Iterative Methods for Linear Systems
Iterative Methods for Linear Systems
Theory and Applications

Maxim A. Olshanskii
University of Houston
Houston, Texas

Eugene E. Tyrtyshnikov
Russian Academy of Sciences
Moscow, Russia

Society for Industrial and Applied Mathematics
Philadelphia
Contents

List of Figures ix
List of Algorithms xi
Preface xiii

1 Krylov Subspace Methods 1
  1.1 Simple iterative methods ....................................... 1
     1.1.1 Preliminaries ........................................... 2
     1.1.2 Simple linear iterative methods ........................ 4
     1.1.3 A few convergence results ............................... 5
     1.1.4 Positive definiteness, A-norm, and A-orthogonality ..... 6
     1.1.5 A convergence result for Hermitian matrices .......... 7
     1.1.6 Relaxation methods ...................................... 8
     1.1.7 Relaxation parameters and Chebyshev polynomials ..... 9
     1.1.8 Simple bounds for non-Hermitian matrices ............ 12
     1.1.9 Quadratic functionals and linear systems ............ 13
     1.1.10 Steepest descent and residual minimization ......... 13
     1.1.11 Fixed point methods .................................... 15
     1.1.12 Convergence and divergence of fixed point methods . 16
     1.1.13 Newton's method ....................................... 16
     1.1.14 The Anderson acceleration .............................. 18
  1.2 Subspaces and iterative methods ................................ 20
     1.2.1 Minimization and projection ............................. 21
     1.2.2 Krylov subspaces ....................................... 21
     1.2.3 Optimal subspaces ....................................... 22
     1.2.4 The optimality of the Krylov subspaces ............... 23
     1.2.5 The minimal residual method, GMRES ................. 24
     1.2.6 The conjugate gradient method ........................ 27
     1.2.7 Krylov subspaces and matrix decompositions ......... 29
     1.2.8 Formal scalar product .................................. 30
     1.2.9 The biconjugate gradient method ....................... 30
     1.2.10 The quasi-minimal residual method ................... 32
     1.2.11 A minimal residual method for nonlinear problems .. 33
     1.2.12 Finite precision computations ......................... 34
  1.3 Analysis of the minimal residual method ...................... 36
     1.3.1 Convergence ............................................ 36
     1.3.2 Positive definite matrices .............................. 36
Contents

1.3.3 Convergence of the restarted method .............. 37
1.3.4 Polynomials and estimates ........................ 39
1.3.5 Polynomials and resolvent ........................ 39
1.3.6 Asymptotic convergence factor .................... 40
1.3.7 Numerical range .................................... 41
1.3.8 An estimate for resolvent .......................... 41
1.3.9 Pseudospectra ....................................... 42
1.3.10 An estimate for normal matrices .......................... 43
1.3.11 Minimal residuals and the Laplace equation ................. 44
1.3.12 The method of logarithmic potential .................... 44
1.3.13 Further analysis of the method of logarithmic potential .... 45

1.4 Analysis of the conjugate gradient method ................. 48
1.4.1 Convergence of the conjugate gradient method ............. 48
1.4.2 The classic and some sharper estimates ...................... 49
1.4.3 The Arnoldi and Lanczos methods .......................... 50
1.4.4 Ritz vectors and Ritz values ................................ 51
1.4.5 Recurrences and polynomials ................................ 52
1.4.6 “Superlinear convergence” and “disappearing” eigenvalues ... 53
1.4.7 Inexact iterations ....................................... 55
1.4.8 An inexact Richardson method .......................... 55
1.4.9 An inexact conjugate gradient method ..................... 57
1.4.10 Preconditioning ......................................... 58
1.4.11 Preconditioning, the Hermitian case ...................... 59
1.4.12 Number of iterations and alternative convergence estimates ... 60
1.4.13 A linear iteration as a preconditioner ................. 62

2 Toeplitz Matrices and Preconditioners 67
2.1 Introduction to Toeplitz Matrices .......................... 67
2.1.1 Toeplitz matrices and polynomials ........................... 67
2.1.2 Circulant matrices and Fourier matrices .................... 69
2.1.3 The fast Fourier transform .................................. 71
2.2 Preconditioners and applications ........................... 73
2.2.1 Circulant preconditioners ................................ 73
2.2.2 Analysis of optimal preconditioners ....................... 75
2.2.3 Analysis of simple preconditioners ....................... 79
2.2.4 Which circulant preconditioner is better? ............... 80
2.2.5 Proof of estimates for the outlier function ............. 84
2.2.6 Are there better circulants? ............................. 87
2.2.7 Multilevel Toeplitz matrices and multilevel circulants .... 92
2.2.8 Applications and remarks ................................ 94

