A$ with $D = I$ and $\omega = 1$. Hence, $r = \rho(B) = \rho(A)$ holds. \square

Lemma C.30. Let $A \geq 0$ be irreducible and B a proper principal submatrix of A , i.e., $B = (a_{\alpha\beta})_{\alpha,\beta \in I'}$ for a nonempty index subset $I' \subsetneq I$. Then $\rho(B) < \rho(A)$ holds.

Proof. The matrix $B' := (b'_{\alpha\beta})_{\alpha,\beta \in I}$ with $b'_{\alpha\beta} = \begin{cases} a_{\alpha\beta} = b_{\alpha\beta} & \text{for } \alpha, \beta \in I' \\ 0 & \text{otherwise} \end{cases}$ is the block-diagonal matrix $\text{blockdiag}(B, 0)$ with respect to the block structure $\{I', I \setminus I'\}$. The identity $\sigma(B') = \sigma(B) \cup \{0\}$ proves that $\rho(B') = \rho(B)$. Obviously, $|B'| = B' \leq A$ is valid. Since B' is reducible, the right-hand side in (C.13b) cannot be satisfied for B' and A ; hence, $\rho(B) = \rho(B') < \rho(A)$ follows. \square

Proof of Theorem C.25. (i) Lemma C.29 shows that $r = \rho(A)$, whereas Lemma C.27 proves that $r = \rho(A) > 0$ is an eigenvalue with a positive eigenvector.

(ii) If $B \not\supseteq A$, the irreducibility of A carries over to B because of $G(B) \supset G(A)$. Since $A = |A| \not\leq B$, we deduce from Lemma C.28 with interchanged roles of A and B that $\rho(A) < r_B$, where $r_B = \rho(B)$ is the value r in (C.12a) belonging to B .

(iii) It remains to show (C.11a): $\lambda = \rho(A)$ is a simple eigenvalue. Let A_γ for $\gamma \in I$ be the principal submatrices associated with the index set $I_\gamma := I \setminus \{\gamma\}$. The derivative of the determinant of $\lambda I - A$ is equal to

$$\frac{d}{d\lambda} \det(\lambda I - A) = \sum_{\gamma \in I} \det(\lambda I - A_\gamma). \quad (\text{C.14})$$

Since $\rho(A_\gamma) < \rho(A)$ by Lemma C.30, we have $\det(\lambda I - A_\gamma) \neq 0$ for all $\lambda > \rho(A)$. The polynomial $\det(\lambda I - A_\gamma) = \lambda^{n-1} + \dots$ tends to $+\infty$ as $\lambda \rightarrow \infty$. Hence, it must be positive in the interval $[\rho(A), \infty)$. From $\det(\lambda I - A_\gamma) > 0$ and (C.14) we conclude that

$$\det(\lambda I - A) > 0 \quad \text{for } \lambda \geq \rho(A).$$

A double zero of $\det(\lambda I - A)$ at $\lambda = \rho(A)$ would lead to a vanishing derivative; hence, $\lambda = \rho(A)$ is only a simple root and thereby also a simple eigenvalue. \square

Exercise C.31. Prove that the eigenvalue $\lambda = \rho(A)$ of an irreducible matrix $A \geq 0$ is the only one with the property $|\lambda| = \rho(A)$. Hint: Prove that the eigenvector x belonging to λ with $|\lambda| = \rho(A)$ yields a vector $y := |x|$ satisfying (C.12b). Apply Lemma C.27.

Exercise C.32. Prove that if $x > 0$ is an eigenvector of an irreducible matrix $A \geq 0$, then it belongs to the eigenvalue $\lambda = \rho(A)$.

Theorem C.25 requires irreducibility of A , which, in particular, is ensured for positive $A > 0$. However, for reducible matrices $A \geq 0$ not all statements of the theorem remain valid.

Exercise C.33. Prove that there are reducible matrices $A \geq 0$, such that $\rho(A)$ is a multiple eigenvalue and the corresponding eigenvectors $x \geq 0$ have components $x_\alpha = 0$.

Finally, we summarise the properties of possibly reducible matrices $A \geq 0$.

Theorem C.34. Let $A \geq 0$. Then the following statements hold:

$$\begin{aligned} 0 < \rho(A) &\text{ is an eigenvalue of } A, \text{ i.e., } \rho(A) \in \sigma(A), \\ \lambda = \rho(A) &\text{ corresponds to a nonnegative eigenvector } x \not\equiv 0, \\ \rho(B) &\geq \rho(A) \text{ for all } B \geq A. \end{aligned} \tag{C.15}$$

Proof. (i) Since the case $n := \#I = 1$ is trivial, assume that $n > 1$. We define $A_\varepsilon := (a_{\alpha\beta} + \varepsilon)_{\alpha,\beta \in I}$ for $\varepsilon > 0$. A_ε is irreducible because of $G(A_\varepsilon) = I \times I$. By Theorem C.25, $\lambda_\varepsilon = \rho(A_\varepsilon)$ is an eigenvalue of A_ε with the eigenvector $x_\varepsilon > 0$, $\|x_\varepsilon\|_\infty = 1$. Since the eigenvalues (as zeros of a polynomial) vary continuously on ε , $\lambda := \lim_{\varepsilon \rightarrow 0} \lambda_\varepsilon = \lim_{\varepsilon \rightarrow 0} \rho(A_\varepsilon) = \rho(A)$ is an eigenvalue of A . The compactness of $\{x : \|x\|_\infty = 1\}$ implies the existence of a convergent subsequence $x_{\varepsilon_\nu} \rightarrow x$ with $\|x\|_\infty = 1$ and $x \geq 0$. $A_{\varepsilon_\nu} x_{\varepsilon_\nu} = \lambda_{\varepsilon_\nu} x_{\varepsilon_\nu}$ yields $Ax = \lambda x$; i.e., $x \not\equiv 0$ is an eigenvector.

(ii) In analogy to A_ε , we define $B_{2\varepsilon}$. From $B_{2\varepsilon} > A_\varepsilon$ and $\rho(B_{2\varepsilon}) > \rho(A_\varepsilon)$, we conclude that $\rho(B) \geq \rho(A)$ for $\varepsilon \rightarrow 0$. \square

Exercise C.35. Prove $\rho(B) \leq \rho(|B|) \leq \rho(A)$ for all $B \in \mathbb{K}^{I \times I}$ with $|B| \leq A \in \mathbb{R}^{I \times I}$. Hint: Perform the limit $A_\varepsilon \rightarrow A$ in Lemmata C.28–C.29.

C.3.3 Diagonal Dominance

Definition C.36. A matrix $A \in \mathbb{K}^{I \times I}$ is *strictly diagonally dominant* if

$$|a_{\alpha\alpha}| > \sum_{\beta \in I \setminus \{\alpha\}} |a_{\alpha\beta}| \quad \text{for all } \alpha \in I, \tag{C.16}$$

weakly diagonally dominant if

$$|a_{\alpha\alpha}| \geq \sum_{\beta \in I \setminus \{\alpha\}} |a_{\alpha\beta}| \quad \text{for all } \alpha \in I,$$

and *irreducibly diagonally dominant* if A is an irreducible and weakly diagonally dominant matrix and if, in addition,

$$|a_{\alpha\alpha}| > \sum_{\beta \in I \setminus \{\alpha\}} |a_{\alpha\beta}| \quad \text{for at least one } \alpha \in I.$$

If A is not irreducible, the following generalisation may help.

Definition C.37. For $A \in \mathbb{K}^{I \times I}$ and $\gamma \in I$, define

$$G_\gamma := \{\beta \in I : \gamma \text{ connected to } \beta \text{ in the matrix graph } G(A)\}.$$

Then we call A *essentially diagonally dominant* if A is weakly diagonally dominant and if condition (C.17) applies for all $\gamma \in I$:

$$|a_{\alpha\alpha}| > \sum_{\beta \in I \setminus \{\alpha\}} |a_{\alpha\beta}| \quad \text{for at least one } \alpha \in G_\gamma. \quad (\text{C.17})$$

Note that condition (C.17) implies $G_\gamma \neq \emptyset$ for all $\gamma \in I$.

Exercise C.38. For $A \in \mathbb{K}^{I \times I}$, let (C.8) be the (I_1, \dots, I_k) -block decomposition into irreducible principal submatrices. Prove that G_γ is the union of some of the index subsets I_j . Using G_γ for a new block structure, we obtain a block-diagonal matrix with diagonal blocks being irreducibly diagonally dominant.

Exercise C.39. Prove: (a) For irreducible matrices, the essential and irreducible diagonal dominance are equivalent.

(b) The following implications hold: strictly diagonally dominant \Rightarrow essentially diagonally dominant \Rightarrow weakly diagonally dominant.

(c) If A is strictly, irreducibly, or essentially diagonally dominant, then the diagonal elements do not vanish: $a_{\alpha\alpha} \neq 0$.

(d) The matrix of the model problem in §1.2 is irreducibly diagonally dominant, but not strictly diagonally dominant. Hint: Exercise C.16d.

In the next section, diagonal dominance is used as a criterion for the M-matrix property. However, it can also be used to prove positive definiteness.

Lemma C.40. Let $A \in \mathbb{K}^{I \times I}$ be Hermitian with a positive diagonal: $a_{\alpha\alpha} > 0$. If A is strictly, essentially, or irreducibly diagonally dominant, then A is also positive definite. A sufficient condition for positive semidefiniteness is that $A = A^H$ be weakly diagonally dominant with $a_{\alpha\alpha} > 0$.

Proof. By Lemma C.3, it is sufficient to prove that all eigenvalues are positive or nonnegative, respectively.

Let $\lambda \in \mathbb{R}$ be any eigenvalue of A and e a corresponding eigenvector. Let $\alpha \in I$ be the index with $|e_\alpha| = \|e\|_\infty$. Without loss of generality, we may assume that $e_\alpha = \|e\|_\infty = 1$. The α -th component of the system $\lambda e = Ae$ is

$$\lambda = \lambda e_\alpha = (Ae)_\alpha = \sum_{\beta \in I} a_{\alpha\beta} e_\beta = a_{\alpha\alpha} + \sum_{\beta \in I \setminus \{\alpha\}} a_{\alpha\beta} e_\beta.$$

$\lambda \geq 0$ follows from

$$\lambda \geq \underbrace{a_{\alpha\alpha}}_{>0} - \sum_{\beta \in I \setminus \{\alpha\}} |a_{\alpha\beta}| \underbrace{|e_\beta|}_{\leq \|e\|_\infty = 1} \geq |a_{\alpha\alpha}| - \sum_{\beta \in I \setminus \{\alpha\}} |a_{\alpha\beta}| \geq 0,$$

where the last step uses the weak diagonal dominance.

Assuming strict diagonal dominance, we obtain $\lambda > 0$. For the other types of diagonal dominance, one has to use the Gershgorin circles (cf. Varga [376]). The complete proof is in Hackbusch [193, Criterion 4.3.24]. \square

C.4 M-Matrices

The term ‘M-matrix’ was introduced by Ostrowski [301] in 1937 with ‘M’ abbreviating ‘Minkowski’ referring to Minkowski’s determinant.

C.4.1 Definition

Definition C.41 (M-Matrix). A matrix $A \in \mathbb{R}^{I \times I}$ is called an M-matrix if

$$a_{\alpha\alpha} > 0 \quad \text{for all } \alpha \in I, \tag{C.18a}$$

$$a_{\alpha\beta} \leq 0 \quad \text{for all } \alpha \neq \beta, \tag{C.18b}$$

$$A \text{ regular and } A^{-1} \geq 0. \tag{C.18c}$$

The sign pattern (C.18a,b) is easy to check, whereas the verification of $A^{-1} \geq 0$ is more difficult. For this purpose, we shall provide other equivalent or sufficient criteria. Matrices with the property (C.18c) are called *inverse positive* or *monotone*. Hence, M-matrices form a subclass of the inverse positive matrices.

The next lemma shows that the conditions (C.18a–c) are redundant.

Lemma C.42. *The conditions (C.18b,c) imply (C.18a).*

Proof. For an indirect proof, assume that $a_{\gamma\gamma} \leq 0$ for some $\gamma \in I$. Let s^γ be the γ -th column of the matrix A , while e^γ denotes the γ -th unit vector. Multiplying e^γ by $I = A^{-1}A$, we obtain $A^{-1}s^\gamma = e^\gamma$. Using $a_{\gamma\gamma} \leq 0$ and condition (C.18b), $s^\gamma \leq 0$ holds. The property $A^{-1} \geq 0$ in (C.18c) implies that $e^\gamma = A^{-1}s^\gamma \leq 0$ in contradiction to $(e^\gamma)_\gamma = 1$. \square

Exercise C.43. Assume (C.18c) and $b \leq b'$ for the right-hand sides of $Ax = b$ and $Ax' = b'$. Prove that $x \leq x'$.

Exercise C.44. Prove that, in general, the product $A = A_1 A_2$ of two M-matrices A_1 and A_2 is no more an M-matrix, although A is always inverse positive. Hint: choose tridiagonal matrices as an example.

C.4.2 M-Matrices and the Jacobi Iteration

We recall that the iterations matrix M^{Jac} of the pointwise Jacobi iteration (cf. (3.7a)) is defined by $M = I - D^{-1}A$, where $D = \text{diag}\{a_{\alpha\alpha} : \alpha \in I\}$ is the diagonal part of A . Instead of (C.18a–c), we consider the following three conditions:

$$a_{\alpha\alpha} > 0 \quad \text{for all } \alpha \in I, \quad (\text{C.19a})$$

$$M := I - D^{-1}A \geq 0, \quad (\text{C.19b})$$

$$\rho(M) < 1. \quad (\text{C.19c})$$

Inequality (C.19a)—which is a repetition of (C.18a)—ensures that D^{-1} in (C.19b) be well defined. Under the assumption (C.19a), the statements (C.18b) and (C.19b) are equivalent since

$$M_{\alpha\beta} = \delta_{\alpha\beta} - a_{\alpha\beta}/a_{\alpha\alpha} \geq 0 \quad \text{for all } \alpha, \beta \in I \quad (\text{C.19b}')$$

(note that $M_{\alpha\beta} = 0$ for $\alpha = \beta$, while (C.18b) applies for $\alpha \neq \beta$). Therefore, the main difference between the conditions (C.18a–c) and (C.19a–c) are the last conditions (C.18c) and (C.19c). Note that inequality (C.19c) characterises the convergence of the Jacobi iteration.

Theorem C.45. *Both conditions, (C.18a–c) and (C.19a–c) are equivalent definitions of an M-matrix.*

Proof. (i) (C.18a–c) \Rightarrow (C.19a–c): As seen above, (C.18a,b) implies (C.19a,b). Statement (C.19a) is equivalent to $D \geq 0$ and $D^{-1} \geq 0$. For $A' := D^{-1}A$, we find the nonnegative inverse $A'^{-1} = A^{-1}D \geq 0$. (C.19b') shows that $M \geq 0$.

By Theorem C.34, $\lambda := \rho(M) \in \sigma(M)$ belongs to an eigenvector $x \not\equiv 0$. $Mx = \lambda x$ leads to $A'^{-1}(1 - \lambda)x = x$. Since $A'^{-1} \geq 0$ is regular and $x \not\equiv 0$ holds, the inequality $1 - \lambda > 0$ must hold, implying $0 \leq \rho(M) = \lambda < 1$ and (C.19c).

(ii) (C.18a–c) \Leftarrow (C.19a–c): We apply Theorem B.29. Since $\rho(M) < 1$, the geometric sum $(I - M)^{-1} = \sum_{\nu=0}^{\infty} M^{\nu}$ converges and is nonnegative because $M \geq 0$. Then $0 \leq (I - M)^{-1}D^{-1} = (D^{-1}A)^{-1}D^{-1} = A^{-1}DD^{-1} = A^{-1}$ proves (C.18c). \square

The following property is the discrete analogue of the maximum principle of second-order elliptic differential equations (cf. Hackbusch [193, Theorem 2.3.3]). Condition (C.20a) replaces $A^{-1} \geq 0$ in (C.18c) by $A^{-1} > 0$, while (C.20b) corresponds to (C.19c) with the additional requirement that M be irreducible.

Corollary C.46. Let $A \in \mathbb{R}^{I \times I}$ satisfy (C.18a,b) (or equivalently (C.19a,b)). Define D and M as in Theorem C.45. Then the following statements (C.20a) and (C.20b) are equivalent:

$$A \text{ regular and } A^{-1} > 0, \quad (\text{C.20a})$$

$$\rho(M) < 1, \quad A \text{ or } M \text{ irreducible.} \quad (\text{C.20b})$$

Proof. (i) The graphs $G(A)$ and $G(M)$ coincide up to the (uninteresting) diagonal edges (α, α) . Therefore, M is irreducible if and only if A is irreducible.

(ii) $(C.20a) \Rightarrow (C.20b)$: If the matrix A is reducible, there is a block structure $\{I_1, I_2\}$ with $A^{21} = 0$ (cf. (C.7)). Then the inverse $C := A^{-1}$ has the blocks $C = \begin{bmatrix} (A^{11})^{-1} & -(A^{11})^{-1}A^{12}(A^{22})^{-1} \\ 0 & (A^{22})^{-1} \end{bmatrix}$ with $C^{21} = 0$ in contradiction to $A^{-1} > 0$. Hence, A and M are irreducible.

(iii) $(C.20a) \Leftarrow (C.20b)$. Following part (ii) of the previous proof, we use the representation $A^{-1} = \sum_{\nu=0}^{\infty} M^\nu D^{-1}$. This proves $A^{-1} > 0$, since, by Remark C.17e, $\sum_{\nu=0}^{\infty} M^\nu$ is positive. \square

Remark C.47. Irreducible M-matrices have a positive inverse.

Proof. By Theorem C.45, M-matrices satisfy (C.19a–c). Adding the irreducibility, (C.19c) becomes (C.20b). Since Corollary C.46 states the equivalence to (C.20a), $A^{-1} > 0$ is proved. \square

C.4.3 M-Matrices and Diagonal Dominance

The following theorem shows that the diagonal dominance of a matrix, together with the sign conditions (C.18a,b), is sufficient for the M-matrix property. Usually, the theorem is proved by using Gershgorin circles (cf. Hackbusch [193, Criterion 4.3.4] and Varga [376]). Here, however, we use the results of the Perron–Frobenius theory described in §C.3.2.

Theorem C.48. (a) Let the matrix $A \in \mathbb{K}^{I \times I}$ be strictly or essentially or irreducibly diagonally dominant. Then the Jacobi iteration matrix $M := I - D^{-1}A$ (D : diagonal of A) satisfies

$$\rho(M) < 1. \quad (\text{C.21})$$

(b) If furthermore, the sign conditions (C.18a,b) are satisfied, A is an M-matrix.

Proof. (i) By Exercise C.39c, D is regular. Hence M is well-defined. $M' := |M|$ has the entries

$$M'_{\alpha\beta} = 0 \quad \text{for } \alpha = \beta, \quad M'_{\alpha\beta} = |a_{\alpha\beta}/a_{\alpha\alpha}| \quad \text{for } \alpha \neq \beta.$$

In part (ii) we shall show that $\rho(M') < 1$. By Exercise C.35, $\rho(M) < 1$ also holds and proves (C.21). If, in addition, the conditions (C.18a,b) are satisfied, M fulfills condition (C.19b): $M \geq 0$. Since (C.19a,c) are also valid, Theorem C.45b shows that A is an M-matrix.

(ii) By construction, $M' \geq 0$ holds. Hence, an eigenvector $x \geq 0$ with $\|x\|_\infty = 1$ belongs to $\lambda := \rho(M') \in \sigma(M')$. Let $\alpha \in I$ be an index with $x_\alpha = 1$. We want to show that $\lambda < 1$ or $x_\gamma = 1$ for all $\gamma \in G_\alpha$, i.e., for all γ connected to α . Obviously, for an inductive proof, it is sufficient to show this assertion for those γ

that are directly connected to α , i.e., for γ with $(\alpha, \gamma) \in G(M')$. Because of the weak diagonal dominance,

$$\lambda = \lambda x_\alpha = (M'x)_\alpha = \left(\sum_{\beta \in I \setminus \{\alpha\}} |a_{\alpha\beta}| x_\beta \right) / |a_{\alpha\alpha}| \leq \left(\sum_{\beta \in I \setminus \{\alpha\}} |a_{\alpha\beta}| \right) / |a_{\alpha\alpha}| \leq 1$$

holds. The equality $\lambda = 1$ can be valid only if $x_\beta = 1$ for all β with $a_{\alpha\beta} \neq 0$, i.e., for all β with $(\alpha, \beta) \in G(A) \supset G(M')$. This completes the induction proof. If $\lambda < 1$, $\rho(M') < 1$ is shown. Otherwise, $x_\gamma = 1$ must hold for all $\gamma \in G_\alpha$. By Exercise C.39a–b, A is essentially diagonally dominant. According to this definition, there is an index $\gamma \in G_\alpha$ such that $|a_{\gamma\gamma}| > \sum_{\beta \in I \setminus \{\gamma\}} |a_{\gamma\beta}|$. Since $x_\gamma = x_\beta = 1$ for $\gamma, \beta \in G_\alpha$,

$$\lambda = \lambda x_\gamma = (M'x)_\gamma = \left(\sum_{\beta \in I \setminus \{\gamma\}} |a_{\gamma\beta}| x_\beta \right) / |a_{\gamma\gamma}| = \left(\sum_{\beta \in I \setminus \{\gamma\}} |a_{\gamma\beta}| \right) / |a_{\gamma\gamma}| < 1$$

proves the intermediate assertion $\lambda = \rho(M') < 1$ in (i). \square

Using only the weak diagonal dominance, part (ii) of the proof already shows $\lambda \leq 1$ proving the following.

Corollary C.49. Let $A \in \mathbb{K}^{I \times I}$ be a weakly diagonally dominant matrix. The Jacobi iteration matrix $M := I - D^{-1}A$ satisfies

$$\rho(M) \leq 1.$$

Remark C.50. A regular matrix $A \in \mathbb{R}^{I \times I}$ with (C.18b) and $\sum_{\beta \in I} a_{\alpha\beta} \geq 0$ for all $\alpha \in I$ is an M-matrix.

Proof. (i) Using the sign condition (C.18b), $\sum_{\beta \in I} a_{\alpha\beta} \geq 0$ implies the weak diagonal dominance of A and, in addition, $a_{\alpha\alpha} \geq 0$. If $a_{\alpha\alpha} = 0$, also $a_{\alpha\beta} = 0$ must hold for $\beta \neq \alpha$ (cf. (C.18b)). However, a regular matrix cannot contain a zero row $a_{\alpha\beta} = 0$ for all $\beta \in I$. Hence, $a_{\alpha\alpha} > 0$ and therefore (C.18a) is proved.

(ii) We conclude from $M \geq 0$ that $\lambda = \rho(M) \in \sigma(M)$, while Corollary C.49 shows that $\lambda \leq 1$. Assume that $\lambda = 1$. Let x be the corresponding eigenvector. $Mx = x$ is equivalent to $D^{-1}Ax = 0$ and $Ax = 0$ in contradiction to the regularity of A . Hence $\rho(M) < 1$ holds, and Theorem proves the M-matrix property of A . \square

Lemma C.51. For any M-matrix A , there is $A' := \Delta^{-1}A\Delta$ with a diagonal matrix $\Delta \geq 0$ such that the strict inequalities $\sum_{\beta \in I} a'_{\alpha\beta} > 0$ hold for all $\alpha \in I$.

Proof. (i) First, we assume that A is irreducible, so that $A^{-1} > 0$. By Theorem C.25, there is a positive eigenvector $x > 0$ with $A^{-1}x = \lambda x$, $\lambda > 0$. The vector x can be written as $x = \Delta \mathbf{1}$, where $\mathbf{1} \in \mathbb{R}^I$ has the coefficients $\mathbf{1}_\alpha = 1$ ($\alpha \in I$) and $\Delta = \text{diag}\{x_\alpha : \alpha \in I\}$. Equation $A^{-1}x = \lambda x$ is equivalent to $\frac{1}{\lambda} \Delta \mathbf{1} = \frac{1}{\lambda} x = Ax = A\Delta \mathbf{1}$ and $\Delta^{-1}A\Delta \mathbf{1} = \Delta^{-1}\frac{1}{\lambda} \mathbf{1} > 0$, i.e., $\sum_{\beta \in I} a'_{\alpha\beta} > 0$.

(ii) Decompose A as in (C.8) with irreducible diagonal blocks A^{ii} . For simplicity, we assume a 2×2 block structure $\begin{bmatrix} A^{11} & A^{12} \\ 0 & A^{22} \end{bmatrix}$. According to part (i), there are diagonal matrices Δ_i with $\Delta_i^{-1}A^{ii}\Delta_i\mathbf{1} > 0$. Set $\Delta := \text{blockdiag}\{\Delta_1, \varepsilon\Delta_2\}$. For sufficiently small $\varepsilon > 0$, we have $\Delta^{-1}A\Delta = \begin{bmatrix} A^{11} & \varepsilon A^{12} \\ 0 & A^{22} \end{bmatrix} \mathbf{1} > 0$. Analogously, one treats the general case of (C.8). \square

Applying Corollary C.46, we obtain the following.

Corollary C.52. Let the irreducibly diagonally dominant matrix $A \in \mathbb{R}^{I \times I}$ satisfy the sign conditions (C.18a,b). Then A is an M-matrix.

C.4.4 Further Criteria

In the following, we describe situations in which M-matrices generate new ones.

Theorem C.53. Let $A \in \mathbb{R}^{I \times I}$ be an M-matrix and let $B \geq A$ satisfy (C.18b): $b_{\alpha\beta} \leq 0$ for $\alpha \neq \beta$. Then B is also an M-matrix. Further, the inequalities

$$0 \leq B^{-1} \leq A^{-1} \quad (\text{C.22})$$

hold. If, in addition, A is irreducible and $B \neq A$, then even $0 \leq B^{-1} < A^{-1}$ holds.

Proof. (i) Let $M = I - D^{-1}A$ and $M_B = I - D_B^{-1}B$ be the respective Jacobi iteration matrices. One verifies that $0 \leq D_B^{-1} \leq D^{-1}$ and $0 \leq M_B \leq M$. $A^{-1} = (\sum_{\nu=0}^{\infty} M^\nu)D^{-1}$ and $B^{-1} = (\sum_{\nu=0}^{\infty} M_B^\nu)D_B^{-1}$ prove (C.22).

(ii) According to Remark C.47, $A^{-1} > 0$ holds for an irreducible A . Set $A(\lambda) := A + \lambda(B - A)$. For $0 \leq \lambda \leq 1$, we have $A = A(0) \leq A(\lambda) \leq A(1) = B$. The derivative

$$C(\lambda) := \frac{d}{d\lambda} A(\lambda)^{-1} = -A(\lambda)^{-1}(B - A)A(\lambda)^{-1}$$

is nonpositive: $C(\lambda) \leq 0$, because $A(\lambda)^{-1} > 0$ (cf. (C.22)) and $B - A \geq 0$. The particular choice $\lambda = 0$ yields $C(0) = -A^{-1}(B - A)A^{-1} < 0$, since any vector $x \geq 0$ leads to $A^{-1}x > 0$, $(B - A)A^{-1}x > 0$, and $A^{-1}(B - A)A^{-1}x > 0$; hence, $C(0) < 0$ (cf. Exercise C.20b). The inequalities $C(0) < 0$ and $C(\lambda) \leq 0$ prove $A^{-1} > A(\lambda)^{-1} \geq B^{-1}$ for all $0 < \lambda \leq 1$. \square

Theorem C.54. Any principal submatrix of an M-matrix is again an M-matrix. More precisely: If $B = (a_{\alpha\beta})_{\alpha,\beta \in I'}$ for $I' \subset I$, then B is an M-matrix with $0 \leq (B^{-1})_{\alpha\beta} \leq (A^{-1})_{\alpha\beta}$ for $\alpha, \beta \in I'$. If, furthermore, A is irreducible and $I' \subsetneq I$ is a nonempty subset, then the strengthened inequality $(B^{-1})_{\alpha\beta} < (A^{-1})_{\alpha\beta}$ is valid for $\alpha, \beta \in I'$.

Proof. Define $B' \in \mathbb{R}^{I \times I}$ by the entries $b'_{\alpha\beta} := \begin{cases} a_{\alpha\beta} & \text{for } \alpha, \beta \in I' \text{ or } \alpha = \beta \in I \\ 0 & \text{otherwise} \end{cases}$.

B' has the form $\text{blockdiag}\{B, D_2\}$, with D_2 being the diagonal of the block

$A^{22} = (a_{\alpha\beta})_{\alpha,\beta \in I \setminus I'}$. Hence, $B'^{-1} = \text{blockdiag}\{B^{-1}, D_2^{-1}\}$ holds. We apply Theorem C.53 to B' and obtain $0 \leq B'^{-1} \leq A^{-1}$ or $0 \leq B'^{-1} < A^{-1}$, respectively. A restriction to the first block yields the assertion. \square

Exercise C.55. Prove: (a) 2×2 matrices A are M-matrices if and only if (C.18a,b) and $\det(A) > 0$ hold.

(b) M-matrices A have a positive determinant: $\det(A) > 0$. Hint: Discuss the determinant of $A(\lambda) := D + \lambda(A - D)$ for $0 \leq \lambda \leq 1$ with $D := \text{diag}\{A\}$.

(c) All principal minors of an M-matrix are positive. Hint: Apply Theorem C.54.