3 Multigrid Preconditioners 99
3.1 The introductory section ................................. 99
3.1.1 Preliminaries ........................................ 100
3.1.2 A two-point boundary value problem ....................... 100
3.1.3 Finite element method: A two-point boundary value problem ... 101
3.1.4 Finite element method: Systems of algebraic equations ... 101
3.1.5 The Poisson equation ................................ 103
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1.6</td>
<td>Approximation properties of finite elements</td>
<td>105</td>
</tr>
<tr>
<td>3.1.7</td>
<td>Stiffness matrices</td>
<td>106</td>
</tr>
<tr>
<td>3.2</td>
<td>Two-grid iteration</td>
<td></td>
</tr>
<tr>
<td>3.2.1</td>
<td>The smoothing property of basic iterative methods</td>
<td>108</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Coarse grid correction</td>
<td>109</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Prolongation and restriction</td>
<td>110</td>
</tr>
<tr>
<td>3.2.4</td>
<td>Coarse grid operator</td>
<td>111</td>
</tr>
<tr>
<td>3.2.5</td>
<td>Iteration matrix</td>
<td>112</td>
</tr>
<tr>
<td>3.2.6</td>
<td>Convergence analysis for the model problem</td>
<td>112</td>
</tr>
<tr>
<td>3.2.7</td>
<td>Two-grid preconditioner</td>
<td>114</td>
</tr>
<tr>
<td>3.3</td>
<td>Multigrid iteration</td>
<td></td>
</tr>
<tr>
<td>3.3.1</td>
<td>V-, W- and F-cycles</td>
<td>117</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Complexity of a multigrid preconditioner</td>
<td>118</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Iteration matrix</td>
<td>119</td>
</tr>
<tr>
<td>3.3.4</td>
<td>Convergence of the multigrid iteration</td>
<td>121</td>
</tr>
<tr>
<td>3.3.5</td>
<td>Convergence of the W-cycle</td>
<td>122</td>
</tr>
<tr>
<td>3.3.6</td>
<td>Smoothing and approximation properties</td>
<td>123</td>
</tr>
<tr>
<td>3.3.7</td>
<td>The case of a self-adjoint problem</td>
<td>124</td>
</tr>
<tr>
<td>3.4</td>
<td>Convergence analysis</td>
<td></td>
</tr>
<tr>
<td>3.4.1</td>
<td>The Poisson problem: Approximation property</td>
<td>127</td>
</tr>
<tr>
<td>3.4.2</td>
<td>The Poisson problem: Smoothing property</td>
<td>129</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Less standard cases</td>
<td>131</td>
</tr>
<tr>
<td>3.4.4</td>
<td>The case of less than full regularity</td>
<td>134</td>
</tr>
<tr>
<td>3.4.5</td>
<td>Variational framework</td>
<td>135</td>
</tr>
<tr>
<td>3.4.6</td>
<td>Nonconforming finite elements and $U_{k-1} \not\subset U_k$</td>
<td>138</td>
</tr>
<tr>
<td>3.4.7</td>
<td>Multigrid for nonconforming elements</td>
<td>140</td>
</tr>
<tr>
<td>4</td>
<td>Preconditioners by Space Decomposition</td>
<td>145</td>
</tr>
<tr>
<td>4.1</td>
<td>Space decomposition framework</td>
<td></td>
</tr>
<tr>
<td>4.1.1</td>
<td>Space decomposition and preconditioning</td>
<td>145</td>
</tr>
<tr>
<td>4.1.2</td>
<td>Sequential version</td>
<td>147</td>
</tr>
<tr>
<td>4.1.3</td>
<td>The strengthened Cauchy–Schwarz inequality</td>
<td>148</td>
</tr>
<tr>
<td>4.1.4</td>
<td>Eigenvalue analysis of space decomposition preconditions</td>
<td>149</td>
</tr>
<tr>
<td>4.2</td>
<td>Grid decomposition methods</td>
<td></td>
</tr>
<tr>
<td>4.2.1</td>
<td>Additive multigrid method</td>
<td>152</td>
</tr>
<tr>
<td>4.2.2</td>
<td>BPX preconditioner</td>
<td>153</td>
</tr>
<tr>
<td>4.2.3</td>
<td>Local grid refinement</td>
<td>155</td>
</tr>
<tr>
<td>4.2.4</td>
<td>Hierarchical basis preconditioner</td>
<td>156</td>
</tr>
<tr>
<td>4.3</td>
<td>Domain decomposition methods</td>
<td></td>
</tr>
<tr>
<td>4.3.1</td>
<td>Alternating Schwarz method</td>
<td>158</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Fitting alternating Schwarz to space decomposition framework</td>
<td>159</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Multiplicative and additive Schwarz preconditioners</td>
<td>160</td>
</tr>
<tr>
<td>4.3.4</td>
<td>Nonoverlapping domain decomposition preconditioners</td>
<td>161</td>
</tr>
<tr>
<td>4.3.5</td>
<td>A matrix form of the Poincaré–Steklov operator</td>
<td>163</td>
</tr>
<tr>
<td>4.3.6</td>
<td>A Neumann–Neumann preconditioner</td>
<td>164</td>
</tr>
<tr>
<td>4.4</td>
<td>Convergence analysis for the Poisson problem</td>
<td></td>
</tr>
<tr>
<td>4.4.1</td>
<td>BPX preconditioner</td>
<td>165</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Hierarchical basis preconditioner</td>
<td>169</td>
</tr>
</tbody>
</table>
## Contents

4.4.3 Multiplicative and additive Schwarz preconditioners 171
4.4.4 Nonoverlapping domain preconditioners 172

5 Some Applications 175

5.1 Multigrid preconditioners for singular-perturbed problems 175
5.1.1 Singular-perturbed partial differential equations 175
5.1.2 What is a robust method? 177
5.1.3 Conditions on smoothing iterations 177
5.1.4 Bibliographical remarks 178
5.1.5 The linear reaction-diffusion problem 179
5.1.6 ILU decomposition, ILU smoother 182
5.1.7 Anisotropic diffusion equation 185
5.1.8 Upwind schemes for the convection-diffusion equations 190
5.1.9 A smoother for the convection-diffusion equations 192
5.1.10 The streamline-diffusion finite element method 193
5.1.11 Analysis of a multigrid preconditioner for convection-diffusion equations 198
5.1.12 Linear elasticity 201

5.2 Preconditioners for certain problems of fluid mechanics 205
5.2.1 The Stokes problem 206
5.2.2 The algebraic Stokes system 209
5.2.3 Some useful operator norms 211
5.2.4 The Vanka smoother for the Stokes system 212
5.2.5 Distributive iterations as smoothers 214
5.2.6 Smoother of Braess and Sarazin 215
5.2.7 Inexact Uzawa smoother 217
5.2.8 Approximation property 218
5.2.9 Inner-outer iterations 220
5.2.10 Block-diagonal preconditioner 220
5.2.11 Augmented-Lagrangian preconditioners I 222
5.2.12 Augmented-Lagrangian preconditioners II 224

Bibliography 231

Index 245
List of Figures

3.1 Basis functions for piecewise linear elements. .......................... 102
3.2 An example of the domain triangulation and enumeration of vertices for piecewise linear elements. .......................... 105
3.3 Prolongation operator for piecewise linear finite elements. 111
3.4 Restriction operator for piecewise linear finite elements. 111
3.5 V-cycle. .......................... 118
3.6 W-cycle. .......................... 118
3.7 F-cycle. .......................... 119
3.8 The stencil of the projection \( r_k \) from the fine to the coarse grids in the case of piecewise linear finite elements. 129
3.9 (a) L-shaped domain; (b) domain with a cut. 135
3.10 Nonconforming \( P_1 \) elements. (a) Coarse grid; (b) fine grid. 139
3.11 Nonconforming \( P_2 \) element. (a) Coarser grid; (b) finer grid. 140
4.1 One cycle of the additive method. .......................... 153
4.2 Hierarchical basis. .......................... 157
4.3 Nonhierarchical basis. .......................... 157
4.4 Decomposition of \( \Omega \) into two overlapping domains for the alternating Schwarz method. .......................... 159
5.1 (a) The solution to the convection-diffusion equation (5.48) on \([0,0.6]\) for \( \varepsilon = 10^{-2} \); (b) the central finite differences solution (5.50), \( h = \frac{1}{20} \); (c) the upwind finite differences solution (5.52). 191
5.2 Basis of weakly divergence-free functions for \( P_2 - P_0 \) elements. 203
5.3 An example of LBB stable pair \((isoP_2 - P_0)\). 209
5.4 Nonzero pattern of matrix from (5.119). 213
# List of Algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Method</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>GMRES method</td>
<td>26</td>
</tr>
<tr>
<td>1.2</td>
<td>CG method</td>
<td>28</td>
</tr>
<tr>
<td>1.3</td>
<td>Bi-CG method</td>
<td>31</td>
</tr>
<tr>
<td>1.4</td>
<td>PCG method</td>
<td>59</td>
</tr>
<tr>
<td>3.1</td>
<td>Two-grid preconditioner</td>
<td>115</td>
</tr>
<tr>
<td>3.2</td>
<td>Multigrid preconditioner</td>
<td>117</td>
</tr>
</tbody>
</table>
Preface

Systems of linear algebraic equations are ubiquitous in mathematical modeling and scientific computing, and iterative methods are indispensable for the numerical treatment of such systems. What makes iterative solvers so efficient, and what has maintained their state-of-the-art status for almost half of a century? Foremost, these are the optimality property of the Krylov subspaces, which define iterations, and the concept of preconditioning, which gives the approach flexibility and imbues it with mathematical ideas from various fields of applications.

Iterative methods for linear algebraic systems comprise a rapidly developing field of research. The growing variety of applications makes even a nearly complete review of existing preconditioning techniques an overwhelming task. This book gives a mathematically rigorous, consistent introduction to the fundamental and most popular Krylov subspace iterative methods. We review the theory of the Krylov subspaces, including the optimality property, and introduce iterative algorithms following the classic minimization approach. Our analysis of the Krylov subspace methods is concise and rigorous, covering convergence bounds in terms of optimal polynomials, matrix resolvents, numerical range, logarithmic potentials, and alternative estimates. The preconditioning is naturally tied to specific classes of matrices and particular applications which lead to linear systems of equations. We restrict ourselves to the structured (Toeplitz) matrices and to large sparse matrices arising from discretizations of partial differential equations.