Gauss elimination will play an important role in the following proof. Its basic operation is the elimination of an entry $a_{\beta\alpha}$ ($\alpha \neq \beta$) by subtracting the α -th row. The corresponding transformation $A \mapsto A'$ is described by the matrix $T^{\beta\alpha}$:

$$A' = T^{\beta\alpha} A \text{ with } T_{\beta\alpha}^{\beta\alpha} = -\frac{a_{\beta\alpha}}{a_{\alpha\alpha}}, \quad T_{\nu\nu}^{\beta\alpha} = 1 \ (\nu \in I), \quad T_{\nu\mu}^{\beta\alpha} = 0 \ (\nu \neq \mu). \quad (\text{C.23})$$

The usual Gauss elimination (without pivoting) requires an ordered index set I and performs eliminations in the succession $(\beta, \alpha) = (2, 1), (3, 1), \dots, (n, 1), (3, 2), (4, 2), \dots, (n, 2), \dots, (n, n-1)$ below the diagonal, resulting in an upper triangular matrix U . The diagonal elements $p_i = u_{ii}$ of U are the pivot elements. The following considerations are even simpler if we eliminate all off-diagonal entries at $(\beta, \alpha) = (2, 1), (3, 1), \dots, (n, 1), (1, 2), (3, 2), \dots, (n, 2), (1, 3), \dots$, leading to the diagonal matrix $D = \text{diag}\{p_i : i \in I\}$ of the pivot elements. Let H_i be the principal minor of A :

$$H_i := \det((a_{k\ell})_{1 \leq k, \ell \leq i}) \quad \text{for } 1 \leq i \leq n, \quad H_0 := 1.$$

A simple consideration shows that

$$p_i = H_i / H_{i-1} \quad (1 \leq i \leq n), \quad (\text{C.24})$$

provided that $H_{i-1} \neq 0$. This implies that the above elimination process can be performed without pivoting if $H_{i-1} \neq 0$ for all i (cf. Gantmacher [144, p. 36]).

The statement of Exercise C.55c can be extended as follows.

Theorem C.56. Under assumption (C.18b), A is an M-matrix if and only if all principal minors are positive.

Proof. (i) Since Exercise C.55c proves one direction, it remains to show that positive principal minors imply the M-matrix property. The diagonal entries $a_{\alpha\alpha}$ are the determinants of the principal 1×1 submatrices $(a_{\alpha\alpha})$. This ensures (C.18a): $a_{\alpha\alpha} > 0$. Because of (C.24), Gauss elimination needs no pivoting.

(ii) First, we prove that the elimination step (C.23) preserves the sign conditions (C.18a,b). A and A' differ only in the β -th row. Since $\varkappa := T_{\beta\alpha}^{\beta\alpha} = -\frac{a_{\beta\alpha}}{a_{\alpha\alpha}} \geq 0$, the entries $a_{\beta\delta}$ for $\delta \neq \alpha$ become smaller: $a'_{\beta\delta} := a_{\beta\delta} + \varkappa a_{\alpha\delta} \leq 0$, while $a'_{\beta\alpha} = 0$ again satisfies (C.18b). The only problem is raised by condition (C.18a): does $a'_{\beta\beta} > 0$ hold again? As seen above, the diagonal element decreases with each elimination step. Since at the end of the elimination, it represents the pivot element p_i , Eq. (C.24) with $H_\beta, H_{\beta-1}^{-1} > 0$ proves the inequality $a'_{\beta\beta} \geq p_\beta > 0$.

(iii) Performing the elimination (above and below the diagonal), we obtain the diagonal matrix D of the positive pivot elements. Denoting the elimination matrices $T^{\beta\alpha}$ for the respective index pairs by T_1, T_2, \dots, T_N , one arrives at the representation

$$T_N T_{N-1} \cdots T_1 A = D, \quad \text{i.e., } A^{-1} = D^{-1} T_N T_{N-1} \cdots T_1. \quad (\text{C.25})$$

Since, according to (ii), all intermediate matrices fulfil the conditions (C.18a,b), $T_i \geq 0$ holds. Together with $D^{-1} \geq 0$, the missing M-matrix property (C.18c), $A^{-1} \geq 0$, is obtained. \square

The close connection between M-matrices and positive definite matrices is underlined by comparing Theorem C.56 and the next result.

Remark C.57. A Hermitian matrix is positive definite if and only if all principal minors are positive.

Proof. (i) By Lemma C.4c, all principal submatrices are positive definite. Hence, it is sufficient to show that $\det(A) > 0$ holds for positive definite A . This follows from $\det(A) = \prod \lambda_i$ and the positivity $\lambda_i > 0$ of all eigenvalues $\lambda_i \in \sigma(A)$ (cf. Lemma C.3).

(ii) The determinant of $A(\lambda) := A + \lambda I$ can be expanded into

$$\det(A) + \sum_i \lambda \det A_i(\lambda),$$

where $A_i(\lambda)$ is the principal submatrix of $A(\lambda)$ for the index set $I_i := I \setminus \{i\}$. The analogous expansion of the determinants of $A_i(\lambda)$ yields a polynomial $p(\lambda) = \det A(\lambda) = \sum a_\nu \lambda^\nu$ with positive coefficients (e.g., $a_0 = \det(A) > 0$). Hence, $A + \lambda I$ is regular for all $\lambda > 0$. Since its eigenvalues are $\lambda_i + \lambda$, all eigenvalues $\lambda_i \in \sigma(A)$ must be positive. According to Lemma C.3, A is positive definite. \square

Remark C.57 describes one of the numerous characterisations of M-matrices. The interested reader may find fifty different characterisations in Berman–Plemmons [45]. Combining Theorem C.56 and Remark C.57 yields the following.

Theorem C.58. *A positive definite matrix satisfying the sign condition (C.18b) is an M-matrix. On the other hand, a Hermitian M-matrix is positive definite.*

We remark that a matrix $A \in \mathbb{R}^{I \times I}$ with $A > 0$ and the sign condition (C.18b) is called a *Stieltjes matrix*. The next lemma continues the discussion of the Gauss elimination.

Lemma C.59. *If A is an M-matrix and A' is obtained by one Gauss elimination step (C.23), then A' is again an M-matrix.*

Proof. Choose the ordering of the indices such that $\alpha = 1$ and $\beta = 2$. Then $T^{\beta,\alpha}$ in (C.23) describes the first step T_1 of the complete elimination process $T_N T_{N-1} \cdots T_1 A = D$ (cf. (C.25)). $A'^{-1} = (T_1 A)^{-1} \geq 0$ follows from $T_i \geq 0$, $D^{-1} \geq 0$, and $(T_1 A)^{-1} = D^{-1} T_N T_{N-1} \cdots T_2$. Since the conditions (C.18a,b) are already proved in part (ii) of the proof of Theorem 16, A' is an M-matrix. \square

C.5 H-Matrices

The term ‘M-matrix’ can be generalised as follows. The following construction of B changes the signs of the entries $a_{\alpha\beta}$ in such a way that $b_{\alpha\alpha} \geq 0$ and $b_{\alpha\beta} \leq 0$. The letter ‘H’ refers to Hadamard (cf. Ostrowski [301]).

Definition C.60. $A \in \mathbb{C}^{I \times I}$ is called an *H-matrix* if $B := |D| - |A - D|$ with $D := \text{diag}\{A\}$ is an M-matrix.

Remark C.61. (a) Assume that the diagonal part D of A is regular. If A is strictly, irreducibly, or essentially diagonally dominant, A is an H-matrix.

(b) The matrix $M_B := I - |D|^{-1}B = |D|^{-1}|A - D|$ satisfies $\rho(M_B) < 1$.

Proof. Set $B := |D| - |A - D|$ as in Definition C.60 and $M := I - |D|^{-1}B = |D|^{-1}|A - D|$. Apply Theorem C.48b to $B := |D| - |A - D|$ instead of A to prove part (a). For part (b), use Theorem C.48a. \square

The counterpart of Lemma C.59 can easily be verified.

Lemma C.62. If A is an H-matrix and A' is obtained by one Gauss elimination step (C.23), then A' is again an H-matrix.

C.6 Schur Complement

Let $I = \{I_1, I_2\}$ be a block structure, The blockwise elimination of the block-matrix $M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$ leads to $\begin{bmatrix} I & B' \\ 0 & S \end{bmatrix}$ with $B' := A^{-1}B$ and the *Schur complement* $S := D - CA^{-1}B$, provided that A is regular. The inverse of M has the representation

$$M^{-1} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}BTCA^{-1} & A^{-1}BT \\ -TCA^{-1} & T \end{bmatrix} \quad \text{with } T := S^{-1}$$

provided that S is regular. The following statements are elementary.

Remark C.63. Let A be regular. (a) M is regular if and only if S is regular.

(b) If S is regular, $T = S^{-1}$ is the restriction $M^{-1}|_{I_2 \times I_2}$.

Proposition C.64. Let M be an M-matrix, H-matrix, or positive definite matrix. Then S has the same property.

Proof. Blockwise elimination represents the product of all elementary eliminations (C.23) with indices α corresponding to the columns of the first block and $\beta \in I \setminus \{\alpha\}$. Multiple applications of Lemma C.59 prove the M-matrix property of S . Use Lemma C.62 for the H-matrix property.

If $M > 0$, Lemma C.4 proves that $M^{-1} > 0$ and $A > 0$ also holds. By Remark C.63b, $T = M^{-1}|_{I_2 \times I_2} > 0$ holds implying $T^{-1} = S > 0$. \square

Appendix D

Hierarchical Matrices

Abstract The fully populated matrices arising from boundary element methods can be approximated by hierarchical matrices. This reduces the storage cost to almost linear complexity. Another important property of hierarchical matrices is the almost linear cost of matrix operations, including matrix inversion and LU decomposition. In the case of finite element methods discretising elliptic problems, the matrices are sparse, but the inverse of the factors of the LU decomposition can be approximated with any accuracy. The computational cost for accuracy ε is $\mathcal{O}(n \log^*(n) \log^*(\frac{1}{\varepsilon}))$. In Section D.1, we introduce the idea of the block-structured matrices using low-rank matrix blocks and illustrate the operation cost by a model problem. Constructing hierarchical matrices is the subject of Section D.2. In §D.2.9, we explain how fully populated matrices discretising integral operators can be approximated by hierarchical matrices and why the error decreases exponentially with increasing local rank. The matrix operations are described in Section D.3.

D.1 Introduction

D.1.1 Fully Populated Matrices

General fully populated $n \times n$ matrices are described by n^2 entries. For large n , this fact causes problems not only for storing the matrix but also for evaluating all entries. Moreover, matrix operations lead to $\mathcal{O}(n^2)$ or $\mathcal{O}(n^3)$ arithmetic operations.

If a matrix is sparse (cf. §1.7), only the storage cost and the cost of matrix-vector multiplication is proportional to n . This fact has led to the impression that more costly operations as matrix-matrix multiplications, matrix inversions, or LU decompositions should be avoided for large-scale matrices. Note that products of sparse matrices are less sparse and higher powers of a sparse matrix become fully populated. The inverse of a sparse matrix is in general fully populated (cf. Corollary C.46). Except for band matrices, the LU decomposition leads to a fill-in, although, in this case, sparsity can partially be saved.

The technique of hierarchical matrices (\mathcal{H} -matrix technique) applies to matrices related to elliptic problems. This includes two types of matrices. The first group are discretisations of elliptic problems formulated by the integral equation method. The fully populated matrices that arise are an ideal object of this technique. The second group are usual finite element discretisations of elliptic boundary value problems. Then all matrix operations, in particular the LU decomposition, can be performed with a cost¹ of $\mathcal{O}(n \log^* n)$.

The \mathcal{H} -matrix technique introduces an additional error. The reduction of the storage cost to $\mathcal{O}(n \log^* n)$ as well as the \mathcal{H} -matrix operations go together with a *truncation error*. However, this error can easily be controlled. Usually, it is chosen of a similar or of smaller size than the already existing discretisation error.

For solving a linear system $Ax = b$, the LU decomposition $A \approx LU$ is exceptionally important as shown in §13.4. Since computing LU requires only the data of the matrix A , the associated iteration belongs to the class of purely algebraic methods (see Definition 2.2b and case (i) in Remark 7.7).

Remark D.1. The availability of cheap matrix operations enables further applications, e.g., computing matrix functions (cf. Higham [222]) like the matrix exponential $\exp(A)$ and solving matrix equations as the Lyapunov, Sylvester, or Riccati equation.

The construction of hierarchical matrices consists of two ingredients: (i) a very particular block decomposition, and (ii) the use of low-rank matrices for the matrix blocks. The block decomposition is individually constructed for each matrix A (cf. §D.2.3). A typical example of the block structure that arises is shown in Figure D.1 (middle) for a 128×128 matrix. Note that the size of the blocks increases with their distance from the diagonal. The number of blocks is bounded by $\mathcal{O}(n)$. Each block is filled by a low-rank matrix (cf. §D.1.2). The analysis shows that (i) storage of low-rank matrices and matrix operations involving low-rank matrices are cheap and that (ii) truncation to rank r leads to an error decreasing exponentially with respect to r . Given a tolerance ε , the required rank is $r = \mathcal{O}(\log^* 1/\varepsilon)$.

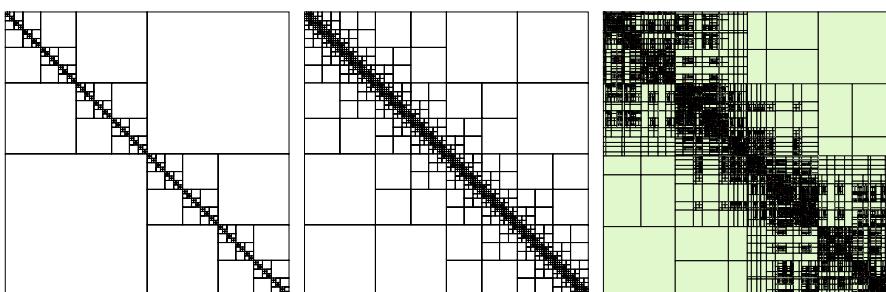


Fig. D.1 Left: simple block partition \mathcal{H}_7 . Middle: admissible block partition. Right: partition for a real-life application of size 447488.

¹ The asterisk in \log^* replaces some exponent.

D.1.2 Rank- r Matrices

Let I and J be index sets. If $M \in \mathbb{K}^{I \times J}$ satisfies $\text{rank}(M) \leq r$, there are factors A and B such that

$$M = AB^T \quad \text{with } A \in \mathbb{K}^{I \times r}, \quad B \in \mathbb{K}^{J \times r}. \quad (\text{D.1})$$

For instance, the QR decomposition in Remark A.26b produces the factorisation $M = QR$ with an orthogonal matrix $A := Q$ and an upper triangular matrix $B^T := R$. On the other hand, if M satisfies (D.1), then $\text{rank}(M) \leq r$ follows.

Let $a^{(\ell)}$ and $b^{(\ell)}$ be the columns of A and B respectively. Then (D.1) is equivalent to

$$M = \sum_{\ell=1}^r a^{(\ell)} b^{(\ell)\top}. \quad (\text{D.2})$$

The number r in (D.2) is called the *representation rank*, even if $\text{rank}(M) < r$.

In the following, it is important that A and B are not only existing but explicitly available. The mapping $(A, B) \mapsto M = AB^T$ describes the representation of M by its factors (rank- r format). In contrast, $M \in \mathcal{F}$ describes the *full format*, i.e., M is described in the standard way by its entries.

Definition D.2 ($\mathcal{F}, \mathcal{R}_r$). (a) The set \mathcal{F} is formed by all matrices $M \in \mathbb{K}^{I \times J}$ which are explicitly given by their entries M_{ij} ($i \in I$, $j \in J$).

(b) The set \mathcal{R}_r is formed by all matrices² M in (D.1) with explicitly known factors $A \in \mathcal{F}$ and $B \in \mathcal{F}$.

Remark D.3. The storage size of $M \in \mathcal{F} \cap \mathbb{K}^{I \times J}$ is $\#I \#J$, while the storage size of $M \in \mathcal{R}_r \cap \mathbb{K}^{I \times J}$ is $r (\#I + \#J)$.

Operations involving matrices from \mathcal{R}_r are much cheaper than those for fully populated matrices:

$$\begin{aligned} \text{matrix-vector multiplication: } Mx &= A \cdot (B^T x), \\ \text{matrix-matrix multiplication: } M' M'' &= A' \cdot ((B'^T \cdot A'') \cdot B''^T) \end{aligned} \quad (\text{D.3})$$

Let $M \in \mathcal{R}_r \cap \mathbb{K}^{I \times J}$, $M' \in \mathcal{R}_{r'} \cap \mathbb{K}^{I \times J}$, $M'' \in \mathcal{R}_{r''} \cap \mathbb{K}^{J \times K}$. The respective numbers of arithmetic operations $(+, -, *)$ are

$$\begin{aligned} N_{\text{MV}} &= 2r (\#I + \#J) - \#I - r, \\ N_{\text{MM}} &= 2r' r'' (\#J + \#K) - r' (\#K + r''). \end{aligned}$$

Note that $M := M' M'' \in \mathcal{R}_r$ has the representation AB^T with the explicitly available factors $A := A' \in \mathbb{K}^{I \times r'}$ and $B := B'' \cdot (A''^T \cdot B') \in \mathbb{K}^{K \times r'}$.

The (exact) addition of low-rank matrices increases the representation rank, but does not require any arithmetic operators. The sum of $M' = A'B'^T \in \mathcal{R}_{r'}$ and $M'' = A''B''^T \in \mathcal{R}_{r''}$ yields $M \in \mathcal{R}_{r'+r''}$:

² It would be more precise to consider triples (M, A, B) with $M = AB^T$, but this would complicate the notation.

$$M = M' + M'' = AB^\top \quad \text{with} \quad \begin{cases} A := [A' \ A''] \in \mathbb{R}^{I \times \{1, \dots, r' + r''\}}, \\ B := [B' \ B''] \in \mathbb{R}^{J \times \{1, \dots, r' + r''\}}. \end{cases} \quad (\text{D.4})$$

Although the computational cost is zero, the storage cost increases. Therefore, after one or more additions, the result has to be truncated to smaller rank. Optimal truncation uses the singular value decomposition (SVD) of M (cf. Proposition A.45). In general, SVD is too costly for large index sets I, J (cf. Remark A.44). However, the representation of $M \in \mathcal{R}_r$ by AB^\top enables a much cheaper application of SVD.

Lemma D.4. *Assume $M \in \mathcal{R}_s$ with $M = AB^\top$ and $r < s$. Then the SVD approximation $M_r = A'B'^\top \in \mathcal{R}_r$ can be determined with arithmetic cost*

$$N_{\text{SVD}} \leq 2(\#I + \#J)(2s + r)s + \frac{65}{3}s^3 + \dots .$$

Proof. Determine the QR decompositions $X = Q_X R_X$ and $Y = Q_Y R_Y$ involving $Q_X \in \mathbb{R}^{I \times s}$, $Q_Y \in \mathbb{R}^{J \times s}$, $R_X, R_Y \in \mathbb{R}^{s \times s}$ (cf. Remark A.26b). Compute the product $P := R_X R_Y^\top \in \mathbb{R}^{s \times s}$. Let $P = U \Sigma V^\top$ be the singular value decomposition of P with $\Sigma = \text{diag}\{\sigma_1, \dots, \sigma_s\}$ and set $\Sigma_r = \text{diag}\{\sigma_1, \dots, \sigma_r, 0, \dots\}$. Then, $M_r = Q_X U \Sigma_r V^\top Q_Y^\top$ is the desired SVD truncation. To obtain again a representation of the form $A'B'^\top$, set $A' := Q_X U \Sigma$ and $B' := Q_Y V$. Summation of the amount of work yields N_{SVD} described above (cf. [198, Remark 2.18]). \square

Corollary D.5. *In the case of $s = 2$ and $r = 1$, truncation can be performed by $9\#I + 8\#J + 19$ arithmetic operations (cf. [198, Corollary 2.19b]).*

D.1.3 Model Format

Since the block structure in Figure D.1 (middle) is already rather involved, we introduce a much simpler structure which, nevertheless, shows the same characteristics. For $p \in \mathbb{N}_0$, $n := 2^p$, and $k \geq 1$, we define a recursion for the set $\mathcal{H}_p(r) \subset \mathbb{R}^{n \times n}$ of matrices. For $p = 0$ (i.e., $n = 1$), all $\mathcal{H}_p(r)$ are 1×1 matrices (i.e., $\mathcal{H}_p(r)$ is equivalent to \mathbb{K}). For $p > 0$, the matrices $M \in \mathcal{H}_p(r)$ are characterised by

$$M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}, \quad M_{11}, M_{22} \in \mathcal{H}_{p-1}(r), \quad M_{12}, M_{21} \in \mathcal{R}_k. \quad (\text{D.5})$$

The block structures are  ($p = 0$),  ($p = 1$),  ($p = 2$),  ($p = 3$).

Figure D.1 (left) shows the case of $n = 2^7 = 128$. Each block b of the depicted block decompositions contains a matrix block $M|_b$ of $\text{rank}(M|_b) \leq r$. Here we use the following notation.

Notation D.6. Let $M \in \mathbb{K}^{I \times J}$ be any matrix. A general block b is the Cartesian product $b = \tau \times \sigma$ of some nonempty subsets $\tau \subset I$ and $\sigma \subset J$. Hence b is a subset of $I \times J$, while the corresponding matrix block is

$$M|_b := (M_{\alpha\beta})_{\alpha \in \tau, \beta \in \sigma} \in \mathbb{K}^{\tau \times \sigma}.$$

Larger values of the representation rank r in $\mathcal{H}_p(r)$ improve the approximation of general matrices by $M \in \mathcal{H}_p(r)$, but for following considerations about the matrix operations the value of r is less relevant. Therefore we fix $r = 1$.

Let $N_{\text{bl}}(p)$ be the number of blocks in $\mathcal{H}_p(r)$. Using $N_{\text{bl}}(0) = 1$ and the recursion $N_{\text{bl}}(p) = 2 + 2N_{\text{bl}}(p - 1)$, we obtain

$$N_{\text{bl}}(p) = 3n - 2 \quad (n = 2^p).$$

Similar recursive proofs can be used for the following results (cf. [198, §3]). We require that all matrix blocks $M|_b$ of $M \in \mathcal{H}_p(1)$ be \mathcal{R}_1 matrices. Then the storage size is

$$S_p = n + 2n \log_2 n.$$

Matrix-vector multiplication $M \cdot x$ requires computing the products $y_{11} := M_{11}x_1$, $y_{12} := M_{12}x_2$, $y_{21} := M_{21}x_1$, $y_{22} := M_{22}x_2$ (cf. (D.5)) and the sums $y_{11} + y_{12}$ and $y_{21} + y_{22}$. The required work is

$$N_{\text{MV}}(p) = 4n \log_2 n - n + 2.$$

Different from the following operations, matrix-vector multiplication yields an *exact* result.

\oplus_1 (\oplus_r) denotes addition followed by truncation to rank 1 (r). Using Corollary D.5, we obtain the following cost of addition involving at least one hierarchical matrix.

Lemma D.7. *Let $n = 2^p$. The formatted additions $\oplus_1 : \mathcal{H}_p \times \mathcal{H}_p \rightarrow \mathcal{H}_p$ as well as $\oplus_1 : \mathcal{H}_p \times \mathcal{R}_1 \rightarrow \mathcal{H}_p$ and $\oplus_1 : \mathcal{R}_1 \times \mathcal{H}_p \rightarrow \mathcal{H}_p$ require*

$$17n \log_2 n + 39n - 38$$

operations.

Let $n = 2^p$. We distinguish three kinds of matrix-matrix multiplications:

- (1) $A \cdot B \in \mathcal{R}_1$ for $A, B \in \mathcal{R}_1$ with the cost $N_{R \cdot R}(p)$,
- (2a) $A \cdot B \in \mathcal{R}_1$ for $A \in \mathcal{R}_1$ and $B \in \mathcal{H}_p$ with the cost $N_{R \cdot H}(p)$,
- (2b) $A \cdot B \in \mathcal{R}_1$ for $A \in \mathcal{H}_p$ and $B \in \mathcal{R}_1$ with the cost $N_{H \cdot R}(p)$,
- (3) $A \cdot B \in \mathcal{H}_p$ for $A, B \in \mathcal{H}_p$ with the cost $N_{H \cdot H}(p)$.

In cases (1) and (2a,b), the results are exact since the operation reduces to matrix-vector multiplication.

In the first case, the solution is $N_{R \cdot R}(p) = 3n - 1$ (compute the product by $(ab^\top)(cd^\top) = a \cdot ((b^\top c)d^\top)$).

In the case of $A \in \mathcal{H}_p$ and $B \in \mathcal{R}_1$, we use $A \cdot ab^\top = (Aa) \cdot b^\top$; i.e., the result is $a'b^\top \in \mathcal{R}_p$ with $a' := Aa$. This requires one matrix-vector multiplication $A \cdot a$. The cost amounts to $N_{H \cdot R}(p) = 4n \log_2 n - n + 2$.

Similarly, for $B \in \mathcal{R}_p$ and $A \in \mathcal{H}_1$ we perform $BA = ab^T \cdot A = a \cdot (A^T)^T$ so that $N_{R \cdot H}(p) = N_{H \cdot R}(p)$.

In the third case of $A, B \in \mathcal{H}_p$, the product AB is of the form

$$\begin{aligned} & \left[\begin{matrix} \mathcal{H}_{p-1} & \mathcal{R}_{p-1} \\ \mathcal{R}_{p-1} & \mathcal{H}_{p-1} \end{matrix} \right] \cdot \left[\begin{matrix} \mathcal{H}_{p-1} & \mathcal{R}_{p-1} \\ \mathcal{R}_{p-1} & \mathcal{H}_{p-1} \end{matrix} \right] \\ &= \left[\begin{matrix} \mathcal{H}_{p-1} \cdot \mathcal{H}_{p-1} + \mathcal{R}_{p-1} \cdot \mathcal{R}_{p-1} & \mathcal{H}_{p-1} \cdot \mathcal{R}_{p-1} + \mathcal{R}_{p-1} \cdot \mathcal{H}_{p-1} \\ \mathcal{R}_{p-1} \cdot \mathcal{H}_{p-1} + \mathcal{H}_{p-1} \cdot \mathcal{R}_{p-1} & \mathcal{R}_{p-1} \cdot \mathcal{R}_{p-1} + \mathcal{H}_{p-1} \cdot \mathcal{H}_{p-1} \end{matrix} \right]. \end{aligned}$$

On level $p - 1$, all three types of multiplications appear. The multiplication $\mathcal{H}_{p-1} \cdot \mathcal{H}_{p-1}$ of the third type requires an approximation by \odot . Finally, addition via \oplus_1 has to be performed. Counting the operations, we derive the recursion

$$\begin{aligned} N_{H \cdot H}(p) &= 2N_{H \cdot H}(p-1) + 2N_{R \cdot R}(p-1) + 2N_{H \cdot R}(p-1) \\ &\quad + 2N_{R \cdot H}(p-1) + 2N_{H+R}(p-1) + 2N_{R+R}(p-1). \end{aligned}$$

Inserting the known quantities $N_{R \cdot R}, N_{H \cdot R}, N_{H+R}, N_{R+R}$, we obtain $N_{H \cdot H}(p) = 2N_{H \cdot H}(p-1) + 25pn + 32n - 32$. Together with the starting value $N_{H \cdot H}(0) = 1$, we get the following operation costs:

$$\begin{aligned} N_{H \cdot H}(p) &= \frac{25}{2}n \log_2^2 n + \frac{89}{2}n \log_2 n - 31n + 32, \\ N_{H \cdot R}(p) &= N_{R \cdot H}(p) = 4n \log_2 n - n + 2, \\ N_{R \cdot R}(p) &= 3n - 1. \end{aligned}$$

Finally, we want to approximate the inverse M^{-1} of a matrix $M \in \mathcal{H}_p$. For this purpose, we define the inversion mapping $inv : D_p \subset \mathcal{H}_p \rightarrow \mathcal{H}_p$ recursively (D_p : domain of inv). For $p = 0$, we define $inv(M) := M^{-1}$ as the exact inverse of the 1×1 -matrix M , provided that $M \neq 0$. Let inv be defined on $D_{p-1} \subset \mathcal{H}_{p-1}$. The (exact) inverse of M with the block structure (D.5) is

$$M^{-1} = \begin{bmatrix} M_{11}^{-1} + M_{11}^{-1}M_{12}S^{-1}M_{21}M_{11}^{-1} & -M_{11}^{-1}M_{12}S^{-1} \\ -S^{-1}M_{21}M_{11}^{-1} & S^{-1} \end{bmatrix} \quad (\text{D.6})$$

involving the *Schur complement* $S := M_{22} - M_{21}M_{11}^{-1}M_{12}$ (cf. §C.6). The representation (D.6) and therefore the following algorithm also requires M_{11} to be regular. If M and M_{11} are regular, then the Schur complement S is also regular.

Again, the inversion of M at level p involves two inversions (M_{11}^{-1} and S^{-1}) at level $p - 1$. This yield a similar recursion formula as for $N_{H \cdot H}(p)$. Its solution is

$$N_{\text{inv}}(p) = \frac{25}{2}n \log_2^2 n + \frac{55}{2}n \log_2 n - 69n + 70.$$

Concerning a detailed derivation of the mentioned recursions see [198, §3.7].

D.2 Construction

In the following, we are searching for a suitable block decomposition. It is impossible to consider *all* block partitions for two reasons. First, the number of all possible partitions is far too large; second, the matrix operations causes further restrictions. It turns out that in both aspects a tree structure is very helpful. In §D.2.1, we introduce *cluster trees* $T(I)$ and $T(J)$ providing vector blocks in \mathbb{K}^I and \mathbb{K}^J . The *block cluster tree* $T(I \times J)$ constructed in §D.2.2 provides a variety of matrix blocks of all sizes. The choice of the final partition is described in §D.2.3. In §D.2.9, we illustrate the discretisation of an integral operator and the estimate of the approximation error.