The book consists of five chapters. Chapter 1 starts with necessary preliminaries and reviews basic iterative methods. Further, we introduce the Krylov subspaces and prove an important optimality result. Fundamental Krylov subspace methods are defined and considered from algebraic, geometrical, and nonlinear optimization viewpoints. This is followed by the comprehensive convergence analysis of the Krylov subspace methods in a general case and the case of Hermitian matrices. Besides classical estimates, several special topics important in practice are considered, including superlinear convergence and inexact iterations. When we apprehend what properties of the coefficient matrix account for the rate of convergence, we may multiply the original system by some nonsingular matrix, called a preconditioner, so that the new coefficient matrix possesses better properties. Chapter 1 closes with a simple observation, which motivates much of the subsequent discussion: any linear consistent iteration defines a preconditioner.

In Chapter 2, we introduce a class of Toeplitz matrices, review their basic properties, and study preconditioners for this class of structured matrices. Chapter 3 deals with multigrid (MG) preconditioners for systems of linear algebraic equations arising from discretizations of partial differential equations. For reference purposes, we review basic principles of a finite element method, taking a two-point boundary value problem and the Poisson equations as model examples. Classic geometric MG preconditioners, in particular, V- and W-cycles are introduced, and a complete analysis is given in the framework of smoothing and approximation properties. Two-dimensional Poisson equations
Chapter 4 describes preconditioners based on space decompositions. These are multi-level, hierarchical basis, and domain decomposition (including alternating Schwarz, additive Schwarz, and nonoverlapping domains) preconditioners. These preconditioners are analyzed in a uniform framework of space decomposition methods. As in Chapter 3, we use the example of the discrete Poisson equations to apply a general analysis. More examples are collected in Chapter 5, where we show how the ideas and techniques from the previous chapters apply to the numerical solution of many practical problems. Most of these problems primarily arise in computational physics. Such selection exclusively reflects the authors’ own experience and does not indicate any limitation on applicability or efficiency of preconditioned Krylov subspace solvers in other areas of scientific computing.

We believe this book is a useful addition to many excellent texts on iterative methods of linear algebra and their applications, including those by Axelsson [9], Bramble [35], Greenbaum [89], Hackbusch [96, 97], Liesen and Strakoš [131], Marchuk and Kuznetsov [141], Saad [173], Shaidurov [181], Trottenberg, Oosterlee, and Schuller [196], and Varga [210]. It complements other textbooks with its comprehensive and consistent analysis, yet it is still accessible to students with only basic knowledge of linear algebra and calculus. The book includes several topics on Krylov subspace solvers rarely available in texts on numerical linear algebra. Besides this, it also contains all essentials of the multigrid and space decomposition preconditioners via elementary mathematical tools. The final chapter collects many examples of important applications, where the techniques and analysis of previous chapters result in remarkably efficient practical solvers. Most of the material from this chapter had been previously found only in the form of journal publications.

Although the text includes several advanced topics, we strived to achieve a textbook-level of exposition, rather than to write a research monograph. The material included in the book was used by the authors to teach courses on numerical linear algebra and special topics at Moscow State University, the Moscow Institute of Physics and Technology, and the University of Houston, and to give shorter series of lectures at the graduate level. Our teaching experiences and personal tastes largely determined the style of presentation and the topics selection. We complement each section of the book with a list of exercises. Solving these problems is recommended, though not mandatory, for a better understanding of the text.

Tribute to Chebyshev and Krylov

Numerical linear algebra owes a lot to the great Russian mathematicians Pafnuty Lvovich Chebyshev (1821–1894) and Alexei Nikolaevich Krylov (1863–1945). In this text on iterative solvers, these two important names can be found in many instances, and we take this opportunity to remember a part of their scientific legacy.

Pafnuty Chebyshev is often considered the founding father of the Russian mathematical school. The marvelous Chebyshev polynomials are defined as monic polynomials on a closed interval with minimal deviation from zero. They appeared in his original memoir of 1857 in French under the title “Sur les questions de minima qui se rattachent à la représentation approximative des fonctions” and were published in the journal Memoirs of the Imperial Academy of Sciences, St. Petersburg. Chebyshev’s polynomials are recognized as a major milestone that ushered in the beginning of approximation theory. They are essential for the convergence analysis of iterative methods for systems of algebraic equations.
Alexei Krylov was a naval engineer and applied mathematician. He claimed that shipbuilding was his primary profession, or, better to say, the application of mathematics to various questions of maritime science. In linear algebra, the name of Alexei Krylov is given to a special sequence of subspaces in $\mathbb{C}^n$, which is fundamental in most practical numerical methods pervasive in applications, such as conjugate gradient (CG), generalized minimal residual (GMRES), Lanczos bidiagonalization algorithm, etc.