D.2.1 Cluster Trees

D.2.1.1 General Structure

The tree $T(I)$ will describe the decomposition of the index set I into subsets. The elements (*vertices*) of the tree $T(I)$ are called *clusters* and denoted by τ or σ . If a cluster τ is not decomposed further, it is a leaf. The *set of leaves* is denoted by $\mathcal{L}(T(I))$. Otherwise, τ is decomposed into disjoint subsets τ_1, \dots, τ_s . These τ_j are called the *sons* of τ . The set of sons is denoted by

$$S(\tau) = \{\tau_1, \dots, \tau_s\}.$$

Besides the general properties of a tree, we require

$$\begin{aligned} I &= \text{root}(T(I)), \quad \tau \neq \emptyset \quad \text{for all } \tau \in T(I), \\ \#S(\tau) > 1 \text{ and } \bigcup_{\sigma \in S(\tau)} \sigma &= \tau \text{ (disjoint union)} \quad \text{for all } \tau \in T(I) \setminus \mathcal{L}(T(I)). \end{aligned}$$

One concludes that $\tau \subset I$ holds for all $\tau \in T(I)$.

The left and middle partitions in Figure D.1 have blocks of size 1×1 close to the diagonal. For practical purpose, it is advantageous to avoid too small blocks. Therefore we fix some $n_{\min} \in \mathbb{N}$ (e.g., $n_{\min} = 16$ or $n_{\min} = 32$) and require

$$\#\tau \leq n_{\min} \quad \text{if and only if} \quad \tau \in \mathcal{L}(T(I)).$$

Usually, a binary tree is preferred, i.e., $\#S(\tau) = 2$ for $\tau \in T(I) \setminus \mathcal{L}(T(I))$. An exception is mentioned in §13.2.

Each vertex of a tree has a level-number which is defined recursively by zero for the root I and $\text{level}(\sigma) = \text{level}(\tau) + 1$ for $\sigma \in S(\tau)$. We introduce the notation

$$T^{(\ell)}(I) := \{\tau \in T(I) : \text{level}(\tau) = \ell\}, \quad \text{depth}(T(I)) := \max\{\text{level}(\tau) : \tau \in T(I)\} \tag{D.7}$$

If necessary, we write $\text{level}_{T(I)}(\tau)$ to refer to the underlying tree.

D.2.1.2 Concrete Construction

Finite element or difference discretisations use *nodal points* (grid points). Each index $i \in I$ is connected with a nodal point $\xi_i \in \mathbb{R}^d$. This leads us to the following construction. A subset $\tau \subset I$ corresponds to the set of nodal points

$$X_\tau := \{\xi_i : i \in \tau\} \subset \mathbb{R}^d.$$

Next, we consider d -dimensional cuboids $Q = [a_1, b_2] \times [a_2, b_2] \times \dots \times [a_d, b_d]$ containing X_τ . The smallest Q containing X_τ is called the *bounding box* of X_τ and denoted by

$$Q_{\min}(X_\tau).$$

Let $\xi_{i,j}$ ($1 \leq j \leq d$) denote the components of $\xi_i \in X_\tau$. Then $Q_{\min}(X_\tau) = [a_1, b_2] \times \dots \times [a_d, b_d]$ holds with $a_j := \min_{i \in \tau} \{\xi_{i,j}\}$ and $b_j := \max_{i \in \tau} \{\xi_{i,j}\}$.

The general construction of a binary cluster tree $T(I)$ is of the following form:

- (a) Start with the root I associated with the box $Q_I := Q_{\min}(X_I)$ and set $\tau := I$.
- (b) If $\#\tau \leq n_{\min}$, stop the decomposition, otherwise continue.
- (c) Split Q_τ into two disjoint subboxes Q_τ^1 and Q_τ^2 . This induces the subsets $\tau_k := \{i \in \tau : \xi_i \in Q_\tau^k\}$ for $k = 1, 2$, representing the two sons of τ . Continue the algorithm at item (b) with τ_k ($k = 1, 2$) instead of τ . The box Q_{τ_k} has to satisfy $X_{\tau_k} \subset Q_{\tau_k} \subset Q_\tau^k$.

The concrete construction depends on the way how Q_τ is divided into Q_τ^1 and Q_τ^2 and how Q_{τ_k} is defined. Note that step (c) is performed only if $\#\tau > n_{\min} \geq 1$.

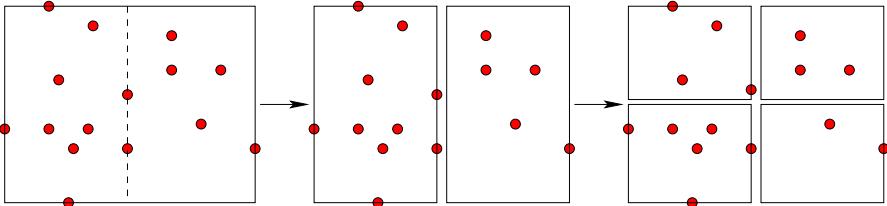


Fig. D.2 Regular geometric partition.

(A) Regular Geometric Bisection. Let $Q_\tau = [a_1, b_2] \times \dots \times [a_d, b_d]$. Define j as the index corresponding to the largest side length $b_j - a_j$. Set $m_j := \frac{a_j + b_j}{2}$. Divide Q_τ into

$$\begin{aligned} Q_\tau^1 &:= [a_1, b_2] \times \dots \times [a_j, m_j] \times \dots \times [a_d, b_d], \\ Q_\tau^2 &:= [a_1, b_2] \times \dots \times (m_j, b_j] \times \dots \times [a_d, b_d] \end{aligned}$$

and set $Q_{\tau_k} := Q_\tau^k$ ($k = 1, 2$). After ℓ steps, all boxes are similar and have the volume $\text{vol}(Q_\tau) = 2^{-\ell} \text{vol}(Q_I)$.

It may happen that X_τ is completely contained, e.g., in Q_τ^1 . Then $\tau_1 := \tau$ and $\tau_2 := \emptyset$ hold. In this case, τ_2 is omitted as a son and τ has only one son³ $\tau_1 = \tau$.

(B) Geometric Bisection by Bounding Boxes. Determine Q_τ^k and τ_k as above and set $Q_{\tau_k} := Q_{\min}(X_{\tau_k})$ ($k = 1, 2$). In this case, one verifies that⁴ $Q_{\tau_1} \neq \emptyset$ and $Q_{\tau_2} \neq \emptyset$ (cf. Fig. D.3).

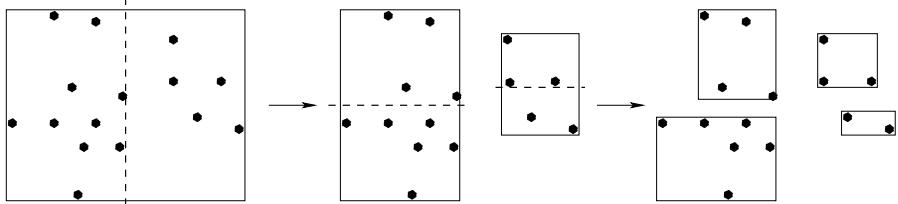


Fig. D.3 Geometric partition by bounding boxes.

In the first two approaches the boxes are bisected. The subboxes Q_τ^1 and Q_τ^2 have the same size, but there is no guaranty that τ_1 and τ_2 are of a similar cardinality. In contrast, the next approach ensures that $|\#\tau_1 - \#\tau_2| \leq 1$, while Q_{τ_1} and Q_{τ_2} may have different size.

(C) Cardinality-Based Bisection. Let Q_τ and j as in case (A). Sort all indices of τ such that $\tau = \{i_1, i_2, \dots, i_{\#\tau}\}$ and $\xi_{i_1,j} \leq \xi_{i_2,j} \leq \dots \leq \xi_{i_{\#\tau},j}$. Set $\mu := \lceil \#\tau/2 \rceil$ and define $\tau_1 := \{i_1, i_2, \dots, i_\mu\}$ and $\tau_2 := \{i_{\mu+1}, \dots, i_{\#\tau}\}$. The boxes are chosen as the minimal ones: $Q_{\tau_k} := Q_{\min}(X_{\tau_k})$ ($k = 1, 2$).

D.2.1.3 Cost of $T(I)$

The implementation requires a representation of the clusters $\tau \in T(I)$. Provided that $T(I)$ is a binary tree, the number of clusters is equal to

$$\#T(I) = 2\#\mathcal{L}(T(I)) - 1.$$

In the worst case, $\#\mathcal{L}(T(I)) = \#I$ holds. Since the expectation value of the size of $\tau \in \mathcal{L}(T(I))$ is $3n_{\min}/4$, the mean value is

$$\#T(I) \approx \frac{3}{2} \frac{\#I}{n_{\min}}.$$

The ordering of indices in I can be introduced so that all clusters are identified by a pair of integers $(\alpha_\tau, \beta_\tau)$, since $\tau = \{i_k \in I : \alpha_\tau \leq k \leq \beta_\tau\}$. Therefore, storage of $T(I)$ requires about $3\#I/n_{\min}$ integers. The ordering of I involves additional $\#I$ integers as labels of $i \in I$.

³ For a precise notation, one has to change the identifier of the vertices of the tree to distinguish the father τ from the son τ_1 (cf. [198, Remark 5.1 and §A.4]).

⁴ For simplicity we assume that the nodal points satisfy $\xi_i \neq \xi_j$ for $i \neq j$ (cf. [198, (5.21a,b)]).

D.2.2 Block Cluster Tree

Matrices in $\mathbb{K}^{I \times J}$ correspond to the index set $I \times J$. Correspondingly, we have to describe a block decomposition of $I \times J$. For this purpose, we do not need a new construction but obtain the tree $T(I \times J)$ —now called *block cluster tree*—directly from $T(I)$ and $T(J)$.

Definition D.8. A (level-conserving⁵) block cluster tree is constructed as follows.

(1) $I \times J$ is the root.

(2) The recursion starts with the block $b = \tau \times \sigma$ for $\tau = I$ and $\sigma = J$.

(2a) Define the set of sons of $b = \tau \times \sigma$ by

$$S(b) := \begin{cases} \emptyset & \text{if } S_{T(I)}(\tau) = \emptyset \text{ or } S_{T(J)}(\sigma) = \emptyset, \\ \{\tau' \times \sigma' : \tau' \in S_{T(I)}(\tau), \sigma' \in S_{T(J)}(\sigma)\} & \text{otherwise.} \end{cases}$$

(2b) Apply (2a,b) recursively to all sons of b , provided that $S(b) \neq \emptyset$.

Remark D.9. (a) The *level-conserving* property is described by the identity

$$\text{level}_{T(I \times J)}(b) = \text{level}_{T(I)}(\tau) = \text{level}_{T(J)}(\sigma) \text{ for } b = \tau \times \sigma \in T(I \times J).$$

(b) We have $\text{depth}(T(I \times J)) \leq \min\{\text{depth}(T(I)), \text{depth}(T(J))\}$ and

(c) $\tau \times \sigma \in \mathcal{L}(T(I \times J))$ if and only if $\min\{\#\tau, \#\sigma\} \leq n_{\min}$, implying $\text{rank}(M|_b) \leq n_{\min}$ holds for all $b \in \mathcal{L}(T(I \times J))$.

D.2.3 Partition

Let P be a set of blocks $b \subset I \times J$. We say that P is a (*block*) *partition* of $I \times J$ if all elements of P are disjoint and if $\bigcup_{b \in P} b = I \times J$ (cf. Fig. D.1). As mentioned before, we are not looking at *all* partitions but only at those contained in the block cluster tree: $P \subset T(I \times J)$.

Lemma D.10. There is a one-to-one correspondence between partitions $P \subset T(I \times J)$ and subtrees $T' \subset T(I \times J)$ with $\text{root}(T') = I \times J$. The tree corresponding to P is $T(I \times J; P) := \{b \in T(I \times J) : b' \subset b \text{ for some } b \in P\}$. The inverse mapping is given by $P = \mathcal{L}(T')$.

Proof. (i) Let $P \subset T(I \times J)$ be given. For any $b \in P$, omit the sons of b in $T(I \times J)$, i.e., replace $S(b)$ by the empty set. Hence, b is a leaf of the subtree $T' := T(I \times J; P)$ and $P = \mathcal{L}(T')$ holds.

(ii) For any subtree $T' \subset T(I \times J)$ with $\text{root}(T') = I \times J$, we verify by induction on the depth of T' that $\mathcal{L}(T') \subset T' \subset T(I \times J)$ is a set of disjoint blocks and that $\bigcup_{b \in \mathcal{L}(T')} b = I \times J$. Hence, $P := \mathcal{L}(T') \subset T(I \times J)$ is a partition of $I \times J$. \square

⁵ For more general trees see [198, §§5.5.2–5.5.3].

We can define an order relation between partitions. We say that P_1 is *coarser* than P_2 if for all $b'' \in P_2$ there is a $b' \in P_1$ with $b'' \subset b'$. P_1 is *finer* than P_2 if P_2 is coarser than P_1 .

Concerning the partition of a matrix, we have two contradicting requirements. Let us fix a certain local rank r of the matrix blocks. First, the blocks must be small enough to allow a good rank- r approximation. Since $\sum_{b \in P} \#b = \#(I \times J)$, a small size $\#b$ must be compensated by a large number $\#P$ of blocks; however, the larger $\#P$ the larger the required storage is.

D.2.4 Admissible Blocks

The study of the singularity functions of elliptic boundary value problems shows that an optimal partition should use blocks satisfying the following admissibility condition.

In §D.2.1.2, nodal points ξ_i and sets X_τ are introduced. Now, we replace $\xi_i \in \mathbb{R}^d$ with a subset $X_i \subset \mathbb{R}^d$. The computations may use $X_i := \{\xi_i\}$, but the precise analysis of finite element discretisations requires⁶ $X_i := \text{supp } \phi_i$, where ϕ_i denotes the finite element basis function in (E.7b). X_τ is re-defined by⁷

$$X_\tau := \bigcup_{i \in \tau} X_i \subset \mathbb{R}^d. \quad (\text{D.8})$$

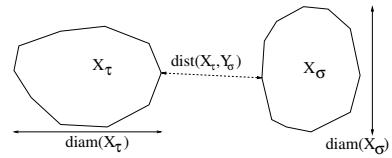


Fig. D.4 Supports X_τ and X_σ .

The corresponding sets related to a second index set J and clusters $\sigma \in T(J)$ are denoted⁸ by Y_i and Y_σ , respectively. This allows us to define a diameter of a cluster and the distance between two clusters (cf. Fig. D.4):

$$\text{diam}(\tau) := \max\{\|x' - x''\| : x', x'' \in X_\tau\}, \quad \tau \subset I, \quad (\text{D.9a})$$

$$\text{dist}(\tau, \sigma) := \min\{\|x - y\| : x \in X_\tau, y \in Y_\sigma\}, \quad \tau \subset I, \sigma \subset J. \quad (\text{D.9b})$$

Definition D.11 (η -admissibility of a block). Let $\eta > 0$. The clusters $\tau \subset I$ and $\sigma \subset J$ are associated with supports X_τ and X_σ . Then the block $b = \tau \times \sigma$ is called η -admissible⁹ if

$$\min\{\text{diam}(\tau), \text{diam}(\sigma)\} \leq \eta \text{dist}(\tau, \sigma). \quad (\text{D.10})$$

If η is a fixed value, we use the term *admissibility* of b without referring to η .

⁶ The support of a function $f : X \rightarrow Y$ is defined by $\text{supp}(f) := \overline{\{x : f(x) \neq 0\}}$.

⁷ The later analysis requires convex sets; i.e., the sets X_τ should be replaced by their convex hulls. Since, finally, we shall replace X_τ by its bounding box (cf. §D.2.5), convexity will be guaranteed.

⁸ Even if $I = J$, the Petrov–Galerkin discretisation of Definition E.7 may use different ansatz functions ϕ_i ($i \in I$) and test functions ψ_j ($j \in J$) so that $X_i \neq Y_i$.

⁹ For variants of the admissibility condition see [198, (5.7a–c) and §5.2.3].

We cannot require that all blocks $b \in P$ of a partition be admissible. Consider, e.g., a diagonal block $b = \tau \times \tau$. Then $\text{diam}(\tau)$ is positive, while $\text{dist}(\tau, \tau) = 0$, so that (D.10) cannot hold for $\eta > 0$. For such blocks, we cannot expect to find low-rank approximations. Instead we use the full representation of the matrix block $M|_b$, but we require that these blocks be small in the sense that $b = \tau \times \sigma$ with $\min\{\#\tau, \#\sigma\} \leq n_{\min}$. The latter condition implies that $b \in \mathcal{L}(T(I \times J))$. We combine both requirements in the following definition:

$$\text{adm}^*(b) := \begin{cases} \text{true} & \text{if (D.10) holds or if } b \in \mathcal{L}(T(I \times J)), \\ \text{false} & \text{otherwise.} \end{cases} \quad (\text{D.11})$$

A partition $P \subset T(I \times J)$ is called admissible if $\text{adm}^*(b)$ holds for all $b \in P$.

Now we want to find the *minimal* admissible partition $P \subset T(I \times J)$ which is the coarsest partition so that $\text{adm}^*(b)$ holds for all $b \in P$. The construction of the set P uses the equivalent formulation $P = \mathcal{L}(T')$ with $T' = T(I \times J; P)$ (cf. Lemma D.10). T' and P are the result of the call

$$T' := \emptyset; P := \emptyset; \text{MinAdmPart}(T', P, I \times J)$$

of the following recursion:

```
procedure MinAdmPart( $T', P, b$ );
begin  $T' := T' \cup \{b\}$ ;
  if  $\text{adm}^*(b)$  then  $P := P \cup \{b\}$  else for all  $b' \in S(b)$  do MinAdmPart( $T', P, b'$ )
end;
```

D.2.5 Use of Bounding Boxes for X_τ

The functions $\text{diam}(\tau)$ and $\text{dist}(\tau, \sigma)$ defined above are difficult to evaluate. In the general finite element case, X_τ is a union of triangles. The computational work for determining $\text{diam}(\tau)$ and $\text{dist}(\tau, \sigma)$ increases with the number of involved corner points. For application in practice, we introduce the bounding boxes of X_τ :

$$Q_\tau := Q_{\min}(X_\tau)$$

(cf. §D.2.1.2). Obviously, the inequalities

$$\text{diam}(Q_\tau) \geq \text{diam}(\tau), \quad \text{dist}(Q_\tau, Q_\sigma) \leq \text{dist}(\tau, \sigma)$$

hold, while evaluating $\text{diam}(Q_\tau)$ and $\text{dist}(Q_\tau, Q_\sigma)$ is trivial. Computing Q_τ from $Q_{\tau'}$ and $Q_{\tau''}$ for $\tau = \tau' \dot{\cup} \tau''$ is also simple. In (D.12) we replace $\text{adm}^*(b)$ with $\text{adm}_Q^*(b)$ which uses

$$\min\{\text{diam}(Q_\tau), \text{diam}(Q_\sigma)\} \leq \eta \text{dist}(Q_\tau, Q_\sigma)$$

instead of (D.10). One verifies the implication¹⁰ $\text{adm}_Q^*(b) \Rightarrow \text{adm}^*(b)$.

¹⁰ This statement shows that the partition might be finer than necessary. Therefore it makes sense to coarsen the partition. This technique checks whether a coarser partition with (almost) the same error bounds exists without increasing the storage size (cf. [198, §6.7.2]).

D.2.6 Set of Hierarchical Matrices

Let $P \subset T(I \times J)$ be an admissible partition. Then the *near-field* P^- and the *far-field* P^+ are defined by

$$P^- := \{b = \tau \times \sigma \in P : \min\{\#\tau, \#\sigma\} \leq n_{\min}\}, \quad P^+ := P \setminus P^-.$$

In Figure D.1 (right) the dark blocks belong to P^- .

Definition D.12 (hierarchical matrix). Let I and J be index sets, $T(I \times J)$ a block cluster tree, and P a partition. Furthermore, a local rank distribution is given by the function

$$r : P \rightarrow \mathbb{N}_0. \quad (\text{D.13})$$

Then the set $\mathcal{H}(r, P) \subset \mathbb{R}^{I \times J}$ of hierarchical matrices (with respect to partition P and rank distribution r) consists of all matrices $M \in \mathbb{R}^{I \times J}$ with

$$\text{rank}(M|_b) \leq r(b) \quad \text{for all } b \in P^+.$$

More precisely, $M|_b \in \mathcal{R}_{r(b)}$ (cf. Definition D.2) is required for all blocks $b \in P^+$; i.e., the factors A_b, B_b of the representation $M|_b = A_b B_b^\top$ be explicitly available. Matrix blocks $M|_b$ corresponding to the small blocks $b \in P^-$ are implemented as full matrices: $M|_b \in \mathcal{F}(b)$ (cf. Definition D.2).

Remark D.13. (a) The standard choice of the function (D.13) is a constant $r \in \mathbb{N}_0$. Then we say that the hierarchical matrix has the *local rank* r .

(b) A variable rank $r(b)$ is in particular needed for the adaptive choice of the local ranks (cf. Remark D.19).

Remark D.14. Assume that $M = M^\top \in \mathcal{H}(r, P) \subset \mathbb{R}^{I \times I}$ is a symmetric matrix, and that P is symmetric (i.e., $b = \tau \times \sigma \in P \Leftrightarrow \sigma \times \tau \in P$). Then the factors A_b and B_b of $M|_b = A_b B_b^\top$ and $M|_{b'} = (M|_b)^\top = B_b A_b^\top$ for $b = \tau \times \sigma \in P^+$ and $b' = \sigma \times \tau$ have to be stored only once. The same statement holds for $b \in P^-$.

D.2.7 \mathcal{H}^2 -Matrices

Without going into details we mention that the set of \mathcal{H}^2 -matrices, which satisfy additional conditions, leads to even less storage cost and less computational work of the matrix operations (cf. [198, §8], Börm [54], Börm–Reimer [55]).

D.2.8 Storage

In the following, we introduce the quantity C_{sp} , which is crucial for estimating the storage cost and the computational cost of the matrix operations discussed later. Let $T(I \times J)$ be the block cluster tree corresponding to $T(I), T(J)$, and let P be

the partition. For any $\sigma \in T(J)$, there should be only a few blocks $b = \tau \times \sigma \in P$ containing σ as a factor. The quantities

$$\begin{aligned} C_{\text{sp},l}(\tau, P) &:= \#\{\sigma \in T(J) : \tau \times \sigma \in P\} \quad \text{for } \tau \in T(I), \\ C_{\text{sp},r}(\sigma, P) &:= \#\{\tau \in T(I) : \tau \times \sigma \in P\} \quad \text{for } \sigma \in T(J) \end{aligned}$$

describe how often the clusters τ and σ appear as columns or rows in the blocks of the partition P . Define

$$C_{\text{sp}}(P) := \max \left\{ \max_{\tau \in T(I)} C_{\text{sp},l}(\tau, P), \max_{\sigma \in T(J)} C_{\text{sp},r}(\sigma, P) \right\}.$$

For instance, the format in the middle of Figure D.1 has the constant $C_{\text{sp}}(P) = 6$, while the simpler hierarchical format on the left side yields the sparsity constant $C_{\text{sp}}(P) = 2$. These constants are independent of the size of the matrices.

Proposition D.15. *Assume that the regular geometric bisection is used for generating $T(I)$. Then the finite element matrices for a sequence of grids with uniform shape regularity have a uniformly bounded C_{sp} .*

Proof. See [198, §6.4.3 and Theorem 6.24] or [161]. \square

Lemma D.16. *The number of blocks in partition P is bounded by*

$$\#P \leq (2 \min\{\#I, \#J\} - 1) C_{\text{sp}}(P).$$

Proof. We estimate by

$$\begin{aligned} \#P &= \sum_{\tau \times \sigma \in P} 1 = \sum_{\tau \in T(I)} \#\{\sigma \in T(J) : \tau \times \sigma \in P\} \leq \sum_{\tau \in T(I)} C_{\text{sp}}(P) \\ &\leq (2\#I - 1) C_{\text{sp}}(P). \end{aligned}$$

Interchanging the roles of τ and σ , we also obtain the bound $(2\#J - 1) C_{\text{sp}}(P)$. \square

Lemma D.17 (storage). *The storage cost $S_{\mathcal{H}}(r, P)$ of matrices in $\mathcal{H}(r, P)$ is bounded by*

$$C_{\text{sp}}(P) \cdot \max\{n_{\min}, r\} \cdot [(\text{depth}(T(I)) + 1) \#I + (\text{depth}(T(J)) + 1) \#J].$$

Proof. $S_{\mathcal{H}}(r, P)$ is the sum of the storage cost of all blocks $b = \tau \times \sigma \in P$:

$$S_{\mathcal{H}}(r, P) \underset{\text{Remark D.3}}{=} r \sum_{\tau \times \sigma \in P^+} (\#\tau + \#\sigma) + \sum_{\tau \times \sigma \in P^-} \#\tau \cdot \#\sigma.$$

Because of $\min\{\#\tau, \#\sigma\} \leq n_{\min}$ for $\tau \times \sigma \in P^-$, we have

$$\begin{aligned} \#\tau \#\sigma &= \min\{\#\tau, \#\sigma\} \max\{\#\tau, \#\sigma\} \\ &\leq \min\{\#\tau, \#\sigma\} (\#\tau + \#\sigma) \leq n_{\min} (\#\tau + \#\sigma), \end{aligned}$$

proving that

$$S_{\mathcal{H}}(r, P) \leq \max\{n_{\min}, r\} \sum_{\tau \times \sigma \in P} (\#\tau + \#\sigma).$$

The definition of $C_{\text{sp},l}(\tau, P)$ and $C_{\text{sp}}(P)$ yields

$$\begin{aligned} \sum_{\tau \times \sigma \in P} \#\tau &= \sum_{\tau \in T(I)} \left[\#\tau \sum_{\sigma: \tau \times \sigma \in P} 1 \right] = \sum_{\tau \in T(I)} \#\tau C_{\text{sp},l}(\tau, P) \\ &\leq \sum_{\ell=0}^{\text{depth}(T(I))} C_{\text{sp}}(P) \sum_{\tau \in T^{(\ell)}(I)} \#\tau \leq C_{\text{sp}}(P) (\text{depth}(T(I)) + 1) \#I \end{aligned}$$

with $T^{(\ell)}(I)$ defined in (D.7). Combining this estimate with the similar inequality $\sum_{\tau \times \sigma \in P} \#\sigma \leq C_{\text{sp}}(P) (\text{depth}(T(J)) + 1) \#J$ proves the desired bound. \square

D.2.9 Accuracy

D.2.9.1 Discretisation of an Integral Operator

We consider a typical example. The integral operator \mathcal{K} is defined by

$$(\mathcal{K}f)(x) := \int_0^1 \sigma(x, y) f(y) dy \quad (\text{cf. } \S 11.9.1).$$

For discretisation, we introduce an equidistant grid $x_\nu = \nu h$ ($h = 1/N$) with collocation points $\xi_\nu = (\nu - \frac{1}{2})h$. The piecewise constant basis functions $b_\nu(x) = \begin{cases} 1 & x \in [x_{\nu-1}, x_\nu] \\ 0 & \text{otherwise} \end{cases}$ are used for the ansatz $f = \sum_{j=1}^N a_j b_j$. Collocation¹¹ of $\mathcal{K}f = g$ yields the equations

$$\left(\mathcal{K} \left(\sum_{j=1}^N a_j b_j \right) \right) (\xi_i) = g(\xi_i) \quad (1 \leq i, j \leq N).$$

The sets X_i and Y_j in (D.8) are $X_i = \{\xi_i\}$ and $Y_j = \text{supp}(b_j) = [x_{j-1}, x_j]$. Let $\tau \times \sigma \in P^+$ be an admissible block. This means that $X_\tau = [a, b]$ and $Y_\sigma = [c, d]$ satisfy (D.10). The entries of the discretisation matrix K are

$$K_{ij} = \int_0^1 \sigma(\xi_i, y) b_j(y) dy = \int_{x_{j-1}}^{x_j} \sigma(\xi_i, y) dy.$$

Assume that σ has a separable approximation of separation rank r :

$$\sigma(x, y) = \sigma_r(x, y) + R_r \quad \text{with} \quad \sigma_r(x, y) = \sum_{\ell=1}^r \varphi_\ell(x) \psi_\ell(y). \quad (\text{D.14})$$

¹¹ The choice of the discretisation is not essential.

Replacing σ with σ_r , we obtain $K_{ij}^{(r)}$ instead of K_{ij} :

$$K_{ij}^{(r)} = \int_{x_{j-1}}^{x_j} \sigma_r(\xi_i, y) dy = \sum_{\ell=1}^r \underbrace{\varphi_\ell(\xi_i)}_{=:a_i^{(\ell)}} \underbrace{\int_{x_{j-1}}^{x_j} \psi_\ell(y) dy}_{=:b_j^{(\ell)}}$$

The vectors $a^{(\ell)} = (a_i^{(\ell)})_{i \in \tau}$ and $b^{(\ell)} = (b_j^{(\ell)})_{j \in \sigma}$ form the rank- r matrix

$$(K_{ij}^{(r)})_{(i,j) \in \tau \times \sigma} = K^{(r)}|_{\tau \times \sigma} = \sum_{\ell=1}^r a^{(\ell)} b^{(\ell)\top}.$$

The general integral equation method uses the singularity function $\sigma(x, y)$ of an elliptic differential equation with constant coefficients or its derivatives as kernel function of the integral operator. The integral is taken over a surface. This does not change the fact that a separable approximation of separation rank r leads to a matrix block $K^{(r)}|_{\tau \times \sigma} \in \mathcal{R}_r$.

D.2.9.2 Error Estimate

Let $\sigma(x, y)$ be the singularity function of an elliptic differential equation with constant coefficients. Then $\sigma(x, y)$ is not only analytic for $(x, y) \in X \times Y$, $x \neq y$, but also *asymptotically smooth*; i.e., the partial derivatives satisfy the estimate

$$\left| \frac{\partial^\alpha}{\partial x^\alpha} \frac{\partial^\beta}{\partial x^\beta} \sigma(x, y) \right| \leq C_{\alpha+\beta} |x - y|^{-|\alpha|-|\beta|-s}$$

for all multi-indices $\alpha, \beta \in \mathbb{N}_0^d$, $\alpha + \beta \neq 0$ and for all $x \in X$, $y \in Y$, $x \neq y$. The fixed value $s \in \mathbb{R}$ indicates the strength of the singularity (cf. [198, §E]).

A typical example of an asymptotically smooth function is the function

$$\sigma(x, y) = \log|x - y| \quad \text{for } x, y \in \mathbb{R} (x \neq y).$$

A simple method producing an separable expression is the Taylor¹² expansion. Assume that $x \in X := [a, b]$ and $y \in Y := [c, d]$ with $0 \leq a \leq b < c \leq d \leq 1$. Expansion around $x_0 := \frac{a+b}{2}$ yields

$$\log|x - y| = \sum_{\ell=0}^{r-1} \frac{(x - x_0)^\ell}{\ell!} \frac{d^\ell}{dx^\ell} \log|x_0 - y| + R_r.$$

This is (D.14) with $\varphi_\ell(x) = \frac{(x - x_0)^{\ell-1}}{(\ell-1)!}$ and $\psi_\ell(y) = \frac{d^{\ell-1}}{dx^{\ell-1}} \log|x_0 - y|$. Remainder R_r can be estimated by

$$|R_r| = \left| \sum_{\ell=r}^{\infty} \frac{1}{\ell} \left(\frac{x - x_0}{y - x_0} \right)^\ell \right| \leq \frac{1}{r} \frac{\left| \frac{x - x_0}{y - x_0} \right|^r}{1 - \left| \frac{x - x_0}{y - x_0} \right|} \quad \text{for } x \in [a, b], y \in [c, d].$$

¹² In practice, interpolation is preferred. It is easier to implement and yields better approximations.

Obviously, the error $|K_{ij}^{(r)} - K_{ij}|$ is described by the remainder R_r . Since $X = [a, b]$ and $Y = [c, d]$ are assumed to satisfy η -admissibility (D.10), we have

$$\min\{\text{diam}(\tau), \text{diam}(\sigma)\} = b - a \leq \eta(c - b),$$

provided that $\text{diam}(X) = b - a \leq \text{diam}(Y) = d - c$ (otherwise interchange the roles of x and y). Since $|x - x_0| \leq \frac{b-a}{2}$ and $|y - x_0| \geq \frac{b-a}{2} + c - b$, we obtain the estimates

$$\left| \frac{x - x_0}{y - x_0} \right| \leq \frac{\frac{b-a}{2}}{\frac{b-a}{2} + c - b} = \frac{1}{1 + 2\frac{c-b}{b-a}} \leq \frac{1}{1 + 2/\eta}$$

and

$$|R_r| \leq \frac{1}{r} \frac{\eta + 2}{2} \left(\frac{1}{1 + 2/\eta} \right)^r.$$

This proves that the error R_r decays exponentially with respect to r .

D.2.9.3 Separable Expansion of the Green Function

In the case of an integral operator, the kernel function $\sigma(x, y)$ is explicitly given. In the case of the inversion of a sparse finite element matrix, the inverse is implicitly connected with an integral operator using the Green function $G(x, y)$ instead of σ . Even if the coefficients of the differential operator in $\Omega \subset \mathbb{R}^d$ are nonsmooth (only bounded), one can show that $G(x, y)$ has a separable expansion (D.14) with remainder $\mathcal{O}(R_r) \leq \mathcal{O}(\exp(-c r^{1/(d+1)}))$ (cf. Bebendorf–Hackbusch [39], Faustmann [128], Faustmann–Melenk–Praetorius [129], and [198, §11.3]).

D.3 Matrix Operations

D.3.1 Matrix-Vector Multiplication

Let $M \in \mathcal{H}(r, P)$, $x \in \mathbb{R}^J$, and $y \in \mathbb{R}^I$. Calling $MVM(y, M, x, I \times J)$ produces $y := y + Mx$. It is a recursion in $T(I \times J, P)$:

```
procedure MVM( $y, M, x, b$ );
  if  $b = \tau \times \sigma \in P$  then  $y|_\tau := y|_\tau + M|_b \cdot x|_\sigma$ 
  else for all  $b' \in S(b)$  do MVM( $y, M, x, b'$ );
```

If $b \in P^-$, $M|_b$ is represented as a full matrix and $M|_b \cdot x|_\sigma$ in the second line of *MVM* is the standard matrix-vector multiplication. If $b \in P^+$, $M|_b \in \mathcal{R}_{r(b)}$ holds and the product $M|_b \cdot x|_\sigma$ is performed as in (D.3).

Lemma D.18. *The number N_{MV} of arithmetic operations for matrix-vector multiplication involving a matrix from $\mathcal{H}(r, P)$ can be bounded by the storage cost $S_{\mathcal{H}}(r, P)$ (estimated in Lemma D.17):*

$$S_{\mathcal{H}}(r, P) \leq N_{\text{MV}} \leq 2S_{\mathcal{H}}(r, P).$$

Proof. See [198, Lemma 7.17]. □

D.3.2 Truncations

Truncation to lower rank is an essential part of the following operations. The general notation of the (nonlinear) truncation operator to rank r is \mathcal{T}_r . An additional upper index indicates the type of matrices to which the operator is applied:

$$\mathcal{T}_r^{\mathcal{R} \leftarrow \mathcal{R}} : \mathcal{R}_s \rightarrow \mathcal{R}_r \quad (s > r), \quad \mathcal{T}_r^{\mathcal{R} \leftarrow \mathcal{F}} : \mathcal{F} \rightarrow \mathcal{R}_r, \quad \mathcal{T}_r^{\mathcal{R}} : \mathcal{R}_s \cup \mathcal{F} \rightarrow \mathcal{R}_r.$$

$\mathcal{T}_r^{\mathcal{R} \leftarrow \mathcal{R}}$ is explained in Lemma D.4. $\mathcal{T}_r^{\mathcal{R} \leftarrow \mathcal{F}}$ using the singular value decomposition is only applied to small-sized fully populated matrices. $\mathcal{T}_r^{\mathcal{R} \leftarrow \mathcal{R}}$ is defined as the identity if $s \leq r$.

Remark D.19. Instead of the target rank r we can fix an accuracy of $\varepsilon > 0$ and choose r such that the error is bounded by ε .

The truncation $\mathcal{T}_r^{\mathcal{H}} : \mathcal{H}(s, P) \rightarrow \mathcal{H}(r, P)$ ($s \not\geq r$) of hierarchical matrices is defined by a blockwise truncation:

$$(\mathcal{T}_{r \leftarrow s}^{\mathcal{H}}(M))|_b = \begin{cases} \mathcal{T}_{r(b) \leftarrow s(b)}^{\mathcal{R}}(M|_b) & \text{if } b \in P^+ \\ M|_b & \text{if } b \in P^- \end{cases} \quad \text{for } M \in \mathcal{H}(s, P).$$

D.3.3 Addition

Let $M_1 \in \mathcal{H}(r_1, P)$ and $M_2 \in \mathcal{H}(r_2, P)$ be two hierarchical matrices with the same partition P . The *exact* addition yields $M = M_1 + M_2 \in \mathcal{H}(r_1 + r_2, P)$. For its computation, we have to add all blocks: $M|_b := M_1|_b + M_2|_b$ for $b \in P$. Computational work is only involved for $b \in P^-$, since addition of \mathcal{R}_r matrices is free (cf. (D.4)).

Formatted addition $\oplus_r : \mathcal{H}(r_1, P) \times \mathcal{H}(r_2, P) \rightarrow \mathcal{H}(r, P)$ uses truncation $M|_b := \mathcal{T}_r^{\mathcal{R} \leftarrow \mathcal{R}}(M_1|_b + M_2|_b)$ for all $b \in P^+$.

Lemma D.20. *In the standard case of $r = r_1 = r_2$, the cost is*

$$N_{H+H} \leq 24rS_{\mathcal{H}}(r, P) + 176r^3 \#P^+.$$

Proof. Use Lemma D.4 with $s = 2r$ for $b \in P^+$ and sum $\#b$ over all $b \in P^-$; cf. [198, Lemma 7.20b]. \square

In the case of multiple additions, the cost of $\mathcal{R}_r(\sum_{\nu=1}^q M_{\nu})$ increases with the rank of $\sum M_{\nu}$ and therefore with the number of terms. Instead, the cheaper but possibly less accurate *pairwise truncation* is preferred:

$$\mathcal{T}_{r, \text{pairw}}^{\mathcal{R}} \left(\sum_{\nu=1}^q M_{\nu} \right) = \mathcal{T}_r^{\mathcal{R}} \left(\dots \mathcal{T}_r^{\mathcal{R}} \underbrace{\left(\mathcal{T}_r^{\mathcal{R}} \left(M_1 + M_2 \right) + M_3 \right) + M_4}_{\dots} + \dots \right).$$

D.3.4 Agglomeration

Next, we consider the conversion of a block-structured matrix into a (global) rank- r matrix; e.g., $\boxed{} \mapsto \square$. Let b be the block of the output matrix, whereas the input matrix is split into the blocks $b_i \in S(b)$: $\begin{bmatrix} M_1 & M_2 \\ M_3 & M_4 \end{bmatrix} \in \mathbb{R}^b$ with $M_i \in \mathcal{R}(s, b_i)$. The extension of a matrix $M \in \mathbb{K}^{b'}$ to a larger size \mathbb{R}^b is denoted by the symbol $\cdot|b$:

$$(M|b)_{i,j} = \begin{cases} M_{i,j} & \text{if } (i,j) \in b', \\ 0 & \text{if } (i,j) \in b \setminus b'. \end{cases}$$

Therefore the agglomeration can be written as a summation: $M = \begin{bmatrix} M_1 & M_2 \\ M_3 & M_4 \end{bmatrix} = M_1|b + M_2|b + M_3|b + M_4|b$. Accordingly, truncation $\mathcal{T}_r^{\mathcal{R}}$ (or $\mathcal{T}_{r,\text{pairw}}^{\mathcal{R}}$) can be applied. The truncation error is analysed in Hackbusch [200].

D.3.5 Matrix-Matrix Multiplication

We consider $M = M'M''$ with $M' \in \mathbb{K}^{I \times J}$ and $M'' \in \mathbb{K}^{J \times K}$. The involved block cluster trees $T(I \times J)$ and $T(J \times K)$ share the common structure of $T(J)$. This fact is essential since otherwise the block structures of M' and M'' would not fit together.

The basic idea of the multiplication algorithm is as follows. Using the substructuring $M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}$ of hierarchical matrices, we can divide the multiplication task $M = M'M'' \in \mathcal{H}(r, P)$ with $M' \in \mathcal{H}(r', P')$ and $M'' \in \mathcal{H}(r'', P'')$ in four subtasks:

$$M_{11} = M'_{11}M''_{11} + M'_{12}M''_{21}, \quad M_{12} = \dots, \quad \text{etc.}$$

The terms are of the form $\hat{M} = \hat{M}'\hat{M}''$. It may happen that one of the submatrices \hat{M}' or \hat{M}'' is not substructured but belongs to \mathcal{R}_r or \mathcal{F} . Then, as discussed below, the product $\hat{M} = \hat{M}'\hat{M}''$ can be evaluated. Otherwise, we have to subdivide recursively until one of the following cases applies.

Case 1a) $\hat{M}'' = M''|_b$ for $b \in P^+$, i.e., $\hat{M}'' = \sum a_i b_i^T \in \mathcal{R}_{r(b)}$. Now the multiplication $\hat{M}'\hat{M}''$ reduces to r matrix-vector multiplications $\hat{M}'a_i$.

Case 1b) $\hat{M}'' = M''|_b$ for $b \in P^-$, i.e., $\hat{M}'' \in \mathcal{F}$. Same as Case 1a with vectors a_i being the columns of \hat{M}'' .

Case 2) $\hat{M}' = M'|_b$ for $b \in P$. $\hat{M}^T = \hat{M}''^T \hat{M}'^T$ can be treated as before.

Case 3) The target matrix M contains the block $\hat{M} = M|_b$ for $b = \tau \times \rho \in P$. The other matrices are $\hat{M}' \in \mathbb{K}^{\tau \times \sigma}$ and $\hat{M}'' \in \mathbb{K}^{\sigma \times \rho}$ for some $\sigma \in T(J)$. Then the operation $M|_{\tau \times \rho} \leftarrow M|_{\tau \times \rho} \oplus_r (M'|_{\tau \times \sigma} \odot_r M''|_{\sigma \times \rho})$ is performed by the following procedure. Here, \odot_r indicates that the product is truncated to rank r .

```

procedure MMR( $M, M', M'', \tau, \sigma, \rho$ );
begin
  if  $\tau \times \sigma \in P'$  or  $\sigma \times \rho \in P''$  then
    begin  $Z := M'|_{\tau \times \sigma} M''|_{\sigma \times \rho}$ ;
      if  $\tau \times \rho \subset b \in P^+$  then  $Z := \mathcal{T}_r^R(Z)$  {for a suitable  $b \in T$ }
      end else {the else case corresponds to  $\tau \times \sigma \notin P'$  and  $\sigma \times \rho \notin P''$ }
      begin  $Z|_{\tau \times \rho} := 0$ ;
        for all  $\tau' \in S(\tau), \sigma' \in S(\sigma), \rho' \in S(\rho)$ 
        do MMR( $Z, M', M'', \tau', \sigma', \rho'$ ) {recursion}
      end;
      if  $\tau \times \rho \in P^-$  then  $M|_{\tau \times \rho} := M|_{\tau \times \rho} + Z$  else  $M|_{\tau \times \rho} := \mathcal{T}_r^R(M|_{\tau \times \rho} + Z)$ 
    end;
  end;

```

The general formatted product is written in the form $M := M \oplus_r (M' \odot_r M'')$. The call $MM(M, M', M'', I, J, K)$ of the following procedure produces $M := M \oplus_r (M' \odot_r M'')$. The factors M', M'' are input parameters, whereas M is input and output parameter. The parameters τ, σ, ρ must satisfy $\tau \times \sigma \in T(I \times J, P')$, $\sigma \times \rho \in T(J \times K, P'')$, and $\tau \times \rho \in T(I \times K, P)$.

```

procedure MM( $M, M', M'', \tau, \sigma, \rho$ );
  if  $\tau \times \sigma \notin P'$  and  $\sigma \times \rho \notin P''$  and  $\tau \times \rho \notin P$  then
    for all  $\tau' \in S_{T(I)}(\tau), \sigma' \in S_{T(J)}(\sigma), \rho' \in S_{T(K)}(\rho)$  do
      MM( $M, M', M'', \tau', \sigma', \rho'$ )
  else if  $\tau \times \rho \notin P$  then { $\tau \times \sigma \in P'$  or  $\sigma \times \rho \in P''$  hold}
    begin  $Z := M'|_{\tau \times \sigma} M''|_{\sigma \times \rho}; M|_{\tau \times \rho} := \mathcal{T}_r^H(M|_{\tau \times \rho} + Z)$ 
  end else MMR( $M, M', M'', \tau, \sigma, \rho$ ); { $\tau \times \rho \in P$ }

```

Analysing the computational cost is more involved (details in [198, §7.8.3]). For discussing the asymptotic behaviour, we consider the case $I = J = K$ with $n := \#I$ and assume $\text{depth}(T(I \times I, P)) = \mathcal{O}(\log n)$ and $\#P = \mathcal{O}(n)$ (cf. Lemma D.16). Then we obtain (D.15), where $r := \max\{r', r'', n_{\min}\}$:

$$N_{\text{MM}}(P, r', r'') \leq \mathcal{O}(rn \log(n) (\log(n) + r^2)). \quad (\text{D.15})$$

D.3.6 Inversion and LU Decomposition

The representation (D.6) of the inverse yields a recursion. The involved additions (subtractions) and multiplications are replaced by formatted additions \oplus_r and formatted multiplications \odot_r . Details are given in [198, §7.5 and §7.8.4].

For many purposes, the inversion can be replaced by the LU decomposition which is described in detail in §13.1.

Appendix E

Galerkin Discretisation of Elliptic PDEs

Abstract A standard source of sparse and large linear systems is discretisation of elliptic boundary value problems. Here we consider Galerkin discretisation. The variational formulation of the problem is introduced in Section E.1, while the Galerkin approach follows in Section E.2. Some details about the finite element matrix are mentioned in Section E.3. Section E.4 describes the connections between the continuous differential operator and the discrete problem. The error estimates are discussed in Section E.5. Two discretisations corresponding to Galerkin subspaces $V_{n'} \subset V_n$ lead to characteristic connections between the corresponding matrices (cf. Section E.6).

Details about finite elements can be found, e.g., in Braess [63], Brenner–Scott [82], and Hackbusch [193, 201].

E.1 Variational Formulation of Boundary Value Problems

In the following, U and V are Hilbert spaces. The respective scalar products are denoted by $(\cdot, \cdot)_U$ and $(\cdot, \cdot)_V$. The norms are induced by the scalar products:

$$\|u\|_U = \sqrt{(u, u)_U} \quad \text{and} \quad \|v\|_V = \sqrt{(v, v)_V}.$$

The *dual space* V' is the space of all linear and continuous maps $f : V \rightarrow \mathbb{K}$ (\mathbb{K} denotes the field \mathbb{R} or \mathbb{C}) with the dual norm

$$\|f\|_{V'} = \sup\{|f(v)| : \|v\|_V \leq 1\}.$$

Similarly, the dual space U' of U can be introduced. By the Riesz isomorphism (cf. Riesz [326, §II.30]), any $f \in U'$ corresponds to an element $u_f \in U$ with

$$\|u_f\|_U = \|f\|_{U'} \quad \text{and} \quad (u_f, v)_U = f(v) \quad \text{for all } v \in U.$$

In the case of U , we identify f and u_f so that $U = U'$. However, V and V' are regarded as different spaces.

In the following application we consider the *Gelfand triple*

$$V \subset U = U' \subset V' \quad (\text{continuous and dense embeddings}).$$

The embedding is continuous if $\sup\{\|v\|_U/\|v\|_V : v \in V\} < \infty$. It is dense if V is dense in U . If $V \subset U$ is a continuous and dense embedding, $U \subset V'$ is as well.

In the case of standard scalar differential equation of order $2m$, the space V is the Sobolev space $H^m(\Omega)$ or a subspace as, e.g., $H_0^m(\Omega)$ with vanishing Dirichlet boundary values, while $U = L_2(\Omega)$ (explanation of these spaces in [193, §6.2]).

Finite element discretisations are based on the variational formulation of the boundary value problem, involving a bilinear or sesquilinear form.

Definition E.1. $a(\cdot, \cdot) : V \times V \rightarrow \mathbb{K}$ is a *sesquilinear form* if $\mathbb{K} = \mathbb{C}$ and

$$a(u + \lambda v, w) = a(u, w) + \lambda a(v, w), \quad a(u, v + \lambda w) = a(u, v) + \bar{\lambda} a(u, v + \lambda w)$$

for all $u, v, w \in V$, $\lambda \in \mathbb{K}$. In the case of $\mathbb{K} = \mathbb{R}$, the form $a(\cdot, \cdot)$ is called *bilinear*. The form a is called *symmetric* if $a(u, v) = \overline{a(v, u)}$ for all $u, v \in V$.

For simplicity, we restrict the explanations to $\mathbb{K} = \mathbb{R}$. The bilinear form is assumed to be bounded, i.e.,

$$\|a\| := \sup\{|a(v, w)| : v, w \in V, \|v\|_V \leq 1, \|w\|_V \leq 1\} \quad (\text{E.1})$$

is finite.

Starting from the strong formulation $Lv - f = 0$, multiplying by a *test function* $w \in V$, and applying partial integration, we arrive at the *weak formulation* or *variational formulation*

$$\text{find } v \in V \text{ with } a(v, w) = f(w) \quad \text{for all } w \in V \quad (\text{E.2})$$

(cf. [193, §7]). The Poisson model problem corresponds to (E.2) with

$$V = H_0^1(\Omega), \quad U = L_2(\Omega), \quad a(v, w) = \int_{\Omega} \langle \nabla v, \nabla w \rangle \, dx, \quad f(w) = \int_{\Omega} fw \, dx.$$

Concerning the solvability of problem (E.2), we refer to [193, §6.5]. A simple sufficient condition is coercivity of the form a .

Definition E.2. A bilinear (sesquilinear) form is called *coercive* if there is some $C > 0$ with

$$a(v, v) \geq \frac{1}{C} \|v\|_V^2 \quad \text{for all } v \in V. \quad (\text{E.3})$$

Together with $\|a\| < \infty$ (cf. (E.1)), this condition implies that the energy norm

$$\|v\|_a := \sqrt{a(v, v)}$$

is equivalent to $\|\cdot\|_V$.

Remark E.3. If $a(\cdot, \cdot)$ is symmetric and coercive, the variation problem (E.2) is equivalent to finding the minimiser of $\min\{\frac{1}{2}a(v, v) - f(v); v \in V\}$.

A bounded bilinear form $a(\cdot, \cdot) : V \times V \rightarrow \mathbb{R}$ corresponds to a unique linear operator

$$\begin{aligned} A : V &\rightarrow V' \quad \text{with} \\ a(u, v) &= \langle Av, w \rangle_{V' \times V} = (Au, v)_U \quad \text{for all } u, v \in V. \end{aligned} \tag{E.4}$$

Here, $\langle \varphi, v \rangle_{V' \times V} := \varphi(v)$ denotes the application of the dual map $\varphi \in V'$ to $v \in V$. The scalar product $(\cdot, \cdot)_U : U \times U \rightarrow \mathbb{R}$ restricted to $U \times V$ can be extended continuously to $(\cdot, v)_U = \langle \cdot, v \rangle_{V' \times V}$. Therefore, $(Au, v)_U$ makes sense. The operator A is the weak counterpart of the differential operator. Note that A also contains the boundary condition. Problem (E.2) is equivalent to

$$Au = f.$$

Remark E.4. The norms $\|a\|$ in (E.1) and $\|A\|_{V' \leftarrow V}$ (cf. (B.11)) coincide. Problem (E.2) is solvable for all $f \in V'$ if and only if A is invertible. In the latter case, the solution is bounded by

$$\|v\|_V \leq \|A^{-1}\|_{V \leftarrow V'} \|f\|_{V'}.$$

In the case of (E.3), $\|A^{-1}\|_{V \leftarrow V'} \leq C$ holds.

E.2 Galerkin Discretisation

For discretisation, we introduce a subspace¹ $V_n \subset V$ with $\dim(V_n) = n < \infty$. The norm on V_n is $\|\cdot\|_V$ restricted to V_n . The variational formulation (E.2) can be repeated with V_n instead of V :

$$\text{find } u_n \in V_n \text{ with } a(v, w) = (f, w)_U \quad \text{for all } w \in V_n. \tag{E.5}$$

For a concrete description of the Galerkin solution u_n , a basis $\{\phi_\alpha : \alpha \in I\}$ of V_n must be chosen, where $\#I = n = \dim V_n$. Any function $v_n \in V_n$ has a basis representation

$$P_n x := \sum_{\alpha \in I} x_\alpha \phi_\alpha = v_n \in V_n \tag{E.6}$$

involving a vector

$$x = (x_\alpha)_{\alpha \in I} \in X := \mathbb{R}^I$$

¹ The inclusion $V_n \subset V$ is characteristic for *conforming* Galerkin discretisations. Nonconforming discretisations need an extra analysis. Concerning multigrid applications to nonconforming discretisations, see Brenner [81] and Braess–Dryja–Hackbusch [64].

called the *coefficient vector* of v_n . Equation (E.6) defines the injective linear map $P_n : X \rightarrow V_n \subset V$. It can be interpreted as an invertible mapping $P_n : X \rightarrow V_n$ or as a mapping $P_n : X \rightarrow V$.

We make the ansatz $u_n = P_n x$ for the solution of (E.5). One easily verifies that it is sufficient to test the equation $a(v, w) = (f, w)_U$ with all basis functions $w = \phi_\alpha$, $\alpha \in I$. This leads to the following system of equations:

$$A_n x = f_n \quad \text{with entries} \tag{E.7a}$$

$$A_{n,\alpha\beta} = a(\phi_\beta, \phi_\alpha), \quad f_{n,\alpha} = (f, \phi_\alpha)_U \quad (\alpha, \beta \in I). \tag{E.7b}$$

A standard finite element basis $\{\phi_\alpha\}$ has the property that for any $\alpha \in I$ there are only a few $\beta \in I$ so that the interiors of the supports of ϕ_α and ϕ_β overlap. In this case, Galerkin discretisation is a *finite element discretisation* and the matrix A_n is sparse.

The finite element method uses a tessellation of Ω (e.g., a triangulation) into the *finite elements* (e.g., triangles), together with a set of *nodal points* $\{x_\alpha : \alpha \in I\}$. The standard basis functions are piecewise polynomials (e.g., piecewise linear functions) with the property

$$\phi_\alpha(x_\beta) = \delta_{\alpha\beta} \quad \text{for } \alpha, \beta \in I \quad (\text{cf. (1.11)}).$$

The following setting becomes easier if ϕ_α is scaled by a factor of $h^{-\frac{d}{2}}$ so that²

$$\|\phi_\alpha\|_{L_2(\Omega)} = \mathcal{O}(1) \tag{E.8}$$

holds with respect to the limit $h \rightarrow 0$, where h is the maximum of the diameters of the finite elements. This scaling leads to a matrix of order $\mathcal{O}(h^{-2})$ for second order differential equations as in (1.8). Another consequence is that both the matrix and the differential operator have a smallest eigenvalue of order $\mathcal{O}(1)$.

Exercise E.5. Prove that (E.7b) implies that $\langle A_h x, y \rangle = a(P_h x, P_h y)$ for all $x, y \in \mathbb{K}^I$, where $\langle \cdot, \cdot \rangle$ is the Euclidean scalar product in \mathbb{K}^I .

Several properties of the bilinear form a are inherited by the finite element matrix A_h .

Exercise E.6. Prove: (a) If $a(\cdot, \cdot)$ is symmetric, then A_h is Hermitian.

(b) If $a(\cdot, \cdot)$ is *positive*, i.e., $a(v, v) > 0$ for all $0 \neq v \in V$, then the Hermitian part $\frac{1}{2}(A_h + A_h^\top)$ is positive definite. Note that coercivity of a implies that a is positive.

(c) If $a(\cdot, \cdot)$ is positive and symmetric, then A_h is positive definite.

Unfortunately, in general, the invertibility of A_h is not directly connected with the invertibility of the operator A associated with a . A_h may be regular, although A is singular, and vice versa.

² Concerning the optimality of this choice, see Theorem 7.51.

Definition E.7. A generalisation of the Galerkin discretisation is the Petrov–Galerkin method which uses a test space W_n different from V_n :

$$\text{find } u_n \in V_n \text{ with } a(v, w) = (f, w)_U \quad \text{for all } w \in W_n.$$

Here $\dim(W_n) = \dim(V_n)$ is required. Statements about regularity and stability of the discrete problem are even more difficult than for the Galerkin method.

E.3 Subdomain Problems and Finite Element Matrix

The finite element matrix is defined in (E.7a,b) as a matrix with entries $a(\phi_\beta, \phi_\alpha)$ involving the basis functions ϕ_α . The expression $a(\phi_\beta, \phi_\alpha)$ is an integral over the intersection of the supports of ϕ_α and ϕ_β . The intersection may contain more than one geometric element (triangle, etc.). Therefore, the usual computation of $a(\phi_\beta, \phi_\alpha)$ is split into $a_\nu(\phi_\beta, \phi_\alpha)$, where a_ν involves the integral over one element $\Delta_\nu \in \mathcal{T}$ of the triangulation \mathcal{T} . Let J be the index set in $\mathcal{T} = \{\Delta_\nu : \nu \in J\}$. Define the quantities $b_{\alpha\beta}^{(\nu)} := a_\nu(\phi_\beta, \phi_\alpha)$ for $\alpha, \beta \in I$ and $\nu \in J$ and the corresponding matrices $B^{(\nu)} = (b_{\alpha\beta}^{(\nu)})_{\alpha, \beta \in I}$. Then the finite element matrix is

$$A = \sum_{\nu \in J} B^{(\nu)}.$$

Because of sparsity, $\sum_{\nu \in J}$ contains only $\mathcal{O}(1)$ nonzero terms.

Now we consider a subset $\omega \subsetneq \Omega$. The consistency with \mathcal{T} is expressed by the condition

$$\bar{\omega} = \bigcup_{\nu \in J_0} \overline{\Delta_\nu} \quad \text{for some } J_0 \subsetneq J.$$

Assume that the bilinear form is an integral $a(v, w) = \int_\Omega \dots dx$ over Ω . Replacing Ω by ω , we define

$$a_\omega(v, w) = \int_\omega \dots dx$$

with the same integrand. The variational formulation (E.2) with a_ω describes the boundary value problem in ω with natural boundary conditions on $\partial\omega \setminus \partial\Omega$ and the original boundary conditions on $\partial\omega \cap \partial\Omega$ (cf. Hackbusch [193, §7.5]).

Remark E.8. (a) The finite element matrix A_ω corresponding to $a_\omega(\phi_\beta, \phi_\alpha)$ can easily be computed from $B^{(\nu)}$ by $A_\omega = \sum_{\nu \in J_0} B^{(\nu)}$ involving $J_0 \subsetneq J$.