The Krylov subspaces were first introduced in 1931 in the paper “On numerical solution of the equation of technical sciences for frequencies of small oscillations in material systems.” In this paper, Alexei Krylov introduced a method for computing the coefficients of characteristic polynomials of matrices. Krylov’s idea was to reduce the equation

$$\det(A^T - \lambda I) = \det \begin{bmatrix} a_{11} - \lambda & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} - \lambda & \cdots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{nn} - \lambda \end{bmatrix} = 0 \quad (1)$$

to a more computationally affordable form

$$\det \begin{bmatrix} b_{11} - \lambda & b_{21} & \cdots & b_{n1} \\ b_{12} - \lambda^2 & b_{22} & \cdots & b_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ b_{1n} - \lambda^n & b_{2n} & \cdots & b_{nn} \end{bmatrix} = 0. \quad (2)$$

He obtained (2) from (1) by a few simple transformations of the equations

$$\lambda x_1 = a_{11} x_1 + a_{21} x_2 + \cdots + a_{n1} x_n,$$
$$\lambda x_2 = a_{12} x_1 + a_{22} x_2 + \cdots + a_{n2} x_n,$$
$$\vdots$$
$$\lambda x_n = a_{1n} x_1 + a_{2n} x_2 + \cdots + a_{nn} x_n.$$

The first step is the multiplication of the first equation by $\lambda$ and replacing $\lambda x_1, \ldots, \lambda x_n$ with the right-hand sides of the original equations. The second step is the multiplication of the new equation by $\lambda$ and the substitution as previously, and so on. In this way, one obviously finishes with the equation of the required form (2). Denote by $b_1, \ldots, b_n$ the column vectors built of the coefficients of the rows of the final equations

$$\lambda x_1 = b_{11} x_1 + b_{21} x_2 + \cdots + b_{n1} x_n,$$
$$\lambda^2 x_1 = b_{12} x_1 + b_{22} x_2 + \cdots + b_{n2} x_n,$$
$$\vdots$$
$$\lambda^n x_1 = b_{1n} x_1 + b_{2n} x_2 + \cdots + b_{nn} x_n.$$

Let $A$ be a matrix with the entries $a_{ij}$. Then it is not difficult to see that

$$b_1 = Ar_0, \quad b_2 = A^2 r_0, \quad \ldots, \quad b_n = A^n r_0, \quad \text{where} \quad r_0 = [1, 0, \ldots, 0]^T.$$
and suppose that
\[ \dim L_k = \dim L_{k+1} = k. \]

Then \( L_k \) is an invariant subspace with respect to \( A \), and the method allows one to obtain the coefficients of a polynomial which is a divisor of the characteristic polynomial. More precisely, one can find the minimal polynomial for a restriction of the matrix \( A \) on \( L_k \), and with an appropriate choice of \( r_0 \), it is always possible to get the minimal polynomial for \( A \) in the whole space; see [74].

In the modern analysis of iterative methods, the names Chebyshev and Krylov both are important and naturally stand close, because the Krylov vectors are intrinsically linked with polynomials. To realize this, consider a linear algebraic system \( Ax = b \) with a symmetric positive definite matrix \( A \). Then the CG method, with an initial guess \( x_0 \), constructs the Krylov subspaces \( L_k \) for \( r_0 := b - Ax_0 \) and defines the \( k \)th iterate \( x_k \) such that it solves the minimization problem
\[
\| b - Ax_k \|_{A^{-1}} = \min_{x \in x_0 + L_k} \| b - Ax \|_{A^{-1}},
\]

where
\[
\| x \|_{A^{-1}} := \sqrt{(A^{-1}x, x)}.
\]

Obviously, for any \( x_k \in x_0 + L_k \) we have \( r_k = f_k(A)r_0 \) with \( f_k \in F_k \), where \( F_k \) is the set of all polynomials of degree bounded by \( k \) and with the zero order term equal to 1, i.e., \( f_k(0) = 1 \). From the above minimization property,
\[
\| b - Ax_k \|_{A^{-1}} \leq \min_{f_k \in F_k} \| f_k(A) \|_{A^{-1}} \leq \min_{f_k \in F_k} \max_{m \leq \lambda \leq M} |f_k(\lambda)| \| r_0 \|_{A^{-1}},
\]

where \( m \) and \( M \) are the minimal and maximal eigenvalues of \( A \). Looking for the minimum of the right-hand side, one inevitably resorts to the Chebyshev polynomials. This is the standard way to obtain the classical convergence estimates for the CG and other Krylov subspace methods.

**Acknowledgments**

We are grateful to many colleagues for valuable discussions and comments on the manuscript. We would like to thank Nikolai Zamarashkin and Sergei Goreinov, who read the manuscript and suggested many improvements. We give our special thanks to Yuri Vassilevski for informal and insightful discussions of the domain decomposition methods. Valeria Simoncini read a draft of this book and initiated our cooperation with SIAM's book program. We thank Valeria for her encouragement. The research related to the book was supported in part by RFFI and DFG funding agencies. We also acknowledge the support of the University of Houston through the Small Grants Program, which helped us to finalize the manuscript.
Chapter 1

Krylov Subspace Methods

1.1 Simple iterative methods

This book focuses on methods for numerical and approximate solution of systems of linear algebraic equations. Many numerical algorithms and methods benefit from specific properties of algebraic systems, such as matrix sparsity, symmetry, or sign-definiteness, which can be shared by systems arising from particular applications. In this chapter, however, we strive to consider the most generic situation of a consistent linear algebraic system with a nonsingular, square, complex-valued matrix $A$.