(b) Neither A_ω for $\omega \subsetneq \Omega$ nor $B^{(\nu)}$ can be obtained from A .

(c) If ω_ℓ ($1 \leq \ell \leq L$) are disjoint with $\overline{\Omega} = \bigcup_\ell \overline{\omega_\ell}$, we have $a(\cdot, \cdot) = \sum_\ell a_{\omega_\ell}(\cdot, \cdot)$.

E.4 Relations Between the Continuous and Discrete Problems

The map $P_n : X \rightarrow V_n \subset V$ introduced in (E.6) is called a *prolongation*. It is the first connection between the n -dimensional space $X = \mathbb{R}^I$ and the infinite-dimensional function space V . X becomes a Hilbert space when it is endowed with the Euclidean norm:

$$\langle x, y \rangle_X = \sum_{\alpha \in I} x_\alpha y_\alpha, \quad \|x\|_X = \sqrt{\sum_{\alpha \in I} |x_\alpha|^2}.$$

Because of the scaling (E.8), the constants \underline{C}_P and \bar{C}_P in

$$\underline{C}_P^{-1} \|x\|_X \leq \|P_n x\|_U \leq \bar{C}_P \|x\|_X \quad \text{for all } x \in X \quad (\text{E.9})$$

can be expected to be independent of $\dim(V_n)$. Inequality (E.9) is justified by Proposition E.13.

The prolongation $P_n : X \rightarrow V$ has the adjoint mapping $R_n = P_n^* : V' \rightarrow X'$ which we call the *restriction*. The definition $(P_n x, v)_U = \langle x, R_n v \rangle_X$ with $x = \mathbf{e}_\alpha$ (α -th unit vector) yields the explicit description of

$$R_n = P_n^* : V' \rightarrow X, \quad (R_n v)_\alpha = (\phi_\alpha, v)_U = \int_{\Omega} \phi_\alpha v \, dx \text{ for } \alpha \in I.$$

The matrix A_n in (E.7a,b) is the discrete equivalent of the operator A in (E.4). A direct connection is described below.

Proposition E.9. $A_n = R_n A P_n$ and $f_n = R_n f$ are the quantities in (E.7a,b).

Proof. The first identity is shown by $A_{n,\alpha\beta} = a(\phi_\beta, \phi_\alpha) = a(P_n \mathbf{e}_\beta, P_n \mathbf{e}_\alpha) = (A P_n \mathbf{e}_\beta, P_n \mathbf{e}_\alpha)_U = \langle R_n A P_n \mathbf{e}_\beta, \mathbf{e}_\alpha \rangle_X = h^d (R_n A P_n)_{\alpha\beta}$. The equations

$$f_{n,\alpha} = (f, \phi_\alpha)_U = (f, P_n \mathbf{e}_\alpha)_U = \langle R_n f, \mathbf{e}_\alpha \rangle_X = h^d (R_n f)_\alpha$$

prove the second one. \square

The product $M_n := R_n A P_n$ is called the *mass matrix* or the *Gram matrix* of the basis $\{\phi_\alpha\}$.

Proposition E.10. M_n is positive definite. The extreme eigenvalues of M_n determine the best bounds in (E.9):

$$\underline{C}_P = \sqrt{\|M_n^{-1}\|} = 1/\sqrt{\lambda_{\min}(M_n)}, \quad \bar{C}_P = \sqrt{\|M_n\|} = \sqrt{\lambda_{\max}(M_n)}. \quad (\text{E.10a})$$

Furthermore,

$$\|P_n\|_{U \leftarrow X} = \|R_n\|_{X \leftarrow U} = \sqrt{\|M_n\|} = \|M_n^{1/2}\| \quad (\text{E.10b})$$

holds.

Proof. (i) $\bar{C}_P^2 \|x\|_X^2 \stackrel{(E.9)}{\leq} \|P_n x\|_U^2 = (P_n x, P_n x)_U = \langle R_n P_n x, x \rangle_X = \langle M_n x, x \rangle_X = \|M_n\| \|x\|_X^2$.

(ii) The estimate $\|M_n^{-1}\|_2 = \frac{1}{\lambda}$ holds with $\lambda = \lambda_{\min}(M_n)$. The corresponding eigenvector x satisfies

$$\lambda \|x\|_X^2 = \lambda \langle x, x \rangle_X = \langle M_n x, x \rangle_X = (P_n x, P_n x)_U = \|P_n x\|_U^2 \geq \|x\|_X^2 / \underline{C}_P^2.$$

(iii) The last statement follows from (B.21a,b). \square

Conclusion E.11. *The condition of M_n satisfies*

$$\text{cond}(M_n) = \kappa(M_n) \leq (\bar{C}_P \underline{C}_P)^2.$$

For the asymptotic behaviour, we have to consider a family of Galerkin discretisations described by a sequence $\{V_n : n \in \mathbb{N}' \subset \mathbb{N}\}$ of subspaces for an infinite subset \mathbb{N}' . We conclude that $(\bar{C}_P \underline{C}_P)^2$ is uniformly bounded if and only if

$$\sup_{n \in \mathbb{N}'} \text{cond}(M_n) < \infty.$$

The latter property holds under suitable assumptions on the finite elements (cf. [193, Remark 8.8.4], [100, Corollary 3.2]).

Since $M_n = R_n P_n : X \rightarrow X$ is invertible, the mappings

$$\hat{P}_n := P_n (R_n P_n)^{-1}, \quad \hat{R}_n = \hat{P}_n^* := (R_n P_n)^{-1} R_n \quad (\text{E.11a})$$

exist. $Q_n := \hat{P}_n R_n = P_n (R_n P_n)^{-1} R_n : U \rightarrow U$ is the orthogonal projection onto V_n . The identity

$$R_n \hat{P}_n = \hat{R}_n P_n = I : X \rightarrow X \quad (\text{E.11b})$$

proves that \hat{R}_n is the left-inverse of $P_n : X \rightarrow V$, while R_n is the left-inverse of $\hat{P}_n : X \rightarrow V$.

Lemma E.12. *(E.11c) holds with \underline{C}_P in (E.9) and vice versa:*

$$\|\hat{P}_n\|_{U \leftarrow X} = \|\hat{R}_n\|_{X \leftarrow U} = \|M_n^{-1/2}\| \leq \underline{C}_P. \quad (\text{E.11c})$$

Proof. $\|\hat{P}_n\|_{U \leftarrow X}^2 = \|\hat{P}_n^* \hat{P}_n\| = \|\hat{R}_n \hat{P}_n\| = \|(R_n P_n)^{-1} R_n P_n (R_n P_n)^{-1}\| = \|M_n^{-1}\|$. \square

Proposition E.13. *Assume $\sup_{n \in \mathbb{N}'} \text{cond}(M_n) < \infty$, let the basis functions ϕ_α be scaled by (E.8), and assume that M_n is uniformly sparse:*

$$\sup_{n \in \mathbb{N}'} \max_\alpha \#\{\beta : M_{n,\alpha\beta} \neq 0\} < \infty$$

(the latter property is ensured by the shape regularity of the finite element triangulation). Then inequality (E.9) holds.

Proof. (i) $\|\phi_\alpha\|_{L_2(\Omega)} \lesssim 1$ implies $M_{n,\alpha\beta} \sim 1$. Together with the uniform sparsity of M_n , we conclude that $\lambda_{\max}(M_n) \sim 1$ and $\|P_n x\|_U \lesssim \|x\|_X$.

(ii) The first inequality $\underline{C}_P^{-1} \|x\|_X \leq \|P_n x\|_U$ follows from $\|\hat{R}_n v\|_X \leq \underline{C}_P \|v\|_U$ by setting $v = P_n x$. By (E.11c),

$$\|\hat{R}_n v\|_X \lesssim \|v\|_U / \sqrt{\lambda_{\min}(M_n)} = \|v\|_U \sqrt{\text{cond}(M_n) / \lambda_{\max}(M_n)}$$

holds. With $\text{cond}(M_n) \lesssim 1$ and part (i) we arrive at $\|\hat{R}_n v\|_X \lesssim \|v\|_U$. \square

The *inverse estimate* is the inequality

$$\|v\|_V \leq C_{\text{inv}} h^{-m} \|v\|_U \quad \text{for all } v \in V_n, \quad (\text{E.12a})$$

which holds under suitable conditions on the finite elements.

Exercise E.14. Assume (E.11c), (E.12a), and the boundedness

$$|a(u, v)| \leq C_a \|u\|_V \|v\|_V \quad \text{for all } u, v \in V \quad (\text{E.12b})$$

of a (i.e., $\|a\| \leq C_a$, cf. (E.1)). Prove (E.12c) with $C_K := C_a (C_{\text{inv}} \bar{C}_P)^2$:

$$\|A_n\|_2 \leq C_K h^{-2m}. \quad (\text{E.12c})$$

E.5 Error Estimates

If A in (E.4) corresponds to the (differentiation) order $2m$, V is (a subspace of) the Sobolev space $H^m(\Omega)$, whose norm is now denoted by $\|\cdot\|_V = \|\cdot\|_m$. By arguments from approximation theory and under suitable assumptions on V_n , one finds estimates of the form

$$\|u - u_n\|_m \leq C_m h^m \|u\|_{2m} \quad (u, u_n: \text{solutions of (E.2), (E.5)}). \quad (\text{E.13a})$$

$\|\cdot\|_{2m}$ is the norm of the Sobolev space $H^{2m}(\Omega) \cap V$.

The problem (E.2) is called $2m$ -regular if

$$\|u\|_{2m} \leq C_{\text{reg}} \|f\|_U \quad \text{for } u = A^{-1} f \text{ and all } f \in U. \quad (\text{E.13b})$$

The inequalities (E.13a) and (E.13b) yield the error estimate

$$\|u - u_n\|_V = \|u - u_n\|_m \leq Ch^m \|f\|_U. \quad (\text{E.13c})$$

Proposition E.9 shows that $u_n = P_n A_n^{-1} R_n f$; hence, $u - u_n = E_n f$ with

$$E_n := A^{-1} - P_n A_n^{-1} R_n.$$

Inequality (E.13c) can be expressed as

$$\|E_n\|_{V \leftarrow U} \leq Ch^m. \quad (\text{E.13d})$$

If the bilinear (sesquilinear) form a is not symmetric, we require the *adjoint problem*

$$\text{find } v^* \in V \text{ with } a(w, v) = (w, f)_U \quad \text{for all } w \in V$$

to have the same properties (E.13a,b) as the original problem (E.2). Analogously to (E.13d), we obtain

$$\|E_n^*\|_{V \leftarrow U} \leq C^* h^m.$$

As $\|E_n^*\|_{V \leftarrow U} = \|E_n\|_{U \leftarrow V'}$ (by $U = U'$), it follows that

$$\|E_n\|_{U \leftarrow V'} \leq C^* h^m.$$

The identity

$$\begin{aligned} E_n A E_n &= (A^{-1} - P_n A_n^{-1} R_n) A (A^{-1} - P_n A_n^{-1} R_n) \\ &= A^{-1} - 2P_n A_n^{-1} R_n + P_n A_n^{-1} \underbrace{R_n A P_n A_n^{-1} R_n}_{=A_n} \\ &= A^{-1} - P_n A_n^{-1} R_n = E_n \end{aligned}$$

proves

$$\|E_n\|_{U \leftarrow U} \leq \|E_n\|_{U \leftarrow V'} \|A\|_{V' \leftarrow V} \|E_n^*\|_{V \leftarrow U} \leq \|a\| C C^* h^{2m} \quad (\text{E.14})$$

with $\|a\|$ defined in (E.1). Inequality (E.14) is equivalent to

$$\|u - u_n\|_U \leq \|a\| C C^* h^{2m} \|f\|_U.$$

Combining the latter estimate with (E.12c), we obtain

$$\|E_n\|_{U \leftarrow U} \leq C'/\|A_n\|_2$$

with $C' = C C^* (C_a C_{\text{inv}} \bar{C}_P)^2$.

So far, we used the $2m$ -regularity (cf. (E.13b)). In the case of the Poisson equation, this is

$$A^{-1} = -\Delta^{-1} : U = L_2(\Omega) = H^0(\Omega) \rightarrow H^2(\Omega) \cap H_0^1(\Omega).$$

This property holds for the unit square $\Omega = (0, 1) \times (0, 1)$ as for any *convex* domain, but it does not hold, e.g., for domains with re-entrant corners. In the general case, one obtains only statements of the form

$$A^{-1} : H^{-\sigma m}(\Omega) \rightarrow H^{(2-\sigma)m}(\Omega) \cap H_0^m(\Omega) \quad \text{for some } \sigma \in (0, 1)$$

(cf. Hackbusch [193, §9.1]). A similar statement may be assumed for A^* . In this case, the previous estimates must be formulated by other norms. (E.13d) becomes

$$\|E_n\|_{V \leftarrow H^{-\sigma m}} \leq C h^{\sigma m}.$$

Repeating the proof above, we obtain $\|E_n\|_{H^{\sigma m} \leftarrow H^{-\sigma m}} \leq C' h^{\sigma m}$ and

$$\|E_n\|_{H^{\sigma m} \leftarrow H^{-\sigma m}} \leq C'/\|A_n\|_2^\sigma. \quad (\text{E.15})$$

E.6 Relations Between Two Discrete Problems

Now we investigate the case of two Galerkin discretisations corresponding to the subspaces

$$V_{n'} \subset V_n \subset V. \quad (\text{E.16})$$

Typically, the situation $V_{n'} \subset V_n$ arises when finite elements belonging to $V_{n'}$ are refined. The mappings $P_n, R_n, \hat{P}_n, \hat{R}_n, P_{n'}, R_{n'}, \hat{P}_{n'}, \hat{R}_{n'}$ are defined as above. The previous space $X = \mathbb{R}^I$ is now written as either $X_{n'} = \mathbb{R}^{I_{n'}}$ or $X_n = \mathbb{R}^{I_n}$.

Proposition E.15. *If (E.16) holds, then there are mappings $p : X_{n'} \rightarrow X_n$ and $r : X_n \rightarrow X_{n'}$ such that*

$$P_n p = P_{n'}, \quad r R_n = R_{n'}. \quad (\text{E.17})$$

Proof. We recall that $P_n : X_n \rightarrow V_n$ is an isomorphism with the inverse \hat{R}_n (similarly for n'). Therefore

$$p := \hat{R}_n P_{n'} \quad (\text{E.18})$$

satisfies $P_n p = P_n \hat{R}_n P_{n'} = P_{n'}$. The adjoint mapping

$$r := p^* = P_{n'}^* \hat{R}_n^* \underset{(\text{E.11a})}{=} R_{n'} \hat{P}_n$$

satisfies $r R_n = R_{n'}$. \square

The property (E.17) is illustrated by the following commuting diagram:

$$\begin{array}{ccc} V_{n'} & \xrightarrow{\text{id}} & V_n \\ P_{n'} \uparrow\downarrow \hat{R}_{n'} & & P_n \uparrow\downarrow \hat{R}_n \\ X_{n'} & \xrightarrow[p]{} & X_n \end{array} \quad (\text{E.19})$$

Note that $V_{n'} \subset V_n$ allows us to use the identity id as a mapping from $V_{n'}$ into V_n .

Proposition E.16. *The two matrices $A_n = R_n A P_n$, $A_{n'} = R_{n'} A P_{n'}$, and the right-hand sides $f_n = R_n f$, $f_{n'} = R_{n'} f$ (cf. Proposition E.9) are related by*

$$A_{n'} = r A_n p, \quad f_{n'} = r f_n.$$

Proof. $A_{n'} = R_{n'} A P_{n'} = (r R_n) A (P_n p) = r (R_n A P_n) p = r A_n p$. \square

References

1. Alefeld, G.: Zur Konvergenz des Peaceman-Rachford-Verfahrens. *Numer. Math.* **26**, 409–419 (1976)
2. Alefeld, G.: On the convergence of the symmetric SOR method for matrices with red-black ordering. *Numer. Math.* **39**, 113–117 (1982)
3. Alefeld, G., Varga, R.S.: Zur Konvergenz des symmetrischen Relaxationsverfahrens. *Numer. Math.* **25**, 291–295 (1976)
4. Allgower, E., Böhmer, K., Zhen, M.: On a problem decomposition for semilinear nearly symmetric elliptic equations. In: Hackbusch [187], pp. 1–17
5. Andreev, R., Tobler, C.: Multilevel preconditioning and low-rank tensor iteration for space-time simultaneous discretizations of parabolic PDEs. *Numer. Linear Algebra Appl.* **22**, 317–337 (2015)
6. Ansorge, R., Glashoff, K., Werner, B. (eds.): Numerical Mathematics, *ISNM*, Vol. 49. Birkhäuser, Basel (1979). (Hamburg, Jan. 1979)³
7. Ashby, S.F., Manteuffel, T.A., Saylor, P.E.: Adaptive polynomial preconditioning for Hermitian indefinite linear systems. *BIT* **29**, 583–609 (1989)
8. Astrachanov, G.P.: An iterative method of solving elliptic net problems. *USSR Comput. Math. Math. Phys.* **11**, 2, 171–182 (1971)
9. Astrachanov, G.P.: Methods of fictitious domains for a second-order elliptic equation with natural boundary conditions. *USSR Comput. Math. Math. Phys.* **18**, 114–121 (1978)
10. Axelsson, O.: Solution of linear systems of equations: iterative methods. In: Barker [33], pp. 1–51
11. Axelsson, O.: On algebraic multilevel iteration methods for selfadjoint elliptic problems with anisotropy. *Rend. Semin. Mat. Politec. Torino* pp. 31–61 (1991)
12. Axelsson, O.: Iterative Solution Methods. Cambridge Univ. Press (1994). Reprinted 1996
13. Axelsson, O., Barker, V.A.: Finite element solution of boundary value problems. Academic Press, Orlando (1984). Reprinted by SIAM, Philadelphia, 2001
14. Axelsson, O., Brinkkemper, S., Il'in, V.P.: On some versions of incomplete block-matrix factorization iterative methods. *Linear Algebra Appl.* **58**, 3–15 (1984)
15. Axelsson, O., Polman, B.: A robust preconditioner based on algebraic substructuring and two-level grids. In: Hackbusch [185], pp. 1–26
16. Axelsson, O., Vassilevski, P.: Algebraic multilevel preconditioning methods, part I. *Numer. Math.* **56**, 157–177 (1989)
17. Axelsson, O., Vassilevski, P.: Algebraic multilevel preconditioning methods, part II. *SIAM J. Numer. Anal.* **27**, 1569–1590 (1990)
18. Axelsson, O., Vassilevski, P.: A black box generalized CG solver with inner iterations and variable-step preconditioning. *SIAM J. Matrix Anal. Appl.* **12**, 625–644 (1991)

³ The bracket ‘(town, date)’ refers to the town and date of the corresponding conference.

19. Axelsson, O., Vassilevski, P.: Construction of variable-step preconditioners for inner-outer iteration methods. In: Beauwens and de Groen [38], pp. 1–14
20. Aziz, A.K. (ed.): The mathematical foundation of the finite element method with applications to partial differential equations. Academic Press, New York (1972). (Maryland, Jun. 1972)
21. Babuška, I.: Über Schwarzsche Algorithmen in partiellen Differentialgleichungen der mathematischen Physik. ZAMP **37**, 243–245 (1957)
22. Bachem, A., Grötschel, M., Korte, B. (eds.): Mathematical programming, the state of art. Springer, Berlin (1983). (Bonn, Aug. 1982)
23. Bakhvalov, N.S.: On the convergence of a relaxation method with natural constraints on the elliptic operator. USSR Comput. Math. Math. Phys. **6**(5), 101–135 (1966)
24. Ballani, J., Grasedyck, L.: A projection method to solve linear systems in tensor format. Numer. Linear Algebra Appl. **20**, 27–43 (2013)
25. Bank, R.E.: Marching algorithms and block Gaussian elimination. In: Bunch and Rose [86], pp. 293–307
26. Bank, R.E., Chan, T.F.: An analysis of the composite step biconjugate gradient method. Numer. Math. **66**, 295–319 (1993)
27. Bank, R.E., Douglas, C.C.: Sharp estimates for multigrid rates of convergence with general smoothing and acceleration. SIAM J. Numer. Anal. **22**, 617–633 (1985)
28. Bank, R.E., Dupont, T.F.: Analysis of a two-level scheme for solving finite element equations. Report CNA-159, University of Texas at Austin (1980)
29. Bank, R.E., Dupont, T.F.: An optimal order process for solving elliptic finite element equations. Math. Comp. **36**, 35–51 (1981)
30. Bank, R.E., Dupont, T.F., Yserentant, H.: The hierarchical basis multigrid method. Math. Comp. **52**, 427–458 (1988)
31. Bank, R.E., Holst, M.J., Widlund, O.B., Xu, J. (eds.): Domain Decomposition Methods in Science and Engineering XX, *Lect. Notes Comput. Sci. Eng.*, Vol. 91. Springer, Berlin (2013). (San Diego, Feb. 2011)
32. Bank, R.E., Scott, L.R.: On the conditioning of finite element equations with highly refined meshes. SIAM J. Numer. Anal. **26**, 1383–1394 (1989)
33. Barker, V.A. (ed.): Sparse matrix techniques, *Lect. Notes Math.*, Vol. 572. Springer, Berlin (1977). (Copenhagen, Aug. 1976)
34. Bastian, P., Hackbusch, W., Wittum, G.: Additive and multiplicative multi-grid – a comparison. Computing **60**, 345–364 (1998)
35. Baur, U.: Low-rank solution of data-sparse Sylvester equations. Numer. Linear Algebra Appl. **15**, 837–851 (2008)
36. Baur, U., Benner, P.: Factorized solution of Lyapupov equations based on hierarchical matrix arithmetic. Computing **78**, 211–234 (2006)
37. Beauwens, R.: Approximate factorizations with s/p consistently ordered M-factors. BIT **29**, 658–681 (1989)
38. Beauwens, R., de Groen, P. (eds.): Iterative Methods in Linear Algebra. North-Holland, Amsterdam (1992). (Brussels, April 1991)
39. Bebendorf, M., Hackbusch, W.: Existence of \mathcal{H} -matrix approximants to the inverse FE-matrix of elliptic operators with L^∞ -coefficients. Numer. Math. **95**, 1–28 (2003)
40. Benner, P., Breiten, T.: Low rank methods for a class of generalized Lyapunov equations and related issues. Numer. Math. **124**, 441–470 (2013)
41. Benzi, M.: Gianfranco Cimmino's contributions to numerical mathematics. In: Ciclo di Conferenze in Ricordo di Gianfranco Cimmino, March-May 2004, pp. 87–109. Seminario di Analisi Matematica, Dipartimento di Matematica dell'Università di Bologna (2005)
42. Benzi, M., Tůma, M.: A robust incomplete factorization preconditioner for positive definite matrices. Numer. Linear Algebra Appl. **10**, 385–400 (2003)
43. Bercovier, M., Gander, M.J., Kornhuber, R., Widlund, O.B. (eds.): Domain Decomposition Methods in Science and Engineering XVIII, *Lect. Notes Comput. Sci. Eng.*, Vol. 70. Springer, Berlin (2009). (Jerusalem, Jan. 2008)
44. Berg, L.: Lineare Gleichungssysteme mit Bandstruktur. Deutscher Verlag der Wissenschaften, Berlin (1986)

45. Berman, A., Plemmons, R.J.: Nonnegative matrices in the mathematical sciences. Academic Press, New York (1979)
46. Berman, A., Plemmons, R.J.: Cones and iterative methods for best least squares solutions of linear systems. SIAM J. Numer. Anal. **11**, 145–154 (2006)
47. Björck, Å.: Numerical Methods for Least Squares Problems. SIAM, Philadelphia (1996)
48. Björck, Å.: Numerical Methods in Matrix Computations. Springer, Cham (2015)
49. Bjørstad, P.E.: The direct solution of a generalized biharmonic equation on a disk. In: Hackbusch [182], pp. 1–9
50. Bjørstad, P.E., Espedal, M.S., Keyes, D.E. (eds.): Ninth International Conference on Domain Decomposition Methods. ddm.org (1996). (Bergen, June 1996)
51. Bjørstad, P.E., Mandel, J.L.: On the spectra of sums of orthogonal projections with applications to parallel computing. BIT **31**, 76–88 (1991)
52. Bjørstad, P.E., Widlund, O.B.: Iterative methods for the solution of elliptic problems on regions partitioned into substructures. SIAM J. Numer. Anal. **23**, 1097–1120 (1986)
53. Börgers, C., Widlund, O.B.: On finite element domain imbedding methods. SIAM J. Numer. Anal. **27**, 963–978 (1990)
54. Börn, S.: Efficient Numerical Methods for Non-local Operators. EMS, Zürich (2010). Corrected 2nd printing, 2013
55. Börn, S., Reimer, K.: Efficient arithmetic operations for rank-structured matrices based on hierarchical low-rank updates. Comput. Vis. Sci. **16**, 247–258 (2013) [published 2015]
56. Bornemann, F., Deuflhard, P.: The cascadic multigrid method for elliptic problems. Numer. Math. **75**, 135–152 (1996)
57. Bornemann, F., Yserentant, H.: A basic norm equivalence for the theory of multilevel methods. Numer. Math. **64**, 455–476 (1993)
58. Braess, D.: The contraction number of a multigrid method for solving the Poisson equation. Numer. Math. **37**, 387–404 (1981)
59. Braess, D.: The convergence rate of a multigrid method with Gauß-Seidel relaxation for the Poisson equation. Math. Comp. **42**, 505–386 (1984)
60. Braess, D.: Nonlinear Approximation Theory. Springer, Berlin (1986)
61. Braess, D.: On the combination of the multigrid method and conjugate gradients. In: Hackbusch and Trottenberg [206], pp. 52–64
62. Braess, D.: A multigrid method for the membran problem. Comput. Mech. **3**, 321–329 (1988)
63. Braess, D.: Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics, 3rd ed. Cambridge University Press, Cambridge (2007)
64. Braess, D., Dryja, M., Hackbusch, W.: Grid transfer for nonconforming FE-discretisations with application to non-matching grids. Computing **63**, 1–25 (1999)
65. Braess, D., Hackbusch, W.: A new convergence proof for the multigrid method including the V-cycle. SIAM J. Numer. Anal. **20**, 967–975 (1983)
66. Braess, D., Hackbusch, W.: Approximation of $1/x$ by exponential sums in $[1, \infty)$. IMA J. Numer. Anal. **25**, 685–697 (2005)
67. Braess, D., Hackbusch, W.: On the efficient computation of high-dimensional integrals and the approximation by exponential sums. In: DeVore and Kunoth [107], pp. 39–74
68. Braess, D., Hackbusch, W., Trottenberg, U. (eds.): Advances in Multi-Grid Methods, *Notes on Numerical Fluid Mechanics*, Vol. 11. Vieweg, Braunschweig (1985). (Oberwolfach, Dec. 1984)
69. Brakhage, H.: Über die numerische Behandlung von Integralgleichungen nach der Quadraturformelmethode. Numer. Math. **2**, 183–196 (1960)
70. Bramble, J.H., Pasciak, J.E., Schatz, A.H.: The construction of preconditioners for elliptic problems by substructuring, part I. Math. Comp. **47**, 103–134 (1986)
71. Bramble, J.H., Pasciak, J.E., Schatz, A.H.: The construction of preconditioners for elliptic problems by substructuring, part II. Math. Comp. **49**, 1–16 (1987)
72. Bramble, J.H., Pasciak, J.E., Schatz, A.H.: The construction of preconditioners for elliptic problems by substructuring, part III. Math. Comp. **51**, 415–430 (1988)
73. Bramble, J.H., Pasciak, J.E., Schatz, A.H.: The construction of preconditioners for elliptic problems by substructuring, part IV. Math. Comp. **53**, 1–24 (1989)

74. Bramble, J.H., Pasciak, J.E., Wang, J., Xu, J.: Convergence estimates for multigrid algorithms without regularity assumptions. *Math. Comp.* **57**, 23–45 (1991)
75. Bramble, J.H., Pasciak, J.E., Xu, J.: The analysis of multigrid algorithms for nonsymmetric and indefinite elliptic problems. *Math. Comp.* **51**, 389–414 (1988)
76. Bramble, J.H., Pasciak, J.E., Xu, J.: Parallel multilevel preconditioners. *Math. Comp.* **55**, 1–22 (1990)
77. Brandt, A.: Multi-level adaptive solutions to boundary-value problems. *Math. Comp.* **31**, 333–390 (1977)
78. Brenner, S.C.: A new look at FETI. In: Debit et al. [106], pp. 41–51
79. Brenner, S.C.: Convergence of the multigrid V-cycle algorithm for second-order boundary value problems without full elliptic regularity. *Math. Comp.* **71**, 507–525 (2002)
80. Brenner, S.C.: An additive Schwarz preconditioner for the FETI method. *Numer. Math.* **94**, 1–31 (2003)
81. Brenner, S.C.: Convergence of nonconforming V-cycle and F-cycle multigrid algorithms for second order elliptic boundary value problems. *Math. Comp.* **73**, 1041–1066 (2004)
82. Brenner, S.C., Scott, L.R.: *The Mathematical Theory of Finite Element Methods*, 3rd ed. Springer, New York (2008)
83. Buczyński, J.A., Landsberg, J.M.: On the third secant variety. *J. Algebraic Combin.* (2014). Published on-line
84. Buleev, N.I.: Numerical method for solving two- and three-dimensional diffusion equations [in Russian]. *Mat. Sb.* **51**, 227–238 (1960)
85. Bulirsch, R., Grigorieff, R.D., Schröder, J. (eds.): *Numerical treatment of differential equations, Lect. Notes Math.*, Vol. 631. Springer, Berlin (1978). (Oberwolfach, July 1976)
86. Bunch, J.R., Rose, D.J. (eds.): *Sparse Matrix Computations*. Academic Press, New York (1976). (Argonne National Laboratory, Sep. 1975)
87. Buneman, O.: A compact non-iterative Poisson solver. SUIPR Report 294, Stanford University (1969)
88. Buoni, J.J., Varga, R.S.: Theorems of Stein–Rosenberg type. In: Ansorge et al. [6], pp. 65–75
89. Buoni, J.J., Varga, R.S.: Theorems of Stein–Rosenberg type II, optimal paths of relaxation in the complex plane. In: Schultz [334], pp. 231–240
90. Buzbee, B., Golub, G.H., Nielson, C.: On direct methods for solving Poisson's equations. *SIAM J. Numer. Anal.* **7**, 627–656 (1970)
91. Buzdin, A., Wittum, G.: Two-frequency decomposition. *Numer. Math.* **97**, 269–295 (2004)
92. Cai, X.C., Widlund, O.B.: Domain decomposition algorithms for indefinite elliptic problems. *SIAM J. Sci. Statist. Comput.* **13**, 243–258 (1992)
93. Chan, T.F., Glowinski, R., Périaux, J., Widlund, O.B. (eds.): *Domain Decomposition Methods*. SIAM, Philadelphia (1989). (Los Angeles, Jan. 1988)
94. Chan, T.F., Glowinski, R., Périaux, J., Widlund, O.B. (eds.): *Domain Decomposition Methods for Partial Differential Equations*. SIAM, Philadelphia (1990). (Houston, Mar. 1989)
95. Chan, T.F., Kako, T., Kawarada, H., Pironneau, O. (eds.): *Domain Decomposition Methods in Science and Engineering*. ddm.org (2001). (Chiba, Oct. 1999)
96. Ciarlet, P.G., Lions, J.L.: *Handbook of Numerical Analysis*. North-Holland, Amsterdam (1990)
97. Cimmino, G.: Calcolo approssimato per le soluzioni dei sistemi di equazioni lineari. *Ricerca Scientifica ed il Progresso Tecnico* **9**, 326–333 (1938)
98. Concus, P., Golub, G.H.: A generalized conjugate gradient method for nonsymmetric systems of linear equations. In: Glowinski and Lions [152], pp. 56–65
99. Concus, P., Golub, G.H., Meurant, G.A.: Block preconditioning for the conjugate gradient method. *SIAM J. Sci. Statist. Comput.* **6**, 220–252 (1985)
100. Dahmen, W., Faermann, B., Graham, I.G., Hackbusch, W., Sauter, S.A.: Inverse inequalities on non-quasuniform meshes and applications to the mortar element method. *Math. Comp.* **73**, 1107–1138 (2003)
101. Dahmen, W., Kunoth, A.: Multilevel preconditioning. *Numer. Math.* **63**, 315–344 (1992)
102. de Boor, C., Rice, J.R.: Extremal polynomials with applications to Richardson iteration for indefinite linear systems. *SIAM J. Sci. Statist. Comput.* **3**, 47–57 (1982)

103. De Lathauwer, L., De Moor, B., Vandewalle, J.: A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.* **21**, 1253–1278 (2000)
104. de Zeeuw, P.M.: Matrix-dependent prolongations and restrictions in a blackbox multigrid solvers. *J. Comput. Appl. Math.* **33**, 1–27 (1990)
105. de Zeeuw, P.M.: Incomplete line LU as smoother and as preconditioner. In: Hackbusch and Wittum [208], pp. 215–224
106. Debit, N., Garbey, M., Hoppe, R.H.W., Keyes, D.E., Kuznetsov, Y.A., Pélriaux, J. (eds.): Domain Decomposition Methods in Science and Engineering. CIMNE, Barcelona (2002). (Lyon, Oct. 2000)
107. DeVore, R.A., Kunoth, A. (eds.): Multiscale, Nonlinear and Adaptive Approximation. Springer, Berlin (2009) (Günzburg, Oct. 2009)
108. Dick, E., Riemslagh, K., Vierendeels, J. (eds.): Multigrid methods VI, *Lect. Notes Comput. Sci. Eng.*, Vol. 14. Springer, Berlin (2000). (Gent, Sep. 1999)
109. Dickopf, T., Gander, M.J., Halpern, L., Krause, R., Pavarino, L.F. (eds.): Domain Decomposition Methods in Science and Engineering XXII, *Lect. Notes Comput. Sci. Eng.*, Vol. 104. Springer, Cham (2016). (Lugano, Sep. 2013)
110. D’Jakovov, E.G.: The construction of iterative methods based on the use of spectrally equivalent operators. *USSR Comput. Math. Math. Phys.* **6**, 1, 14–46 (1966)
111. Dryja, M.: A finite-capacitance matrix method for elliptic problems on regions partitioned into subregions. *Numer. Math.* **44**, 153–168 (1984)
112. Dryja, M.: An additive Schwarz algorithm for two- and three-dimensional finite element elliptic problems. In: Chan et al. [93], pp. 168–172
113. Dryja, M., Hackbusch, W.: On the nonlinear domain decomposition method. *BIT* **37**, 296–311 (1997)
114. Dryja, M., Widlund, O.B.: Towards a unified theory of domain decomposition algorithms for elliptic problems. In: Chan et al. [94], pp. 3–21
115. Dryja, M., Widlund, O.B.: Multilevel additive methods for elliptic finite element problems. In: Hackbusch [187], pp. 58–69
116. Duff, I.S., Erisman, A.M., Reid, J.K.: Direct Methods for Sparse Matrices. Clarendon Press, Oxford (1989)
117. Duff, I.S., Meurant, G.A.: The effect of ordering on preconditioned conjugate gradients. *BIT* **29**, 635–657 (1989)
118. Eiermann, M., Niethammer, W., Ruttan, A.: Optimal successive overrelaxation iterative methods for p -cyclic matrices. *Numer. Math.* **57**, 593–606 (1990)
119. Eiermann, M., Niethammer, W., Varga, R.S.: A study of semiiterative methods for non-symmetric systems of linear equations. *Numer. Math.* **47**, 505–533 (1985)
120. Eiermann, M., Niethammer, W., Varga, R.S.: Iterationsverfahren für nichtsymmetrische Gleichungssysteme und Approximationsmethoden im Komplexen. *Jber. d. Dt. Math.-Verein.* **89**, 1–33 (1987)
121. Elman, H.C.: Relaxed and stabilized incomplete factorizations for non-self-adjoint linear systems. *BIT* **29**, 890–915 (1989)
122. Erhel, J., Gander, M.J., Halpern, L., Pichot, G., Sassi, T., Widlund, O.B. (eds.): Domain Decomposition Methods in Science and Engineering XXI, *Lect. Notes Comput. Sci. Eng.*, Vol. 98. Springer, Berlin (2014). (Rennes, Jun. 2012)
123. Espig, M., Hackbusch, W., Khachatryan, A.: On the convergence of alternating least squares optimisation in tensor format representations. arXiv (2015)
124. Espig, M., Hackbusch, W., Litvinenko, A., Matthies, H.G., Wähnert, P.: Efficient low-rank approximation of the stochastic Galerkin matrix in tensor formats. *Comput. Math. Appl.* **67**, 818–829 (2014)
125. Espig, M., Hackbusch, W., Rohwedder, T., Schneider, R.: Variational calculus with sums of elementary tensors of fixed rank. *Numer. Math.* **122**, 469–488 (2012)
126. Falcó, A., Nouy, A.: Proper generalized decomposition for nonlinear convex problems in tensor Banach spaces. *Numer. Math.* **121**, 503–530 (2012)
127. Farhat, C., Roux, F.X.: A method of finite element tearing and interconnecting and its parallel solution algorithm. *Int. J. Num. Meth. Engng.* **32**, 1205–1227 (1991)

128. Faustmann, M.: Approximation inverser Finite Elemente- und Randelementematrizen mittels hierarchischer Matrizen. Doctoral thesis, Technische Universität Wien (2015)
129. Faustmann, M., Melenk, J.M., Praetorius, D.: \mathcal{H} -matrix approximability of the inverses of FEM matrices. *Numer. Math.* **131**, 615–642 (2015)
130. Faustmann, M., Melenk, J.M., Praetorius, D.: Existence of \mathcal{H} -matrix approximants to the inverses of BEM matrices: the simple-layer operator. *Math. Comp.* **85**, 119–152 (2016)
131. Fedorenko, R.P.: A relaxation method for solving elliptic difference equations. *USSR Comput. Math. Math. Phys.* **1**, 1092–1096 (1961)
132. Fedorenko, R.P.: The speed of convergence of one iterative process. *USSR Comput. Math. Math. Phys.* **4**, 227–235 (1964)
133. Fischer, B.: Polynomial Based Iteration Methods for Symmetric Linear Systems. Advances in Numerical Mathematics. J. Wiley and Teubner, Stuttgart (1996)
134. Fischer, B., Freund, R.W.: On the constrained Chebyshev approximation problem on ellipses. *J. Approx. Theory* **62**, 297–315 (1990)
135. Fischer, B., Freund, R.W.: Chebyshev polynomials are not always optimal. *J. Approx. Theory* **65**, 261–272 (1991)
136. Fletcher, R.: Conjugate gradient methods for indefinite systems. In: Watson [389], pp. 73–89
137. Forsythe, G.E.: Gauss to Gerling on relaxation. *Math. Tables and Other Aids to Computation* **5**, 255–258 (1951)
138. Forsythe, G.E.: Solving linear algebraic equations can be interesting. *Bull. Amer. Math. Soc.* **59**, 299–329 (1953)
139. Forsythe, G.E., Strauss, E.G.: On best conditioned matrices. *Proc. Amer. Soc.* **6**, 340–345 (1955)
140. Fridman, V.M.: The method of minimum iterations with minimum errors for a system of linear algebraic equations with a symmetrical matrix. *USSR Comput. Math. Math. Phys.* **2**, 362–363 (1963)
141. Frobenius, G.: Über Matrizen aus positiven Elementen. *Sitzungsbericht Akad. Wiss. Phys.-math. Klasse Berlin* pp. 417–476 (1908)
142. Frobenius, G.: Über Matrizen aus positiven Elementen. *Sitzungsbericht Akad. Wiss. Phys.-math. Klasse Berlin* pp. 514–518 (1909)
143. Frommer, A., Szyld, D.B.: H-Splittings and two-stage iterative methods. *Numer. Math.* **63**, 345–356 (1992)
144. Gantmacher, F.R.: Matrizenrechnung, Band I. Deutscher Verlag der Wissenschaften, Berlin (1958)
145. Gantmacher, F.R.: The Theory of Matrices, Vol. 1. AMS Chelsea Publ. (1959)
146. Gauss, C.F.: Nachlass: *Theoria interpolationis : methodo nova tractata* [1805]. In: Werke, Vol. 3, pp. 265–303. Königliche Gesellschaft der Wissenschaft, Göttingen (1866). Reprint by Georg Olms, Hildesheim, 1981
147. Gauss, C.F.: *Supplementum theoriae combinationis observationum erroribus minimis obnoxiae* [1826]. In: Werke, Vol. 4, pp. 55–93. Königliche Gesellschaft der Wissenschaft, Göttingen (1873). Reprint by Georg Olms, Hildesheim, 1981
148. Gauss, C.F.: Brief an Gerling [1823]. In: Werke, Vol. 9, pp. 278–281. Königliche Gesellschaft der Wissenschaft, Göttingen (1903). Reprint by Georg Olms, Hildesheim, 1981. English translation in [137]
149. George, A.: Nested dissection of a regular finite element mesh. *SIAM J. Numer. Anal.* **10**, 345–363 (1973)
150. Glowinski, R., Golub, G.H., Meurant, G.A., Péliaux, J. (eds.): First International Symposium on Domain Decomposition Methods for Partial Differential Equations. SIAM, Philadelphia (1988). (Paris, Jan. 1987)
151. Glowinski, R., Kuznetsov, Y.A., Meurant, G.A., Péliaux, J., Widlund, O.B. (eds.): Fourth International Symposium on Domain Decomposition Methods for Partial Differential Equations. SIAM, Philadelphia (1991). (Moscow, May 1990)
152. Glowinski, R., Lions, J.L. (eds.): Computing Methods in Applied Sciences and Engineering, *Lect. Notes Econ. Math. Syst.*, Vol. 134. Addison-Wesley Publ., Reading, Mass. (1976)

153. Glowinski, R., P\'eriaux, J., Shi, Z.C., Widlund, O.B. (eds.): Domain Decomposition Methods in Science and Engineering. John Wiley & Sons, Strasbourg (1997). (Beijing, May 1995)
154. Golub, G.H.: Direct methods for solving elliptic difference equations. In: Morris [285], pp. 1–19.
155. Golub, G.H., O’Leary, D.P.: Some history of the conjugate gradient and Lanczos algorithms: 1948–1976. *SIAM Rev.* **31**, 50–102 (1989)
156. Golub, G.H., Overton, M.L.: The convergence of inexact Chebyshev and Richardson iterative methods for solving linear systems. *Numer. Math.* **53**, 571–593 (1988)
157. Golub, G.H., Van Loan, C.F.: Matrix Computations, 3rd ed. The Johns Hopkins University Press, Baltimore (1996)
158. Grasedyck, L.: Existence and computation of low Kronecker-rank approximations for large linear systems of tensor product structure. *Computing* **72**, 247–265 (2004)
159. Grasedyck, L.: Existence of a low rank or \mathcal{H} -matrix approximant to the solution of a Sylvester equation. *Numer. Linear Algebra Appl.* **11**, 371–389 (2004)
160. Grasedyck, L.: Nonlinear multigrid for the solution of large-scale Riccati equations in low-rank and \mathcal{H} -matrix format. *Numer. Linear Algebra Appl.* **15**, 779–807 (2008)
161. Grasedyck, L., Hackbusch, W.: Construction and arithmetics of \mathcal{H} -matrices. *Computing* **70**, 295–334 (2003)
162. Grasedyck, L., Hackbusch, W.: A multigrid method to solve large scale Sylvester equations. *SIAM J. Matrix Anal. Appl.* **29**, 870–894 (2007)
163. Grasedyck, L., Hackbusch, W., Khoromskij, B.: Solution of large scale algebraic matrix Riccati equations by use of hierarchical matrices. *Computing* **70**, 121–165 (2003)
164. Grasedyck, L., Hackbusch, W., Kriemann, R.: Performance of \mathcal{H} -LU preconditioning for sparse matrices. *Comput. Methods Appl. Math.* **8**, 336–349 (2008)
165. Grasedyck, L., Kriemann, R., Le Borne, S.: Parallel black box \mathcal{H} -LU preconditioning for elliptic boundary value problems. *Comput. Vis. Sci.* **11**, 273–291 (2008)
166. Grasedyck, L., Kriemann, R., Le Borne, S.: Domain decomposition based \mathcal{H} -LU preconditioning. *Numer. Math.* **112**, 565–600 (2009)
167. Greenbaum, A.: Iterative Methods for Solving Linear Systems. SIAM, Philadelphia (1997)
168. Griebel, M.: Multilevel algorithms considered as iterative methods on semidefinite systems. *SIAM J. Sci. Comput.* **15**, 547–565 (1994)
169. Griebel, M.: Multilevelmethoden als Iterationsverfahren über Erzeugendensystemen. Teubner Skripten zur Numerik. Teubner, Stuttgart (1994)
170. Griebel, M., Oswald, P.: On the abstract theory of additive and multiplicative Schwarz algorithms. *Numer. Math.* **70**, 163–180 (1995)
171. Gunn, J.E.: The solution of elliptic difference equations by semi-explicit iterative techniques. *SIAM J. Numer. Anal.* **2**, 24–45 (1964)
172. Gustafsson, I.: A class of first order factorization methods. *BIT* **18**, 142–156 (1978)
173. Gutknecht, M.H.: A completed theory of the unsymmetric Lanczos process and related algorithms, part I. *SIAM J. Matrix Anal. Appl.* **13**, 594–639 (1992)
174. Gutknecht, M.H.: A completed theory of the unsymmetric Lanczos process and related algorithms, part II. *SIAM J. Matrix Anal. Appl.* **15**, 15–58 (1994)
175. Gutknecht, M.H.: Lanczos-type solvers for nonsymmetric linear systems of equations. *Acta Numerica* **6**, 271–397 (1997)
176. Hackbusch, W.: A fast iterative method solving Poisson’s equation in a general region. In: Bulirsch et al. [85], pp. 51–62
177. Hackbusch, W.: The fast numerical solution of very large elliptic difference schemes. *J. Inst. Maths. Appl.* **26**, 119–132 (1980)
178. Hackbusch, W.: Bemerkungen zur iterativen Defektkorrektur und zu ihrer Kombination mit Mehrgitterverfahren. *Rev. Roumaine Math. Pures Appl.* **26**, 1319–1329 (1981)
179. Hackbusch, W.: Fast numerical solution of time-periodic parabolic problems by a multi-grid method. *SIAM J. Sci. Statist. Comput.* **2**, 198–206 (1981)
180. Hackbusch, W.: On the regularity of difference schemes. *Ark. Mat.* **19**, 71–95 (1981)
181. Hackbusch, W.: On the regularity of difference schemes, part II: regularity estimates for linear and nonlinear problems. *Ark. Mat.* **21**, 3–28 (1983)

182. Hackbusch, W. (ed.): Efficient Solvers for Elliptic Systems, *Notes on Numerical Fluid Mechanics*, Vol. 10. Vieweg, Braunschweig (1984). (Kiel, Jan. 1984)
183. Hackbusch, W.: Multi-grid Methods and Applications, *SCM*, Vol. 4. Springer, Berlin (1985)
184. Hackbusch, W.: Multi-grid methods of the second kind. In: Paddon and Holstein [306], pp. 11–83
185. Hackbusch, W. (ed.): Robust Multi-grid Methods, *Notes on Numerical Fluid Mechanics*, Vol. 23. Vieweg, Braunschweig (1988). (Kiel, Jan. 1988)
186. Hackbusch, W.: The frequency decomposition multi-grid method, part I: application to anisotropic equations. *Numer. Math.* **56**, 229–245 (1989)
187. Hackbusch, W. (ed.): Parallel algorithms for PDEs, *Notes on Numerical Fluid Mechanics*, Vol. 31. Vieweg, Braunschweig (1991). (Kiel, Jan. 1990)
188. Hackbusch, W.: The solution of large systems of BEM equations by the multi-grid and panel clustering technique. *Rend. Semin. Mat. Politec. Torino* pp. 163–187 (1991)
189. Hackbusch, W.: Comparison of different multi-grid variants for nonlinear equations. *ZAMM* **72**, 148–151 (1992)
190. Hackbusch, W.: The frequency decomposition multi-grid method, part II: convergence analysis based on the additive Schwarz method. *Numer. Math.* **63**, 433–453 (1992)
191. Hackbusch, W.: Integral Equations – Theory and Numerical Treatment, *ISNM*, Vol. 128. Birkhäuser, Basel (1995)
192. Hackbusch, W.: Direct domain decomposition using the hierarchical matrix technique. In: Herrera et al. [217], pp. 39–50
193. Hackbusch, W.: Elliptic Differential Equations – Theory and Numerical Treatment, *SSCM*, Vol. 18, 2nd ed. Springer, Berlin (2003)
194. Hackbusch, W.: Multi-Grid Methods and Applications, *SCM*, Vol. 4. Springer, Berlin (2003)
195. Hackbusch, W.: Tensor Spaces and Numerical Tensor Calculus, *SSCM*, Vol. 42. Springer, Berlin (2012)
196. Hackbusch, W.: Numerical tensor calculus. *Acta Numerica* **23**, 651–742 (2014)
197. Hackbusch, W.: The Concept of Stability in Numerical Mathematics, *SSCM*, Vol. 45. Springer, Berlin (2014)
198. Hackbusch, W.: Hierarchical Matrices: Algorithms and Analysis, *SSCM*, Vol. 49. Springer, Berlin (2015)
199. Hackbusch, W.: Solution of linear systems in high spatial dimensions. *Comput. Vis. Sci.* **17**, 111–118 (2015).
200. Hackbusch, W.: New estimates for the recursive low-rank truncation of block-structured matrices. *Numer. Math.* **132**, 303–328 (2016).
201. Hackbusch, W.: Theorie und Numerik elliptischer Differentialgleichungen, 4th ed. Springer Spektrum, Wiesbaden (2016)
202. Hackbusch, W., Khoromskij, B., Kriemann, R.: Direct Schur complement method by domain decomposition based on \mathcal{H} -matrix approximation. *Comput. Vis. Sci.* **8**, 179–188 (2005)
203. Hackbusch, W., Khoromskij, B., Tyrtyshnikov, E.E.: Approximate iterations for structured matrices. *Numer. Math.* **109**, 365–383 (2008)
204. Hackbusch, W., Reusken, A.: Analysis of a damped nonlinear multilevel method. *Numer. Math.* **55**, 225–246 (1989)
205. Hackbusch, W., Trottenberg, U. (eds.): Multigrid Methods, *Lect. Notes Math.*, Vol. 960. Springer, Berlin (1982). (Köln-Porz, Nov. 1981)
206. Hackbusch, W., Trottenberg, U. (eds.): Multigrid Methods II, *Lect. Notes Math.*, Vol. 1228. Springer, Berlin (1986). (Köln, Oct. 1985)
207. Hackbusch, W., Trottenberg, U. (eds.): Multigrid Methods III, *ISNM*, Vol. 98. Birkhäuser, Basel (1991). (Bonn, Oct. 1990)
208. Hackbusch, W., Wittum, G. (eds.): Incomplete Decompositions (ILU) – Algorithms, Theory, and Applications, *Notes on Numerical Fluid Mechanics*, Vol. 41. Vieweg, Braunschweig (1992). (Kiel, Jan. 1992)
209. Hackbusch, W., Wittum, G. (eds.): Fast Solvers for Flow Problems, *Notes on Numerical Fluid Mechanics*, Vol. 49. Vieweg, Braunschweig (1995). (Kiel, Jan. 1994)

210. Hackbusch, W., Wittum, G. (eds.): Multigrid Methods V, *Lect. Notes Comput. Sci. Eng.*, Vol. 3. Springer, Berlin (1998). (Stuttgart, Oct. 1996)
211. Hadjidimos, A.: Successive overrelaxation (SOR) and related methods. *J. Comput. Appl. Math.* **123**, 177–199 (2000)
212. Hageman, L.A., Young, D.M.: Applied Iterative Methods. Academic Press, Orlando (1981)
213. Hanke, M., Neumann, M., Niethammer, W.: On the SOR method for symmetric positive definite systems. *Linear Algebra Appl.* **154**, 457–472 (1991)
214. Håstad, J.: Tensor rank is NP-complete. *J. Algorithms* **11**, 644–654 (1990)
215. Heideman, M.T., Johnson, D.H., Burrus, C.S.: Gauss and the history of the fast Fourier transform. *IEEE ASSP Magazine* **1**(4), 14–21 (1984)
216. Hemker, P.W., Wesseling, P. (eds.): Multigrid Methods IV, *ISNM*, Vol. 116. Birkhäuser, Basel (1994). (Amsterdam, July 1993)
217. Herrera, I., Keyes, D.E., Widlund, O.B., Yates, R. (eds.): Domain Decomposition Methods in Science and Engineering. National Autonomous University of Mexico, Mexico City (2003). (Cocoyoc, Mexico, Jan. 2002)
218. Hestenes, M.R.: Conjugate Direction Methods in Optimization. Springer, New York (1980)
219. Hestenes, M.R., Stiefel, E.: Methods of conjugate gradients for solving linear systems. *J. Res. Nat. Bur. Standards* **49**, 409–436 (1952)
220. Heun, K.: Neue Methoden zur approximativen Integration der Differentialgleichungen einer unabhängigen Veränderlichen. *Z. für Math. und Phys.* **45**, 23–38 (1900)
221. Higham, N.J.: Accuracy and Stability of Numerical Algorithms, 2nd ed. SIAM, Philadelphia (2002)
222. Higham, N.J.: Functions of Matrices, Theory and Computation. SIAM, Philadelphia (2008)
223. Hockney, R.W.: A fast direct solution of Poisson's equation using Fourier analysis. *J. ACM* **12**, 95–113 (1965)
224. Holtz, S., Rohwedder, T., Schneider, R.: The alternating linear scheme for tensor optimization in the tensor train format. *SIAM J. Sci. Comput.* **34**, A683–A713 (2012)
225. Huang, Y., Kornhuber, R., Widlund, O.B., Xu, J. (eds.): Domain Decomposition Methods in Science and Engineering XIX, *Lect. Notes Comput. Sci. Eng.*, Vol. 78. Springer, Berlin (2011). (Zhanjiajie, Aug. 2009)
226. Il'in, V.P.: Somes estimates for conjugate gradient methods. *USSR Comput. Math. Math. Phys.* **16**, 22–30 (1976)
227. Jacobi, C.G.J.: Über eine neue Auflösungsart der bei der Methode der kleinsten Quadrate vorkommenden linearen Gleichungen. *Astron. Nachr.* **32**, 297–306 (1845)
228. Jennings, A., Malik, G.M.: Partial elimination. *J. IMA* **20**, 307–316 (1977)
229. Jovanović, B., Süli, E.: Analysis of finite difference schemes for linear partial differential equations with generalized solutions, *SSCM*, Vol. 46. Springer, London (2014)
230. Kaczmarz, S.: Angenäherte Auflösung von Systemen linearer Gleichungen. *Bulletin de l'Academie Polonaise des Sciences et Lettres, Classe des Sciences Mathématiques et Naturelles, Série A* **35**, 355–357 (1937)
231. Kaczmarz, S.: Approximate solution of systems of linear equations. *Internat. J. Control* **57**, 1269–1271 (1993)
232. Kahan, W.M.: Gauss–Seidel methods of solving large systems of linear equations. Doctoral thesis, University of Toronto, Canada (1958)
233. Kanzow, C.: Numerik linearer Gleichungssysteme – Direkte und iterative Verfahren. Springer, Berlin (2005)
234. Karhunen, K.: Über lineare Methoden in der Wahrscheinlichkeitsrechnung. *Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys.* **37**, 1–79 (1947)
235. Kettler, R.: Analysis and comparison of relaxation schemes in robust multigrid and preconditioned conjugate gradient methods. In: Hackbusch and Trottenberg [205], pp. 502–534
236. Keyes, D.E., Chan, T.F., Meurant, G.A., Scroggs, J.S., Voigt, R.G. (eds.): Fifth International Symposium on Domain Decomposition Methods for Partial Differential Equations. SIAM, Philadelphia (1992). (Norfolk, May 1991)
237. Keyes, D.E., Xu, J. (eds.): Domain Decomposition Methods in Science and Engineering Computing. AMS, Providence (1994). (Penn State, Oct. 1993)

238. Khoromskij, B.: Tensor-structured preconditioners and approximate inverse of elliptic operators in \mathbb{R}^d . *Constr. Approx.* **30**, 599–620 (2009)
239. Kornhuber, R., Hoppe, R.H.W., Péraux, J., Pironneau, O., Widlund, O.B., Xu, J. (eds.): Domain Decomposition Methods in Science and Engineering XV, *Lect. Notes Comput. Sci. Eng.*, Vol. 40. Springer, Berlin (2005). (Berlin, July 2003)
240. Kosmol, P.: Methoden zur numerischen Behandlung nichtlinearer Gleichungen und Optimierungsaufgaben. Teubner, Stuttgart (1989)
241. Kosmol, P., Zhou, X.: The limit points of affine iterations. *Numer. Funct. Anal. Optim.* **11**, 403–409 (1990)
242. Kressner, D., Tobler, C.: Krylov subspace methods for linear systems with tensor product structure. *SIAM J. Matrix Anal. Appl.* **31**, 1688–1714 (2010)
243. Kressner, D., Tobler, C.: Low-rank tensor Krylov subspace methods for parametrized linear systems. *SIAM J. Matrix Anal. Appl.* **32**, 1288–1316 (2011)
244. Kressner, D., Tobler, C.: Preconditioned low-rank methods for high-dimensional elliptic PDE eigenvalue problems. *Comput. Methods Appl. Math.* **11**, 363–381 (2011)
245. Kriemann, R.: Parallel \mathcal{H} -matrix arithmetics on shared memory systems. *Computing* **74**, 273–297 (2005)
246. Kriemann, R., Le Borne, S.: \mathcal{H} -FAINV: hierarchically factored approximate inverse preconditioners. *Comput. Vis. Sci.* **17**, 135–150 (2015)
247. Kronsjö, L.: A note on the ‘nested iteration’ method. *BIT* **15**, 107–110 (1975)
248. Kronsjö, L., Dahlquist, G.: On the design of nested iterations for elliptic difference equations. *BIT* **12**, 63–71 (1972). [The front page of the paper shows wrong data: vol. 11 (1971)]
249. Krylov, A. N.: On the numerical solution of the equation by which in technical questions frequencies of small oscillations of material systems are determined [in Russian]. *Izvestiya Akademii Nauk SSSR, Otdelenie Matematicheskikh i Estestvennykh Nauk* **7**, 491–539 (1931)
250. Kutta, W. M.: Beitrag zur nähерungsweisen Integration totaler Differentialgleichungen. *Zeitschrift für Math. und Phys.* **46**, 435–453 (1901)
251. Kuznetsov, Y.A.: Algebraic domain decomposition methods I. *Sov. J. Numer. Anal. Math. Modelling* **4**, 361–392 (1989)
252. Kuznetsov, Y.A.: Multigrid domain decomposition methods for elliptic problems. *Comput. Methods Appl. Mech. Engrg.* **75**, 185–193 (1989)
253. Kuznetsov, Y.A.: Multigrid domain decomposition methods. In: Chan et al. [94], pp. 290–313
254. Lai, C.H., Bjørstad, P.E., Cross, M., Widlund, O.B. (eds.): Eleventh International Conference on Domain Decomposition Methods. ddm.org (1999). (Greenwich, July 1998)
255. Lanczos, C.: Solution of systems of linear equations by minimized iterations. *J. Res. Nat. Bur. Standards* **49**, 33–53 (1952)
256. Landweber, L.: An iteration formula for Fredholm integral equations of the first kind. *Amer. J. Math.* **73**, 615–624 (1951)
257. Langer, R.E. (ed.): Boundary Problems in Differential Equations. Kluwer Academic Publ., Dordrecht (1960). (Madison, April 1959)
258. Langer, U., Discacciati, M., Keyes, D.E., Widlund, O.B., Zulehner, W. (eds.): Domain Decomposition Methods in Science and Engineering XVII, *Lect. Notes Comput. Sci. Eng.*, Vol. 60. Springer, Berlin (2008). (St. Wolfgang/Strobl, July 2006)
259. Le Borne, S., Grasedyck, L., Kriemann, R.: Domain decomposition based \mathcal{H} -LU preconditioners. In: Widlund and Keyes [399], pp. 661–668. (New York, Jan. 2005)
260. Lebedev, V.I.: On a Zolotarev problem in the method of alternating directions. *USSR Comput. Math. Math. Phys.* **17,2**, 58–76 (1971)
261. Lebedev, V.I., Finogenov, S.A.: On the order of choice of the iteration parameters in the Chebyshev cyclic iteration method. *USSR Comput. Math. Math. Phys.* **11,2**, 155–170 (1971)
262. Lebedev, V.I., Finogenov, S.A.: On the order of choice of the iteration parameters in the Chebyshev cyclic iteration method. *USSR Comput. Math. Math. Phys.* **13,1**, 21–41 (1973)
263. Liebmann, K.O.H.: Die angenäherte Ermittlung harmonischer Funktionen und konformer Abbildung. *Sitzungsberichte der Mathematisch-Naturwissenschaftlichen Abteilung der Bayerischen Akademie der Wissenschaften zu München* **47**, 385–416 (1918)