In the matrix-vector notation, the problem of interest is

$$Ax = b,$$

where $A \in \mathbb{C}^{n \times n}$ is a given matrix, $b \in \mathbb{C}^n$ is a given vector, and $x \in \mathbb{C}^n$ is an unknown vector. We assume that the matrix $A$ is invertible, $\det A \neq 0$, and thus that the system is consistent.

For a general matrix $A$, direct methods for solving (1.1), such as Gauss elimination, require $O(n^3)$ arithmetic operations. This complexity can be often reduced due to a specific structure and properties of matrix $A$, depending on a concrete problem at hand. However, for very large systems arising in real-life applications, direct methods are often computationally nonfeasible. Although direct methods with lower complexity are known, they are more sensitive to roundoff errors and are more involved from the algorithmic viewpoint. In many cases, iterative methods for finding an approximate (up to a prescribed accuracy) solution to (1.1) are more attractive for the following reasons:

(i) A solution, vector $x \in \mathbb{C}^n$, is commonly used to approximate an unknown physical variable or statistic. Therefore, there are no definite reasons for finding the solution $x$ exactly; usually it would suffice to find an approximation to $x$ with some relevant accuracy.

(ii) In many applications, matrices are endowed with some additional structure, e.g., they are sparse (most of the entries are zero) or Toeplitz. Both direct and iterative methods may benefit from such an additional structure. In particular, iterative methods exploit the fact that matrix-vector multiplication is relatively cheap: for sparse matrices its computational complexity is determined by the number of nonzero entries. Moreover, the implementation of an iterative solver requires the storage only of these nonzero entries by their values and position in $A$. Several data formats are currently available for organizing efficient storage and fast access to the entries of a sparse matrix.
We open this chapter with necessary preliminaries and proceed with very basic facts about iterative methods. Further in the chapter, we advance to more sophisticated and efficient approaches and study their properties.

1.1.1 Preliminaries

We need some definitions and elementary facts from linear algebra.

For any elements \( a, b \) of a Hilbert or Euclidean vector space with scalar product \((\cdot, \cdot)\), we shall make frequent use of the Cauchy inequality

\[
|\langle a, b \rangle| \leq (a, a)^\frac{1}{2} (b, b)^\frac{1}{2} = \|a\| \|b\|,
\]

(1.2)

and for complex scalars \( a \) and \( b \), of the Young inequality

\[
|a b| \leq \frac{\varepsilon}{2}|a|^2 + \frac{1}{2\varepsilon} |b|^2, \quad \varepsilon \in (0, \infty).
\]

(1.3)

We say that a matrix norm \( \|\cdot\| \) is consistent with a vector norm \( \|\cdot\|_v \) if for any matrix \( A \in \mathbb{C}^{n \times n} \)

\[
\|A\| = \sup_{\|x\|_v \neq 0} \frac{\|Ax\|_v}{\|x\|_v}.
\]

(1.4)

A norm defined by (1.4) is also called an operator norm of \( A \). If it is not stated explicitly, we shall use throughout the book matrix norms consistent with vector norms. For consistent norms, the equality (1.4) immediately implies (we shall drop the subscript “\( v \)" since it is clear from the context that either matrix or vector norm is used)

\[
\|Ax\| \leq \|A\| \|x\| \quad \forall \ A \in \mathbb{C}^{n \times n}, \ x \in \mathbb{C}^n, \ (1.5)
\]

\[
\|AB\| \leq \|A\| \|B\| \quad \forall \ A, B \in \mathbb{C}^{n \times n}. \quad (1.6)
\]

A complex value \( \lambda = \lambda(A) \in \mathbb{C} \) is an eigenvalue and \( \psi \in \mathbb{C}^n \) is an eigenvector of \( A \) if

\[
A \psi = \lambda \psi, \quad \psi \neq 0.
\]

Any \( A \in \mathbb{C}^{n \times n} \) has at most \( n \) different complex eigenvalues, which are the roots of the characteristic polynomial \( \det(A - \lambda I) = 0 \). The degree of a root is the eigenvalue algebraic multiplicity. All eigenvalues of \( A \) form the spectrum \( \text{sp}(A) \). The value \( \rho(A) := \max_{\lambda \in \text{sp}(A)} |\lambda| \) is called the spectral radius of \( A \). Since eigenvalues are the roots of the characteristic polynomial, one gets

\[
\sum_{i=1}^{n} \lambda_i = \text{tr}(A) \quad \text{and} \quad \prod_{i=1}^{n} \lambda_i = \det(A). \quad (1.7)
\]