264. Liesen, J., Saylor, P.E.: Orthogonal Hessenberg reduction and orthogonal Krylov subspace bases. *SIAM J. Numer. Anal.* **42**, 2148–2158 (2005)
265. Liesen, J., Strakos, Z.: *Krylov Subspace Methods, Principles and Analysis*. Oxford University Press, Oxford (2013)
266. Lions, P.L.: On the Schwarz alternating method I. In: Glowinski et al. [150], pp. 1–42
267. Loève, M.: *Probability Theory II*, 4th ed. Springer, New York (1978)
268. MacLachlan, S.P., Oosterlee, C.W.: Algebraic multigrid solvers for complex-valued matrices. *SIAM J. Sci. Comput.* **30**, 1548–1571 (2008)
269. Mandel, J.L.: A multilevel iterative method for symmetric, positive definite problems. *Appl. Math. Optim.* **11**, 77–95 (1984)
270. Mandel, J.L.: On block diagonal and Schur complement preconditioning. *Numer. Math.* **58**, 79–93 (1990)
271. Mandel, J.L. (ed.): *The Tenth International Conference on Domain Decomposition Methods*. AMS, Providence (1998). (Boulder, Aug. 1997)
272. Manteuffel, T.A.: Adaptive procedure for estimating parameters for the nonsymmetric Tchebychev iteration. *Numer. Math.* **31**, 183–208 (1978)
273. Manteuffel, T.A.: An incomplete factorization technique for positive definite linear systems. *Math. Comp.* **34**, 473–497 (1980)
274. Marchuk, G.: Splitting and alternating direction methods. In: Ciarlet–Lions [96], pp. 197–462
275. Marek, I.: Iterative methods for solving linear systems with a rectangular matrix. Report 8132, Universiteit Nijmegen (1981)
276. Mathew, T.P.A.: *Domain Decomposition Methods for the Numerical Solution of Partial Differential Equations*. Springer, Berlin (2008)
277. Matthies, H.G., Zander, E.: Solving stochastic systems with low-rank tensor compression. *Linear Algebra Appl.* **436**, 3819–3838 (2012)
278. McCormick, S.F. (ed.): *Multigrid Methods*. SIAM, Philadelphia (1987)
279. Meijerink, J.: Iterative methods for the solution of linear equations based on incomplete factorisation of the matrix. Shell Publ. 643, Rijswijk (1983). Published as SPE Conference Paper, doi:10.2118/12262-MS
280. Meijerink, J., van der Vorst, H.A.: An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix. *Math. Comp.* **31**, 148–162 (1977)
281. Meis, T., Marcowitz, U.: *Numerische Behandlung partieller Differentialgleichungen*. Springer, Berlin (1978)
282. Meis, T., Marcowitz, U.: *Numerical Solution of Partial Differential Equations*. Springer, New York (1981)
283. Meurant, G.A.: *The Lanczos and Conjugate Gradient Algorithms: from Theory to Finite Precision Computations*. SIAM, Philadelphia (2006)
284. Mohlenkamp, M.J.: Musings on multilinear fitting. *Linear Algebra Appl.* **438**, 834–852 (2013)
285. Morris, J.L. (ed.): *Symposium on the Theory of Numerical Analysis, Lect. Notes Math.*, Vol. 193. Springer, Berlin (1971). (Dundee, Sep. 1970)
286. Mróz, M.: Domain decomposition method for elliptic mixed boundary value problems. *Computing* **42**, 45–59 (1989)
287. Natterer, F.: *The Mathematics of Computerized Tomography*. J. Wiley and Teubner, Stuttgart (1986)
288. Nepomnyaschikh, S.V.: Domain decomposition and Schwarz methods in a subspace for the approximate solution of elliptic boundary value problems [in Russian]. Doctoral thesis, University Novosibirsk (1986)
289. Neumaier, A., Varga, R.S.: Exact convergence and divergence domains for the symmetric successive overrelaxation iterative (SSOR) method applied to H-matrices. *Linear Algebra Appl.* **58**, 261–272 (1984)
290. Nicolaides, R.: On multiple grid and related techniques for solving discrete elliptic systems. *J. Comput. Phys.* **19**, 418–431 (1975)
291. Niethammer, W.: Relaxation bei komplexen Matrizen. *Math. Z.* **86**, 34–40 (1964)

292. Niethammer, W.: Relaxation bei nichtsymmetrischen Matrizen. *Math. Z.* **85**, 319–327 (1964)
293. Niethammer, W.: The SOR method on parallel computers. *Numer. Math.* **56**, 247–254 (1989)
294. Niethammer, W., Varga, R.S.: The analysis of k -step iterative methods for linear systems from summability theory. *Numer. Math.* **41**, 177–206 (1983)
295. Oertel, K.D., Stüben, K.: Multigrid with ILU-smoothing: systematic tests and improvements. In: Hackbusch [185], pp. 188–199
296. O’Leary, D.P., White, R.E.: Multi-splitting of matrices and parallel solution of linear systems. *SIAM J. Alg. Disc. Meth.* **6**, 630–640 (1985)
297. Opfer, G., Schober, G.: Richardson’s iteration for nonsymmetric matrices. *Linear Algebra Appl.* **58**, 343–361 (1984)
298. Ortega, J.M.: *Introduction to Parallel Vector Solution of Linear Systems*. Plenum Press, New York (1988)
299. Oseledets, I.V.: DMRG approach to fast linear algebra in the TT-format. *Comput. Methods Appl. Math.* **11**, 382–393 (2011)
300. Oseledets, I.V., Tyrtyshnikov, E.E.: TT-cross approximation for multidimensional arrays. *Linear Algebra Appl.* **432**, 70–88 (2010)
301. Ostrowski, A.M.: Über die Determinanten mit überwiegender Hauptdiagonale. *Comment. Math. Helv.* **10**, 69–96 (1937)
302. Ostrowski, A.M.: On the linear iteration procedures for symmetric matrices. *Rend. Math. Appl.* **14**, 140–163 (1954)
303. Oswald, P.: On function spaces related to finite element approximation theory. *Z. Anal. Anwendungen* **9**, 43–64 (1990)
304. Oswald, P.: Multilevel finite element approximation. *Teubner Skripten zur Numerik*. Teubner, Stuttgart (1994)
305. Oswald, P.: On the convergence rate of SOR: a worst case estimate. *Computing* **52**, 245–255 (1994)
306. Paddon, D.J., Holstein, H. (eds.): *Multigrid Methods for Integral and Differential Equations*. Clarendon Press, Oxford (1985). (Bristol, Sep. 1983)
307. Paige, C.C., Saunders, M.A.: Solution of sparse indefinite systems of linear equations. *SIAM J. Numer. Anal.* **12**, 617–629 (1975)
308. Peaceman, D.W., Rachford, H.H.: The numerical solution of parabolic and elliptic differential equations. *J. SIAM* **3**, 28–41 (1955)
309. Pearcey, C.: An elementary proof of the power inequality for the numerical radius. *Michigan Math. J.* **13**, 289–291 (1966)
310. Penzl, T.: Low rank solution of data-sparse Sylvester equations. *Systems Control Lett.* **40**, 139–144 (2000)
311. Perron, O.: Zur Theorie der Matrices. *Math. Ann.* **64**, 248–263 (1907)
312. Proskurowski, W., Widlund, O.B.: On the numerical solution of Helmholtz’s equation by the capacitance matrix method. *Math. Comp.* **30**, 433–468 (1976)
313. Quarteroni, A., Périaux, J., Kuznetsov, Y.A., Widlund, O.B. (eds.): *Domain Decomposition Methods in Science and Engineering*. AMS, Providence (1994). (Como, Jun. 1992)
314. Quarteroni, A., Sacco, R., Saleri, F.: *Numerical Mathematics*, 2nd ed. Springer, Berlin (2007)
315. Quarteroni, A., Valli, A.: *Domain Decomposition Methods for Partial Differential Equations*. Oxford University Press, Oxford (1999)
316. Reich, E.: On the convergence of the classical iterative procedures for symmetric matrices. *Ann. Math. Statist.* **20**, 448–451 (1949)
317. Reid, J.K.: A method for finding the optimum successive overrelaxation factor. *Comput. J.* **9**, 200–204 (1966)
318. Reid, J.K. (ed.): *Large Sparse Sets of Linear Equations*. Academic Press, New York (1971). (Oxford, Apr. 1970)
319. Reid, J.K.: On the method of conjugate gradients for the solution of large sparse systems of linear equations. In: Reid [318], pp. 231–254
320. Reiter, S., Vogel, A., Heppner, I., Rupp, M., Wittum, G.: A massively parallel geometric multigrid solver on hierarchically distributed grids. *Comput. Vis. Sci.* **16**, 151–164 (2013)

321. Reusken, A.: Steplength optimization and linear multigrid methods. *Numer. Math.* **58**, 819–838 (1991)
322. Reusken, A.: On maximum norm convergence of multigrid methods for two-point boundary value problems. *SIAM J. Numer. Anal.* **29**, 1569–1578 (1992)
323. Reusken, A.: The smoothing property for regular splittings. In: Hackbusch and Wittum [208], pp. 130–138
324. Richardson, L.F.: The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam. *Philosophical Transactions of the Royal Society of London, Series A* **210**, 307–357 (1910)
325. Richardson, L.F., Gaunt, A.: The deferred approach to the limit. *Philosophical Transactions of the Royal Society of London, Series A* **226**, 299–361 (1927)
326. Riesz, F., Sz.-Nagy, B.: *Vorlesungen über Funktionalanalysis*, 4th ed. VEB Deutscher Verlag der Wissenschaften, Berlin (1982)
327. Runge, C.D.T.: Ueber die numerische Auflösung von Differentialgleichungen. *Math. Ann.* **46**, 167–178 (1895)
328. Saad, Y.: *Iterative Methods for Sparse Linear Systems*, 2nd ed. SIAM, Philadelphia (2003)
329. Saad, Y., Schultz, M.H.: A generalized minimal residual method for solving nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.* **7**, 856–869 (1986)
330. Samarskii, A.A., Nikolaev, E.S.: *Numerical Methods for Grid Equations*, Vol. II, Iterative Methods. Birkhäuser, Basel (1989)
331. Sauter, S.A., Schwab, C.: *Boundary Element Methods, SSCM*, Vol. 39. Springer, Berlin (2011)
332. Savas, B., Eldén, L.: Krylov-type methods for tensor computations I. *Linear Algebra Appl.* **438**, 891–918 (2013)
333. Schröder, J., Trottenberg, U.: Reduktionsverfahren für Differenzengleichungen I. *Numer. Math.* **22**, 37–68 (1973)
334. Schultz, M. H. (ed.): *Elliptic Problem Solvers*. Academic Press, New York (1981). (Santa Fe, June 1980)
335. Schwab, C., Todor, R.A.: Karhunen-Loëve approximation of random fields by generalized fast multipole methods. *J. Comput. Phys.* **217**, 100–122 (2006)
336. Schwarz, H. A.: Über einen Grenzübergang durch alternierende Verfahren. *Vierteljahrsschrift der Naturforschenden Gesellschaft in Zürich* **15**, 272–286 (1870)
337. Seidel, P.L.: Über ein Verfahren, die Gleichungen, auf welche die Methode der kleinsten Quadrate führt, sowie lineare Gleichungen überhaupt, durch successive Annäherung aufzulösen. *Abhandlungen der Mathematisch-Physikalischen Klasse der Königlich Bayerischen Akademie der Wissenschaften* **11**, 81–108 (1874)
338. Shaidurov, V.V.: *Multigrid Methods for Finite Elements*. Kluwer Academic Publ., Dordrecht (1995)
339. Sheldon, J.: On the numerical solution of elliptic difference equations. *Math. Tables and Other Aids to Computation* **9**, 101–112 (1955)
340. Singh, S.P., Burry, J.W.H., Watson, B. (eds.): *Approximation Theory and Spline Functions*. Reidel Publ., Dordrecht (1984). (Newfoundland, Aug./Sep. 1983)
341. Skeel, R.D.: Iterative refinement implies numerical stability for Gaussian elimination. *Math. Comp.* **35**, 817–832 (1980)
342. Skeel, R.D.: Effect of equilibration on residual size of partial pivoting. *SIAM J. Numer. Anal.* **18**, 449–454 (1981)
343. Smith, B.F., Bjørstad, P.E., Gropp, W.: *Domain Decomposition*. Cambridge University Press, Cambridge (1996)
344. Sonneveld, P.: CGS, a fast Lanczos-type solver for nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.* **10**, 36–52 (1989)
345. Sonneveld, P., Wesseling, P., de Zeeuw, P.M.: Multigrid and conjugate gradient methods as convergence acceleration techniques. In: Paddon and Holstein [306], pp. 117–168
346. Southwell, R.V.: Stress-calculation in frameworks by the method of “systematic relaxation of constraints”, parts I–II. *Proc. Roy. Soc. Edinburgh Sect. A* **151**, 56–95 (1935)

347. Southwell, R.V.: Stress-calculation in frameworks by the method of “systematic relaxation of constraints”, part III. *Proc. Roy. Soc. Edinburgh Sect. A* **153**, 41–76 (1935)
348. Southwell, R.V.: Relaxation methods in engineering science – a treatise on approximate computation. Oxford University Press, London (1940)
349. Southwell, R.V.: Relaxation methods in theoretical physics. Clarendon Press, Oxford (1946)
350. Starke, G.: Optimal ADI parameter for nonsymmetric systems of linear equations. *SIAM J. Numer. Anal.* **28**, 1431–1445 (1991)
351. Starke, G., Niethammer, W.: SOR for $AX - XB = C$. *Linear Algebra Appl.* **154**, 355–375 (1991)
352. Stein, P., Rosenberg, R. L.: On the solution of linear simultaneous equations by iteration. *J. London Math. Soc.* **23**, 111–118 (1948)
353. Stiefel, E.: Über einige Methoden der Relaxationsrechnung. *Z. Angew. Math. Phys.* **3**, 1–33 (1952)
354. Stiefel, E.: Relaxationsmethoden bester Strategie zur Lösung linearer Gleichungssysteme. *Comment. Math. Helv.* **29**, 157–179 (1955)
355. Stoer, J.: Solution of large systems of linear equations by conjugate gradient type methods. In: Bachem et al. [22], pp. 540–565
356. Stone, H.L.: Iterative solution of implicit approximations of multi-dimensional partial differential equations. *SIAM J. Numer. Anal.* **5**, 530–558 (1968)
357. Strakos, Z.: On the real convergence rate of the conjugate gradient method. *Linear Algebra Appl.* **154**, 535–549 (1991)
358. Strang, G.: Variational crimes in the finite element method. In: Aziz [20], pp. 689–710
359. Stüben, K.: Algebraic multigrid (AMG): experiences and comparisons. *Appl. Math. Comput.* **13**, 419–451 (1983)
360. Stüben, K.: A review of algebraic multigrid. *J. Comput. Appl. Math.* **128**, 281–309 (2001)
361. Tanabe, K.: Projection method for solving a singular system of linear equations and its applications. *Numer. Math.* **17**, 203–214 (1971)
362. Tobler, C.: Low-rank tensor methods for linear systems and eigenvalue problems. Doctoral thesis, ETH Zürich (2012)
363. Todd, J.: Applications of transformation theory: a legacy from Zolotarev (1847–1878). In: Singh et al. [340], pp. 207–245
364. Todd, J.: A legacy from E. I. Zolotarev (1847–1878). *Math. Intelligencer* **10**, 50–53 (1988)
365. Törnig, W.: Numerische Mathematik für Ingenieure und Physiker, Band 1: Numerische Methoden der Algebra. Springer, Berlin (1979)
366. Toselli, A., Widlund, O.B.: Domain Decomposition Methods – Algorithms and Theory, *SCM*, Vol. 34. Springer, Berlin (2005)
367. Trottenberg, U., Oosterlee, C.W., Schuller, A.: Multigrid. Academic Press, San Diego (2001)
368. van der Sluis, A.: Condition numbers and equilibrium matrices. *Numer. Math.* **14**, 14–23 (1969)
369. van der Sluis, A.: Condition, equilibration and pivoting in linear algebraic systems. *Numer. Math.* **15**, 74–86 (1970)
370. van der Sluis, A., van der Vorst, H.A.: The rate of convergence of conjugate gradients. *Numer. Math.* **48**, 543–560 (1986)
371. van der Vorst, H.A.: A vectorizable variant of some ICCG methods. *SIAM J. Sci. Statist. Comput.* **3**, 350–356 (1982)
372. van der Vorst, H.A.: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.* **13**, 631–644 (1992)
373. van der Vorst, H.A.: Iterative Krylov Methods for Large Linear Systems. Cambridge University Press (2003)
374. Varga, R.S.: Factorization and normalized iterative methods. In: Langer [257], pp. 121–142
375. Varga, R.S.: Matrix Iterative Analysis. Prentice-Hall, Englewood Cliffs (1962)
376. Varga, R.S.: Geršgorin and his Circles. Springer, Berlin (2004)
377. Vassilevski, P.: Multilevel preconditioning matrices and multigrid V-cycle methods. In: Hackbusch [185], pp. 200–208

378. Vassilevski, P.: Multilevel Block Factorization Preconditioners. Springer, New York (2008)
379. Verfürth, R.: A Review of a Posteriori Error Estimation and Adaptive Mesh-refinement Techniques. J. Wiley and Teubner, Stuttgart (1996)
380. Verfürth, R.: A Posteriori Error Estimation Techniques for Finite Element Methods. Oxford University Press, Oxford (2013)
381. von Mises, R., Pollaczek-Geiringer, H.: Praktische Verfahren der Gleichungsauflösung. ZAMM **9**, 58–77 (1929)
382. Wachspress, E.L.: Optimum alternating-direction-implicit iteration parameters for a model problem. J. Soc. Indust. Appl. Math. **10**, 339–350 (1962)
383. Wachspress, E.L.: The ADI Model Problem. Springer, New York (2013)
384. Wachspress, E.L., Haberler, G.J.: An alternating-direction-implicit iteration technique. J. Soc. Indust. Appl. Math. **8**, 403–424 (1960)
385. Wagner, C.: Tangential frequency filtering decompositions for symmetric matrices. Numer. Math. **78**, 119–142 (1997)
386. Wagner, C.: Tangential frequency filtering decompositions for unsymmetric matrices. Numer. Math. **78**, 143–163 (1997)
387. Wagner, C., Wittum, G.: Adaptive filtering. Numer. Math. **78**, 305–328 (1997)
388. Walker, H.: Implementation of the GMRES method using Householder transformations. SIAM J. Sci. Statist. Comput. **9**, 152–163 (1988)
389. Watson, G.A. (ed.): Numerical Analysis, *Lect. Notes Math.*, Vol. 506. Springer, Berlin (1976). (Dundee, July 1975)
390. Weiler, W., Wittum, G.: Parallel frequency filtering. Computing **58**, 303–316 (1997)
391. Weissinger, J.: Über das Iterationsverfahren. ZAMM **31**, 245–246 (1951)
392. Weissinger, J.: Verallgemeinerungen des Seidelschen Iterationsverfahrens. ZAMM **53**, 155–163 (1953)
393. Wesseling, P.: A robust and efficient multigrid method. In: Hackbusch and Trottenberg [205], pp. 614–630
394. Wesseling, P.: Theoretical and practical aspects of a multigrid method. SIAM J. Sci. Statist. Comput. **3**, 387–407 (1982)
395. Wesseling, P.: An Introduction to Multigrid Methods. Wiley, Chichester (1991)
396. Widlund, O.B.: A Lanczos method for a class of nonsymmetric systems of linear equations. SIAM J. Numer. Anal. **15**, 801–812 (1978)
397. Widlund, O.B.: Iterative substructuring methods: algorithms and theory for elliptic problems in the plane. In: Glowinski et al. [150], pp. 113–128
398. Widlund, O.B.: Optimal iterative refinement methods. In: Chan et al. [93], pp. 114–125
399. Widlund, O.B., Keyes, D.E. (eds.): Domain Decomposition Methods in Science and Engineering XVI, *Lect. Notes Comput. Sci. Eng.*, Vol. 55. Springer, Berlin (2007). (New York, Jan. 2005)
400. Wittum, G.: Distributive Iterationen für indefinit Systeme als Glätter im Mehrgitterverfahren am Beispiel der Stokes- und Navier-Stokes-Gleichungen mit Schwerpunkt auf unvollständigen Zerlegungen. Doctoral thesis, Universität zu Kiel (1986)
401. Wittum, G.: On the robustness of ILU-smoothing. In: Hackbusch [185], pp. 217–239
402. Wittum, G.: Linear iterations as smoothers in multigrid methods: theory with applications to incomplete decompositions. Impact Comput. Sci. Eng. **1**, 180–215 (1989)
403. Wittum, G.: On the robustness of ILU smoothing. SIAM J. Sci. Statist. Comput. **10**, 699–717 (1989)
404. Wittum, G.: An ILU-based smoothing correction scheme. In: Hackbusch [187], pp. 228–240
405. Wittum, G.: Filternde Zerlegungen. Teubner Skripten zur Numerik. Teubner, Stuttgart (1992)
406. Wittum, G., Liebau, F.: On truncated incomplete decompositions. BIT **29**, 719–740 (1989)
407. Xu, J.: Theory of multilevel methods. Ph.D. thesis, Penn State University (1989)
408. Xu, J.: A new class of iterative methods for nonselfadjoint or indefinite problems. SIAM J. Numer. Anal. **29**, 303–319 (1992)
409. Xu, J.: Iterative methods by space decomposition and subspace correction. SIAM Rev. **34**, 581–613 (1992)

410. Xu, J., Zikatnov, L.: Algebraic multigrid methods. *Acta Numerica* **26** (2017). To appear
411. Young, D.M.: Iterative methods for solving partial differential equations of elliptic type. Ph.D. thesis, Harvard University (1950)
412. Young, D.M.: Iterative Solution of Large Linear Systems. Academic Press, Orlando (1971)
413. Young, D.M.: A historical overview of iterative methods. *Comput. Phys. Comm.* **53**, 1–17 (1989)
414. Young, D.M., Huang, R.: Some notes on complex successive overrelaxation. Report CNA-185, University of Texas at Austin (1983)
415. Yserentant, H.: Hierarchical bases of finite element spaces in the discretization of non-symmetric elliptic boundary value problems. *Computing* **35**, 39–49 (1985)
416. Yserentant, H.: On the multi-level splitting of finite element spaces. *Numer. Math.* **49**, 379–412 (1986)
417. Yserentant, H.: Two preconditioners based on the multi-level splitting of finite element spaces. *Numer. Math.* **58**, 163–184 (1990)

List of authors involved in the references above, but not placed as first author.

- | | | |
|------------------------------------|---------------------------------|-------------------------------------|
| Barker, V.A. [13] | Grötschel, M. [22] | Meurant, G.A. [99, 117, 150, |
| Bastian [34] | Gropp, W. [343] | 151, 236] |
| Benner, P. [36] | Habetler, G. J. [384] | Mises, R. von <i>see:</i> von Mises |
| Bjørstadt, P.E. [254, 343] | Hackbusch, W. [34, 39, 64–68, | Neumann, M. [213] |
| Böhmer, K. [4] | 100, 113, 123–125, 161–164] | Nelson, C.W. [90] |
| Boor, C. de <i>see</i> de Boor, C. | Halpern, L. [122, 109] | Niethammer, W. [118, 119, |
| Breiten, T. [40] | Heppner, I. [320] | 120, 213, 351] |
| Brinkkemper, S. [14] | Holst, M.J. [31] | Nikolaev, E.S. [330] |
| Burrus, C.S. [215] | Holstein, H. [306] | Nouy, A. [126] |
| Burry, J.W.H. [340] | Hoppe, R.H.W. [106, 239] | O’Leary, D.P. [155] |
| Chan, T.F. [26, 236] | Huang, R. [414] | Oosterlee, C.W. [268, 367] |
| Cross, M. [254] | Il’in, V.P. [14] | Oswald, P. [170] |
| Dahlquist, G. [248] | Johnson, D.H. [215] | Overton, M.L. [156] |
| de Groen, P. [38] | Kako, T. [95] | Pasciak, J.E. [70–76] |
| De Moor, B. [103] | Kawarada, H. [95] | Pavarino, L.F. [109] |
| de Zeeuw, P.M. [345] | Keyes, D.E. [50, 106, 217, 258, | Périaux, J. [93, 94, 106, 150, |
| Deuflhard, P. [56] | 399] | 151, 153, 239, 313] |
| Discacciati, M. [258] | Khachatryan, A. [123] | Pichot, G. [122] |
| Douglas, C.C. [27] | Khoromskij, B. [163, 202, 203] | Pironneau, J. [95, 239] |
| Dryja, M. [64] | Kornhuber, R. [43, 225] | Plemmons, R.J. [45, 46] |
| Dupont, T.F. [28, 29, 30] | Korte, B. [22] | Pollaczek-Geiringer, H. [381] |
| Eldén, L. [332] | Krause, R. [109] | Polman, B. [15] |
| Erisman, A.M. [116] | Kriemann, R. [164, 165, 166, | Praetorius, D. [129] |
| Espedal, M.S. [50] | 202, 259, 246] | Rachford, H.H. [308] |
| Faermann, B. [100] | Kunoth, A. [101, 107] | Reid, J.K. [116] |
| Finogenov, S. A. [261, 262] | Kuznetsov, Y. [106, 151, 313] | Reimer, T. [55] |
| Freund, R.W. [134, 135] | Landsberg, J.M. [83] | Reusken, A. [204] |
| Gander, M.J. [43, 122, 109] | Le Borne, S. [165, 166] | Rice, J.R. [102] |
| Garbey, M. [106] | Liebau, F. [406] | Riemsdagh, K. [108] |
| Gaunt, A. [325] | Lions, J.L. [96, 152] | Rohwedder, T. [125, 224] |
| Glashoff, K. [6] | Litvinenko, A. [124] | Rose, D.J. [86] |
| Glowinski, R. [93, 94] | Malik, G.M. [228] | Rosenberg, R.L. [352] |
| Golub, G.H. [90, 98, 99, 150] | Mandel, J.L. [51] | Roux, F.X. [127] |
| Graham, I.G. [100] | Manteuffel, T.A. [7] | Rupp, M. [320] |
| Grasedyck, L. [24, 259] | Marcowitz, U. [281, 282] | Ruttan, A. [118] |
| Griegorieff, R.D. [85] | Matthies, H.G. [124] | Sacco, R. [314] |
| Groen, P. de [38] | Melenk, M. [129] | Saleri, F. [314] |

- Sassi, T. [122]
Saunders, M.A. [307]
Sauter, S.A. [100]
Saylor, P.E. [7, 264]
Schatz, A.H. [70, 71, 72, 73]
Schober, G. [297]
Schneider, R. [125, 224]
Schröder, J. [85]
Schuller, A. [367]
Schultz, M.H. [329]
Schwab, C. [331]
Scott, L.R. [32, 82]
Scroggs, J.S. [236]
Shi, Z.C. [153]
Stiefel, E. [219]
Strakos, Z. [265]
Strauss, E.G. [139]
Stüben, K. [295]
Süli, E. [229]
Sz.-Nagy, B. [326]
- Szyld, D.B. [143]
Tobler, C. [5, 242, 243, 244]
Todor, R.A. [335]
Trottenberg, U. [68, 205, 206,
207, 333]
Tůma, M. [42]
Tyrtysnikov, E.E. [203, 300]
Valli, A. [315]
Vandewalle, J. [103]
Varga, R. [3, 88, 89, 119, 120,
289, 294]
Vassilevski, P. [16, 17, 18, 19]
van der Vorst, H.A. [280, 370]
Van Loan, C.F. [157]
Vierendeels, J. [108]
Vogel, A. [320]
Voigt, R.G. [236]
Vorst *see* van der Vorst, H.A.
Wähnert, P. [124]
Wang, J. [74]
- Watson, B. [340]
Werner, B. [6]
Wesseling, P. [216, 345]
White, R.E. [296]
Widlund, O.B. [31, 43, 52, 53,
92, 93, 94, 114, 115, 122, 151,
153, 217, 225, 239, 254, 258,
312, 313, 366]
Wittum, G. [34, 91, 208, 209,
210, 320, 387, 390]
Xu, J. [31, 74–76, 225, 237, 239]
Yates, R. [217]
Young, D.M. [212]
Yserentant, H. [30, 57]
Zander, E. [277]
Zeeuw, P.M. *de see:* de Zeeuw
Zhen, M. [4]
Zhou, X. [241]
Zikatnov, L. [410]
Zulehner, W. [258]