We reserve the angle brackets to denote the Hermitian inner product on \( \mathbb{C}^n \): \( \langle x, y \rangle = x^T \overline{y} \) for \( x = (x_1, \ldots, x_n)^T \in \mathbb{C}^n, \ y = (y_1, \ldots, y_n)^T \in \mathbb{C}^n; \) for \( x, y \in \mathbb{R}^n \) this is the Euclidean inner product. By \( A^T \) we denote the transpose matrix to \( A \): \( A^T = \{a_{ji}\} \) for \( A = \{a_{ij}\} \), and by \( A^* \) we denote the conjugate matrix for \( A, A^* = \{\overline{a_{ij}}\} \). For the Hermitian inner product it clearly holds that

\[
\langle Ax, y \rangle = \langle x, A^* y \rangle = \langle A^* y, x \rangle \quad \forall \ x, y \in \mathbb{C}^n, \ \forall \ A \in \mathbb{C}^{n \times n}. \quad (1.8)
\]

A matrix \( A \) is called symmetric if \( A = A^T \), and \( A \) is Hermitian if \( A = A^* \). For real-valued matrices these definitions are equivalent.
1.1. Simple iterative methods

All eigenvalues of a Hermitian matrix \( A \) are real, and we agree to enumerate them in decreasing order:

\[
\lambda_{\text{max}}(A) = \lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_{n-1}(A) \geq \lambda_n(A) = \lambda_{\text{min}}(A).
\]

An eigenvalue with algebraic multiplicity \( m \) occurs \( m \) times in this sequence. For a Hermitian matrix \( A \) there exists the decomposition

\[
A = QDQ^*,
\]

where \( D \) is a diagonal matrix, with eigenvalues of \( A \) on the main diagonal, and \( Q \) is a nonsingular matrix such that \( Q^* = Q^{-1} \) (such matrices are called unitary). The columns of \( Q \) are the eigenvectors of \( A \). A generalization of (1.9) for an \( n \times m \) matrix \( A \) is the singular value decomposition (SVD)

\[
A = QDV,
\]

where \( Q \) is an \( n \times n \) unitary matrix, the matrix \( D \) is an \( n \times m \) diagonal matrix with nonnegative real numbers on the diagonal, and \( V \) is an \( m \times m \) unitary matrix. The diagonal entries of \( D \) are the singular values of \( A \). Decompositions (1.10) and (1.9) yield that the squares of nonzero singular values are the eigenvalues of \( AA^* \) and \( A^*A \).

The norm

\[
||A||_2 = \sup_{\|x\|_2 \neq 0} \frac{\|Ax\|_2}{\|x\|_2}
\]

induced by the Euclidean vector norm \( \|x\|_2 = \langle x, x \rangle^{\frac{1}{2}} \) is called the spectral norm of \( A \). Using (1.9), one finds the equality

\[
||A||_2 = \sqrt{\lambda_{\text{max}}(AA^*)}.
\]

For a Hermitian matrix \( A \), it therefore holds that \( ||A||_2 = \max_{\lambda \in \text{sp}(A)} |\lambda(A)| \).

We shall also need the Frobenius matrix norm

\[
||A||_F = \left( \sum_{i,j=1}^{n} |a_{ij}|^2 \right)^{\frac{1}{2}}.
\]

The Frobenius norm satisfies (1.6). From (1.7) and the definitions of the norms, we have

\[
||A||_2 \leq \left( \sum_{i=1}^{n} \lambda_i(AA^*) \right)^{\frac{1}{2}} = (\text{tr}(AA^*))^{\frac{1}{2}} = ||A||_F.
\]

We shall make frequent use of the QR-decomposition of a square matrix \( A \): \( A = QR \), where \( Q \) is a unitary matrix and \( R \) is an upper triangular matrix (\( R_{ij} = 0 \) for \( j < i \)). If \( A \) is nonsingular and the diagonal elements of the factor \( R \) are normalized to be real and positive, then the QR-decomposition is unique. The decomposition of a square matrix \( A \) as the product of a lower triangular matrix \( L \) and an upper triangular matrix \( U \) is called the \( LU \)-decomposition, \( A = LU \). The factor \( R \) (or \( L \)) is the particular case of an upper (or lower) Hessenberg matrix: A square matrix \( H \) is called an upper (resp., a lower) Hessenberg matrix if \( H_{ij} = 0 \) for \( j < i - 1 \) (\( H_{ij} = 0 \) for \( i < j - 1 \)).

\(^1\)We call a matrix \( D = \{d_{ij}\} \) diagonal if \( d_{ij} = 0 \) for \( i \neq j \).
Another useful factorization of a matrix $A \in \mathbb{C}^{n \times n}$ follows from the existence of its Jordan normal form, a block-diagonal matrix $J$ such that

$$A = CJC^{-1}, \quad C \in \mathbb{C}^{n \times n}, \quad J = \begin{pmatrix} J_1 & & \\ & J_2 & \\ & & \ddots \\ & & & J_m \end{pmatrix},$$

$$J_k \in \mathbb{C}^{p_k \times p_k}, \quad p_1 + \cdots + p_m = n.$$

The elementary Jordan blocks $J_k$ forming $J$ are two-diagonal matrices of the form

$$J_k = \lambda I + \tilde{I}, \quad \text{with} \quad \lambda \in \text{sp}(A), \quad \tilde{I}_{i,j} = \delta_{i,j}^{-1},$$

where $I$ is the identity matrix in $\mathbb{C}^{p_k \times p_k}$, $\delta$ is the Kronecker symbol, and thus the $p_k \times p_k$ matrix $\tilde{I}$ has 1 on the diagonal above the main diagonal, and all other entries equal zero.