Index

- ADI method, 201
 - commutative case, 205
 - convergence, 203, 204, 207
 - cost, 209
 - cyclic, 111, 207
 - instationary, 204
 - iteration matrix, 203
 - numerical example, 209
- ADI parameter, 205, 314
 - optimal, 206
- adjoint problem, 481
- admissibility, η -, 463
- affine subspace, 179
- agglomeration, 471
- algebra, 419
- algebraic reconstruction technique, 117
- algorithm, *see* method or iteration
 - Buneman, 12
 - cascade, 289
- ALS, 400
- alternate-triangular method, 103
- alternating least squares, 400
- anisotropy, 10, 37
- approximation property, 298, 303, 308, 309
- Arnoldi basis, 259
- Arnoldi method, 261
- ART, 117
- asymptotically smooth, 468
- backward substitution, 372
- band width, 11, 402
- basic iteration, 175, 192, 200, 216, 219, 220
 - damped, 181
- basis
 - Arnoldi, 259
 - finite element, 358, 475
 - hierarchical, 366, 368
- hierarchical, 366
- orthogonal, 226, 256
- orthonormal, 276, 409
- BCG, BiCG, Bi-CGSTAB, 262
- biconjugate gradient method, 262
- bilinear form, 474
 - equivalent, 170
- bisection
 - cardinality-based, 461
 - geometric by bounding boxes, 461
 - regular geometric, 460
- block, 6
 - admissible, 463
- block cluster tree, 462
 - level conserving, 380
- block index set, 407
 - ordered, 407
- block matrix, 6, 407, 408
- block structure, 6, 42, 69, 163, 337, 407
 - chequer-board, 7, 71
 - lexicographical, 7
 - plane-wise, 105
 - row, 105
 - zebra, 71
- block vector, 407
- block-diagonal matrix, 42, 44, 408
- block-diagonal part, 70, 408, 433
- block-tridiagonal matrix, 7, 79, 318, 408
- boundary condition
 - Dirichlet, 4, 328, 474
 - natural, 477
 - Neumann, 140, 329
- boundary element method, 15, 377
- boundary value problem, 4, 325
- bounding box, 460, 464
- BPX method, 369
- breakdown, 238

- lucky, 220
- Buneman algorithm, 12, 326
- canonical format, 390
- capacitance matrix method, 333
- cardinality, 8
- cascade algorithm, 289
- Cauchy–Schwarz inequality, strengthened, 343, 347, 359, 365
- CG method, 234, 235, 237, 398
 - applied to a basic iteration, 241
 - applied to the multigrid iteration, 315
 - convergence, 238–240, 242
 - cost, 245
 - numerical example, 244
 - restarted, 238
- CGS, 262
- Chebyshev method, 111, 187, 190–192, 196–198, 218, 239, 246, 314
 - convergence, 192
 - order improvement, 192
- Chebyshev polynomial, 187, 188–190, 313
- Cholesky decomposition, 12, 122, 219, 371, 434
 - incomplete, 150
- Cimmoni iteration, 118
 - convergence, 119
 - iteration matrix, 118
- cluster tree, 459
 - ternary, 379
- coarse-grid correction, 272, 273, 280, 298, 317, 358, 360
 - damped, 315
- coarse-grid equation, 272, 273, 281, 317, 369
- coarse-grid matrix, 273
- coarsening of a block structure, 464
- coefficient vector, 476
- collocation, 467
- comparison theorem, 142
- composed iteration, 106–108, 246, 327
- computational work, 30
 - ADI method, 209
 - CG method, 245
 - effective, 31, 32, 78, 87, 107, 111, 195, 201, 208, 209, 245, 285
 - frequency filtering method, 319
 - Gauss elimination, 11
 - Gauss–Seidel iteration, 45–47, 87
 - Jacobi iteration, 45–47
 - multigrid iteration, 284–286
 - multigrid iteration of the second kind, 316
 - nested iteration, 291
 - Richardson iteration, 46, 47
 - semi-iteration, 195, 196
- SOR iteration, 45–47, 87
- SSOR iteration, 133
- condition, 167, 217, 221, 223, 391, 421
 - Euclidean, 421
- conjugate gradient method, *see* CG method
 - as smoother, 314
- conjugate gradient squared method, 262
- conjugate residuals, *see* CR method
- conjugate vectors, 226, 235, 242
- consistency, 19, 20
 - semi-iterative, 176
 - test, 32
- consistency error, 5
- contraction number, 25, 26, 58, 59, 280, 301
- convection, 10
- convection-diffusion equation, 283
- convergence, 19, 24
 - Chebyshev method, 192
 - CR method, 254
 - Gauss–Seidel iteration, 56, 77, 78, 83, 86, 144, 145
 - ILU iteration, 159
 - Jacobi iteration, 55, 62, 76, 78, 144, 145
 - Kaczmarz iteration, 118
 - Landweber iteration, 114
 - linear, 26
 - order of, *see* convergence order
 - monotone, 25, 55, 128, 129, 131, 132, 214
 - multigrid iteration, 293, 301, 303, 305, 310–312, 364
 - positive definite iteration, 54
 - quadratic, 321
 - Richardson iteration, 47–53, 63, 74, 75
 - Schwarz iteration, 351
 - additive, 347
 - multiplicative, 349
 - secondary iteration, 108
 - SOR iteration, 56–62, 66, 67, 78–81, 83–85
 - SSOR iteration, 132, 133, 135, 146
 - test, 33
 - two-grid iteration, 276, 280, 301, 306, 308, 309, 315, 360, 361
 - V-cycle, 310, 364
 - W-cycle, 311
- convergence order, 32
 - improved, 66
 - optimal, 170
- convergence rate, 26, 28, 115, 142
 - ADI method, 204
 - asymptotic, 178, 192, 194, 197, 200, 217, 254, 255, 347
 - composed iteration, 110
 - Gauss–Seidel iteration, 62, 77, 78, 86
 - Jacobi iteration, 63, 76

- optimal, 48, 49, 63, 75, 77, 86, 126, 204
order, 32, 170
Richardson iteration, 48, 74
SOR iteration, 81, 86
two-grid iteration, 280
V-cycle, 310
convex hull, 49
core tensor, 394
cost factor, 30, 46, 47, 133, 153, 195
CR method, 250
convergence, 254
numerical example, 255
stabilised, 252
cross-point, 332
- damping, 95, 108, 130, 196, 307, 315, 338
optimal, 109, 110, 114, 221, 223
- decomposition
frequency filtering, 318
LU, *see* LU decomposition
QR, 410
UL of the inverse matrix, 381
- defect, 22, 24, 28, 213, 272
- defect correction, 21
- diagonal block, 43
- diagonal dominance, 145, 158, 160, 161, 443
essential, 444
irreducible, 443
strict, 146, 443
strong, 105
weak, 162, 298, 443
- diagonal part, 402, 433
- diagonalisability, 414
real, 414, 434
simultaneous, 205
- diameter of a cluster, 463
- differential equation, 104, 291, 292, 474
elliptic, 446
ordinary, 171
parabolic, 201, 383, 399
partial, 262
- direct method, direct solver, 3, 10–12, 28, 161, 237, 316, 326, 382
- direct sum, 410
- Dirichlet boundary condition, 4, 328, 474
- Dirichlet–Neumann method, 329
- discrete regularity, 299
- discretisation, 5, 283, 296, 316, 325
finite element, 360, 476
Galerkin, 273, 300, 358, 475
- discretisation error, 20, 34, 287, 292
relative, 287
- dissection method, 378
- distance of two clusters, 463
- domain decomposition, 325, 378
nonoverlapping, 329, 332
overlapping, 327, 357
- domain of an iteration, 18
- dominance, diagonal, *see* diagonal dominance
- dual norm, 473
- dual space, 473
- eigenvalue, 35, 79, 403, 424, 432
generalised, 55
maximal, 35
minimal, 35
- eigenvector, 403, 412–415, 440, 443
orthonormal, 36
- enclosure of the solution, 145, 156
- energy norm, 54–57, 128, 190, 351, 434
- error
consistency, 5
discretisation, 20, 34
iteration, *see* iteration error
- Euclidean norm, 26, 417, 418, 422
- exponential sum, 392
- family of matrices, 10
- far field, 465
- fast Fourier transform, 334
- FETI, 331
- fictitious domains method, 334
- field, 8
far, 465
near, 465
- field of values, 427
- fill-in, 11, 149, 376
- finite element basis, 358
- finite element discretisation, 476
- finite elements, 476
shape regular, 15, 466, 479
- five-point formula, 5, 9, 30, 41, 72, 86, 105, 111, 152, 159, 201, 209, 360
- fixed point, 18, 19
- fixed-point equation, 156–158, 302, 316
- fixed-point iteration, 157, 158, 161, 162
- FOM, 261
- form
bilinear, 409, 474
coercive, 474
positive, 476
sesquilinear, 409, 474
symmetric, 474
- format
canonical, 390
model, 456
 r -term, 390
tensor, 390

- forward substitution, 372
 Fourier analysis, 276
 Fourier transform, fast, 334
 frequency filtering decomposition, 318
 Frobenius norm, 419, 420
 full orthogonalisation method, 261
- Galerkin discretisation, conforming, 475
 Galerkin product, 273, 304, 336, 362
 Gauss elimination, 3, 10, 11, 382, 450–452
 cost, 10
 Gauss, Carl Friedrich, 3
 Gauss–Seidel iteration, 3, 4, 12–14, 35, 39, 40,
 41, 56, 117, 118, 140, 295, 357, 368
 2-cyclic, 78, 296
 backward, 91, 103, 133
 block-, 44, 62, 86, 144, 340
 cost, 87
 chequer-board, 13, 40, 275, 360
 convergence, 56, 77, 78, 83, 86, 144, 145
 convergence rate, 62, 77, 78, 86
 cost, 45–47
 iteration matrix, 40
 lexicographical, 13, 40
 nonlinear, 322
 numerical example, 14, 87
 pointwise, 44
 symmetric, 103, 132, 134, 198, 296
 numerical result, 136
- Gelfand triple, 474
 generalised minimal residual method, 258
 Gershgorin circles, 447
 Givens rotation, 261
 GMRES, 258
 GMRES(m), 261
 gradient method, 215, 218, 225, 234
 applied to a basic iteration, 219–221
 biconjugate, 262
 convergence, 216, 217, 221–223, 240
 direct positive definite case, 223
 preconditioned, 220
 residual oriented, 222
- Gram matrix, 478
 graph
 directed, 435
 of a matrix, 149, 342, 435
 undirected, 435
- Green function, 469
 grid, 5, 9, 268
 grid function, 6, 9, 269, 270, 357
 grid point, 5
 grid size, 5, 11, 268, 269
- \mathcal{H} -LU decomposition, 371
- H-matrix, 145, 146, 155, 452
 Hadamard product, 121
 Helmholtz equation, 160, 193, 254, 255, 295
 Hessian matrix, 212
 hierarchical basis, 366
 hierarchical basis method, 366
 hierarchical matrix, 170, 316, 333, 371, 453,
 465
 LU decomposition of a, 375
 operation, 469
 storage, 465
 truncation, 470
- hierarchy
 grid, 318
 of subspaces, 299
 of systems of equations, 268, 269, 287
- ILU decomposition, 148
 approximative, 162
 blockwise, 163, 318
 existence, 155, 156
 modified, 154, 162, 256
 stability, 157, 163
 truncated, 162
- ILU iteration, 148
 convergence, 159
 damped, 160
 modified, 164, 317
 numerical example, 163
 with enlarged diagonal, 160
- index set, 7
 block, 407
 nonordered, 7
 ordered, 39, 40, 70, 402, 408, 411, 450
- initialisation, 30, 321
 instability, 157, 194, 251
 integral equation, 316
 integral operator, 467
 interpolation
 bilinear, 270
 linear, 270
 inverse estimate, 480
 inversion of a matrix, *see* matrix
 iterate, 23
 iteration, 3, 18, 382
 $\langle \text{name} \rangle$, *see* $\langle \text{name} \rangle$ iteration
 adjoint, 95, 339
 algebraic, 18, 38, 40, 41, 43, 140, 150, 382
 alternating-direction-implicit, *see* ADI
 basic, 175
 composed, 104, 106–108, 246, 327
 consistent, 19, 20, 22, 24, 25, 29, 216
 convergent, 19, 20, 24, 25, 29, 99
 cyclic, 194

- damped, 95
 - convergence, 126
- diagonally left/right-invariant, 120
- domain of an, 18
- extrapolated, 95
- identical, 98, 100
- inconsistent, 20
- invariant, 120
- k -step, 195
- linear, 21
- minimal residual, 228
- nested, *see* nested iteration
- nonexpansive, 138
- nonlinear, 216
- normal form of an
 - first, 21
 - second, 22, 23
 - third, 23
- one-step, 29
- positive definite, 54, 119, 124, 127, 208, 220, 221, 241, 242, 338
 - convergence, 54
 - directly, 94, 123, 222, 243
- positive semidefinite, 93
- projection, 93
 - A -orthogonal, 93
- robust, 10, 317
- secondary, 104–106, 110, 246, 262, 282
- semi-, *see* semi-iteration
- smoothing, 268
- subspace, 325, 336
- symmetric, 92, 101, 119, 134, 305, 339, 341, 361
- symmetrised, 101, 125, 129
- truncated, 398
- two-step, 29, 195
- iteration error, 24, 31, 34, 290
- iteration matrix, 21, 25
 - ADI method, 203
 - adjoint iteration, 90
 - composed iteration, 106
 - Gauss–Seidel iteration, 40
 - Jacobi iteration, 38
 - multigrid iteration, 286
 - product iteration, 99
 - Schwarz iteration, 338, 341
 - SOR iteration, 41
 - SSOR iteration, 132
 - two-grid iteration, 274
- Jacobi iteration, 35, 38, 76, 140, 151, 166, 357, 446
 - block-, 43, 56, 63, 64, 144, 340
 - convergence, 56, 62–64
 - semi-iterative, 198
 - convergence, 55, 62, 63, 76–78, 144, 145
 - cost, 45–47
 - damped, 76, 96, 118, 280, 296
 - iteration matrix, 38
 - nonlinear, 322
 - numerical examples, 65
 - semi-iterative, 198
- Jacobi, Carl Gustav, 4, 38
- Jordan block, 412
- Jordan normal form, 412
- Kaczmarz iteration, 116
 - as smoothing iteration, 283, 296
 - convergence, 116, 118
 - numerical example, 118
- Karhunen–Loëve expansion, 397
- kernel, 98
- Kronecker product, 385
- Kronecker symbol, 8
- Krylov space, 179, 256
- Lagrange parameter, 331
- Landau symbol, xxii
- Landweber iteration, 94
 - convergence, 114
- Laplace operator, 5
- least squares, 3, 94, 214, 423
- level conservation, 380
- level number, 268
- Liebmann method, 39
- LU decomposition, 11, 12, 44, 148, 151, 371, 472
 - cost, 376
 - existence, 154
 - incomplete, *see* ILU decomposition
- lucky breakdown, 220
- Lyapunov equation, 384, 385
- M-matrix, 104, 141, 144, 145, 152, 155, 204, 445, 446–452
- mass matrix, 478
- matrix, 7
 - 2-cyclic, 70, 77, 78, 347
 - weakly, 69–77
 - adjoint, 401, 402, 435
 - band, 11, 12, 151
 - block-diagonal, 408
 - block-tridiagonal, 7, 408
 - coercive, 432
 - commutative, 202, 405, 415, 433
 - consistently ordered, 78
 - cyclic of index two, 70
 - diagonal, 402, 405, 406

- diagonalisable, 185, 414, 415
 - real, 414, 434
- family, 10, 168
- finite element, 376
- fully populated, 11, 15, 284, 435, 453
- Hermitian, **402**, 405, 408, 415, 432, 451
- Hermitian transposed, 401
- Hessenberg, 261
- hierarchical, *see* hierarchical matrix
- indefinite, 193, 253, 254, 283, 295
- inverse positive, 141, 142, 156, **445**
- inversion, 458
- irreducible, **436**, 437, 441–444, 446–449
- iteration, *see* iteration matrix
- M-, *see* M-matrix
- monotone, 445
- nonnegative, 438, 439
- normal, 204, **402**, 414, 415, 424, 427
- numerical radius, 26, 53, **427**
- of the first normal form, 21
- of the second normal form, **22**, 90, 99, 106, 151, 339
- of the third normal form, **23**, 41, 54, 55, 99, 106, 139
- orthogonal, 410
- positive, 438, **438**, 440, 447
- positive definite, 35, 36, 43, 47, 54, 56–59, 62, 197, 202, 203, 211, 213, 421, **431**, 432–435, 444, 451, 452
- positive definite Hermitian part, 55, 125, 128, 129, 147, **432**
- positive semidefinite, 203, 432, 434
- principal sub-, 154, 338, 407, 433, 442, 449
- rank- r , 455
- reducible, **436**, 443
- regular, 17, **402**, 433
- selfadjoint, 435
- similar, 403, 404, 411
 - unitarily, 403
- sparse, 11, 15, 45, 148
- spectrally equivalent, 168, 383
- Stieltjes, 451
- sub-, 407
- symmetric, 465
- symmetrisable, 262
- triangular, 12, 148, **402**, 405, 406, 411, 450
 - block-, 44, **408**
 - strictly, 12, 57, 78, 79, 150, 350, **402**, 405
- tridiagonal, 7, 44, 79, **402**
- unitary, **402**, 410, 411, 414, 415, 422
- weakly p -cyclic, 85, *see* matrix, 2-cyclic
- matrix block, 407
- matrix equation
 - Lyapunov, 384, 385
- Riccati, 384, 399
- Sylvester, 208, 399
- matrix exponential, 394
- matrix graph, *see* graph
- matrix inversion, 472
- matrix norm, **419**, 420, 425
 - associated, 419
 - corresponding, 25, **419**, 420, 424, 425, 434
 - Frobenius, *see* Frobenius norm
- matrix polynomial, 177, 184, 185, 236, 243, 313, 404, 405, 408, 413
- matrix-matrix addition, 470
- matrix-matrix addition, cost, 457
- matrix-matrix multiplication, 457, 471
 - cost, 458
- matrix-vector multiplication, 15, 469
- maximum norm, **417**, 418, 438
- maximum principle, 446
- method, *see* iteration
 - ADI, 201, *see* ADI
 - alternate-triangular, 103
 - Arnoldi, 261
 - biconjugate gradient, 262
 - boundary element, 377
 - BPX, 369
 - capacitance matrix, 333
 - Chebyshev, *see* Chebyshev method
 - conjugate gradient squared, 262
 - direct, *see* direct method
 - Dirichlet–Neumann, 329
 - FETI, 331
 - hierarchical basis, 366
 - mortar, 326
 - nested dissection, 333
 - Neumann–Dirichlet, 334
 - Neumann–Neumann, 335
 - Newton, 320
 - of fictitious domains, 334
 - of orthogonal directions, 256
 - of steepest descent, *see* gradient method
 - of the conjugate directions, 224
 - of the conjugate residuals, *see* CR method
 - of total reduction, 12
 - orthogonalisation, 409
 - Schur complement, 332
 - semi-iterative, *see* semi-iteration
- minimal residual iteration, 228
- minimal residual method, *see* MINRES
 - generalised, *see* GMRES
- minimum function, 180, 185, 413
- Minkowski, Hermann, 445
- MINRES, 261
- model format, 456
- model problem, *see* Poisson model problem

- mortar method, 326
- multi-splitting, 140
- multigrid iteration, **265**, **281**, 313, 359, 398
 - additive, 369, 399
 - algebraic, 317, 335
 - convergence, 293, 301, **303**, 310, 311, 364
 - cost, 284–286, 316
 - damped, 315
 - frequency decomposition, 369
 - history, 317
 - iteration matrix, 286
 - nonlinear, 292, 323
 - numerical example, 282
 - of the second kind, 316
 - symmetric, 304
- multilevel Schwarz iteration, 369
- multiplicity
 - algebraic, **403**, 404, 413
 - geometric, 27, **403**, 404, 412
- natural boundary condition, 477
- near field, 465
- nested dissection method, 333
- nested iteration, 172, 287, 289–292, 316
 - computational work, 291
 - nonlinear, 292, 321
- Neumann boundary condition, 140, 329
- Neumann's series, 426
- Neumann–Dirichlet method, 334
- Neumann–Neumann method, 335
- Newton method, 320
- nine-point formula, **9**, 40, 72, 319
- nodal point, 358, 460, 476
- norm, 417
 - <name>, *see* <name> norm
 - dual, 473
 - equivalent, 418
 - matrix, *see* matrix norm
 - operator, 419
 - submultiplicative, 406, 420, 428
- normal equations, 94, 423
- normal form
 - Jordan, 412
 - Schur, 411
- normality, $B-$, 261
- numerical radius, 26, 53, 107, **427**
- OD method, 256
 - stabilised, 257
- operator
 - associated to a bilinear form, 475
 - integral, 15
 - Laplace, 5
 - nonlocal, 15
- order improvement, 200
 - CG method, 239
 - Chebyshev method, 192, 198
 - modified ILU iteration, 160
 - smoothing factor, 314
 - SOR, 60, 84
 - SSOR iteration, 134
 - two-step iteration, 195
- ordering, 12, 38, 70, 372
 - backward, 91
 - chequer-board, 7, 13, 40, 134
 - four-colour, 40
 - lexicographical, **6**, 7, 11, 13, 43, 66, 136, 148
 - backward, 43
 - red-black, *see* ordering, chequer-board zebra, 43
- ORTHODIR, ORTHOMIN, ORTHORES, 262
- orthogonal polynomials, 187
- orthogonal space, 409
- orthogonal, orthonormal, 409
- orthogonalisation (method), 409
- Ostrowski, Alexander Markowitsch, 445
- overrelaxation method, *see* SOR iteration
- parabolic differential equations, 399
- parallel computation, 38, 40, 45, 162, 166, 326, 333, 344, 369, 381
- partition, 462
 - admissible, 464
- path, 436
 - length of a, 436
- pattern, **149**, 151, 152, 161
 - star, 151
- Perron–Frobenius theory, 440
- perturbation lemma, 295
- Picard iteration, 15, 316
- Poisson model problem, **4**, 9–11, **13**, 30–32, 34, 35, 43, 46, 64, 71, 86, 201, 204, 266, 268, 396, 408, 437, 444, 474
 - numerical example, 65, 66, 87, 88, 255
- polynomial
 - characteristic, 403, 408, 413
 - Chebyshev, 187
 - matrix, 404, 405
 - optimal, **184**, 185, 193, 194, 239, 314
 - orthogonal, 187
- post-, pre-smoothing, 274
- preconditioning, **165**, 166, 241, 335
 - diagonal, 166
- principal vector, 412
- product iteration, 93, **99**, 101, 117, 134, 135, **137**, 202, 208, 268, 274, 282, 318, 319, 338, 360, 365

- projection, 116, 273, 337, **410**
 - A -orthogonal, **138**, 337, 361
 - orthogonal, 308, **410**
- prolongation, 269, 270, 272, 336–338, 478
 - canonical, 299, 302, 360
 - matrix-dependent, 271
 - nine-point, 270
- QR decomposition, 410
- r -term format, 390
- rank, *see* representation rank
 - local, 465
 - representation, 390
 - tensor, 390
- reconstruction technique
 - algebraic, 117
 - simultaneous iterative, 94
- reduction factor, 14, **26**, 65
- regularity, 300, 361
 - 2m-, 480
 - weaker, 303, 311, 481
- relaxation, 4, 39
- relaxation parameter, **41**, 85
 - complex, 83
 - optimal, 81, 86
- representation rank, 390, 455
- residual, 179, 213
- restriction, 271, 336, 478
 - canonical, 299, 302
 - matrix-dependent, 272
 - nine-point, 272
 - trivial, 271
- Reusken’s lemma, 297
- Riccati equation, 384, 399
- Richardson extrapolation, 34
- Richardson iteration, 37, 217, 266, 280, 293–295
 - convergence, 47–53, 62, 63, 74, 75
 - cost, 46, 47
 - nonlinear, 322
 - numerical example, 65
 - semi-iterative, 197, 216, 238
 - smoothing property, 266
- Riesz isomorphism, 473
- row-sum norm, 298, 365, **420**, 438
- Runge-Kutta method, 172
- scalar product, **409**, 435
 - energy, 434
 - Euclidean, 409, 435
- Schur complement, **452**, 458
- Schur complement method, 332
- Schur normal form, 411
- Schwarz inequality, 422
- Schwarz iteration, 325, 338, 340
 - additive, 338, 340, 341, 344, 358
 - convergence, 347
 - iteration matrix, 338, 341
 - multilevel, 369
 - multiplicative, 338, 340, 361, 362
 - convergence, 349–351
- search direction, 212, 213
- Seidel, Phillip Ludwig, 4
- semi-iteration, 175, 181
 - consistent, 177, 178, 181
 - cost, 195
 - linear, 177
 - three-term recursion, 183
- separation rank, 467
- separator, 377
- sesquilinear form, 474
- seven-point formula, 104
- shape regularity, 15, 466, 479
- similarity transformation, 404
 - unitary, 411
- simultaneous iterative reconstruction technique
 - SIRT, 94
- singular value, 416
- singular value decomposition, 416
 - higher order (HOSVD), 394
- singular vector, 416
- smoothing iteration, **266**, 268, 272, 286, 317
 - Gauss–Seidel
 - chequer-board, 360
 - Jacobi, 368, 399
 - post-, 274
 - pre-, 274, 304, 362
 - Richardson, 399
 - semi-iterative, **313**
 - smoothing property, **293**, 294–298, 303, **307**
 - smoothing step, 293
 - smoothness, asymptotic, 468
 - SOR iteration, 4, 14, **41**, 145, 198, 204
 - backward, 103, 134
 - block-, 45, 86
 - convergence, 62, 66, 67
 - convergence, 56–61, 66, 81
 - convergence rate, 86
 - cost, 45–47, 87
 - generalised, 61
 - iteration matrix, 41
 - modified, 135
 - nonlinear, 322
 - numerical example, 14, 66, 67, 88
 - symmetric, *see* SSOR iteration
 - unsymmetric, 136
 - sparse matrix format, 16

- sparsity, 15, 465
uniform, 479
- spectral condition number, **421**
- spectral equivalence, 168, 321, 383
- spectral norm, 420, 422, **424**, 427, 434
- spectral radius, **24**, **406**, 408, 424, 425, 427
- spectrum, 403, 404, 408
- splitting
additive, 139, 145, 151, 202
multi-, 140
P-regular, 147
regular, 141–145, 156, 159, 204
weakly, 141
- SSOR iteration, 103, 132, 164, 296
2-cyclic, 134
convergence, 132, 133, 135, 146
cost, 133
iteration matrix, 132
numerical example, 136, 223
semi-iterative, 199
- stability, 157
- star notation, 9
- star pattern, 151
- starting value (of an iteration), 12, 17–19, 29, 33, 288
- steepest descent, 213
- stencil, *see* star
- step size, *see* grid size
- Stieltjes matrix, 451
- stochastic coefficients, 397
- stopping criterion, 34, 321
- storage cost, 10, 11, 133, 237, 469
- submultiplicativity, 420, 428
- subspace iteration, 325, **336**, 369
- substitution
backward, 372
forward, 372
- successive displacement, 39
- superconvergence, 240
- support, 463
- SVD, *see* singular value decomposition
- Sylvester equation, 208, 399
- SYMMQL, 257
- system of equations
linear, 17
nonlinear, 292, 319, 320
parametrised, 396
- tensor
core, 394
elementary, 388
- tensor product, 385
- tensor rank, 390
- tensor representation, 390
- tensor space
algebraic, 387
- topological, 389
- test of consistency, 32
- test of convergence, 33
- test vector, 154, 318, 319
- theorem
Cayley-Hamilton, 185, 413
comparison, 142
Ostrowski, 57
Perron–Frobenius, 440
Rouché, 194
Stein–Rosenberg, 144
Young, 81
- three-term recursion, 29, 183, 190, 195, 247
- time-stepping method, 171
quasi-, 172
- total reduction method, 12
- transformation
left, **112**, 114, 219
right, **115**, 117
similarity, 122, 404, 425
two-sided, **119**, 262
- triangle inequality, 417
inverse, 418
- tridiagonal part, 402
- truncated iteration, 398
convergence, 399
- truncation, 470
matrix, 416
- two-grid iteration, 274
contraction number, 304
convergence, 276, 280, 301, 306, 308, 309, 315, 360
iteration matrix, 274
nonlinear, 323
numerical example, 275
- two-step iteration, 195
- UL decomposition of the inverse matrix, 381
- underrelaxation method, 41
- V-cycle, 281, 362
convergence, 310, 364
numerical example, 282
- variational formulation, 474
- vector, 8
principal, 412
singular, 416
unit, 8
- vector block, 407
- W-cycle, 281, 285
convergence, 311
numerical example, 282
- weak formulation, 474
- Young, David M. Jr., 4