### 1.1.2 Simple linear iterative methods

First, we consider a basic iterative method in the following form.

Assume we are given an approximation $x^i \in \mathbb{C}^n$ to the unknown solution $x$ of (1.1); integer $i \geq 0$ is an index. We find a new (hopefully, better) approximation $x^{i+1}$ from

$$W(x^{i+1} - x^i) = b - Ax^i, \quad i = 0, 1, 2, \ldots$$

To use (1.11) in practice, one has to define the following:

(a) An initial guess $x^0$ (for example, the zero vector is a good choice in many instances).

(b) An auxiliary invertible matrix $W$ (the simplest, although not always the best, choice is to set $W$ equal the identity matrix, $W = I$).

(c) A stopping criterion (one may assume that $x^i$ is a good approximation to $x$ and terminate iterations (1.11) if $\|b - Ax^i\| \leq \varepsilon$ for some $\varepsilon > 0$ and a norm $\|\cdot\|$).

The role of auxiliary matrix $W$ will become clearer later. The matrix $W$ (or $W^{-1}$) is sometimes called a preconditioner. The idea of preconditioners and preconditioning is central to this book and will be further discussed in Section 1.4. One step of (1.11), $i \to i + 1$, is called an iteration.

The error vector is the difference $e^i := x - x^i$ between the unknown solution $x$ to (1.1) and given iterate $x^i$. Clearly, one is interested in getting the error close to the zero vector. The method (1.11) is said to converge if the error $e^i := x - x^i$ tends to zero in some norm:

$$\|e^i\| \to 0, \quad i \to \infty.$$  

The definition is independent of the choice of any particular vector norm. We say that the method has a geometric convergence with the convergence factor $q \in [0, 1)$ if the error is reduced on each iteration at least by a factor of $q$ in some norm:

$$\|e^{i+1}\| \leq q\|e^i\|, \quad i = 0, 1, 2, \ldots$$

From (1.1) and (1.11) we conclude that the error vectors satisfy the recursion

$$e^{i+1} = (I - W^{-1}A)e^i, \quad i = 0, 1, 2, \ldots$$
1.1. Simple iterative methods

Here and further \( I \) denotes the identity matrix. The matrix \( S := I - W^{-1}A \) from (1.12) is called the iteration matrix of the method (1.11). The iteration matrix is an important notion, since this matrix is responsible for the error evolution. In linear iterative methods the error \( e^i \) depends linearly on \( e^0 \). From (1.12) we get \( e^i = S^i e^0 \), and hence (1.11) is the linear method.

One easily checks that \( x^{i+1} \) and \( x^i \) satisfy the identity

\[
x^{i+1} = S x^i + g, \quad i = 0, 1, 2, \ldots,
\]

where \( g \) is the vector dependent on the right-hand side \( b \) but independent of \( i \). The reverse implication is also true: If (1.13) holds for a consistent linear iterative method (an iterative method is called consistent if \( x^i = x \) implies \( x^{i+1} = x \)), then the matrix \( S \) from (1.13) is the iteration matrix of the method.

1.1.3 A few convergence results

Now we recall a few results about the convergence of iteration (1.11). These results are the direct consequence of the error relation (1.12).

**Theorem 1.1.** Assume that for an iteration matrix the estimate \( \|S\| \leq q < 1 \) holds in some operator norm; then the method (1.11) geometrically converges for any initial guess \( x^0 \) and

\[
\|e^i\| \leq q \|e^0\|, \quad i = 0, 1, 2, \ldots
\]

in the corresponding vector norm.

**Proof.** It directly follows from (1.12) that

\[
\|e^i\| = \|S e^{i-1}\| \leq \|S\| \|e^{i-1}\| \leq q \|e^{i-1}\|.
\]

Applying the estimate recursively leads to (1.14). \( \square \)

**Theorem 1.2.** The method (1.11) converges for any \( x^0 \) iff \( \rho(S) < 1 \). Moreover, for large enough \( i \) the spectral norm of the iteration matrix is bounded as

\[
\|S^i\|_2 \leq c_p i^{p-1} \rho(S)^{i-p+1},
\]

where \( p \) is the largest size of all elementary Jordan blocks of the normal Jordan form of \( S \). The constant \( c_p \) is independent of \( i \) but may depend on \( p \).

**Proof.** In this proof, \( \| \cdot \| \) denotes the spectral matrix norm. From the definition of the Jordan normal form we have \( S^i = C^{-1} J^i C \), with a matrix \( C \) independent of \( i \). Thus (1.6) yields \( \|S^i\| \leq c \|J^i\| \), with a constant \( c \) independent of \( i \). Since the Jordan form \( J \) has the block-diagonal structure, its norm is maximum over corresponding norms of elementary Jordan blocks forming \( J \). The same is true for \( J^i \). Thus, we consider an arbitrary Jordan block \( \tilde{J} \in \mathbb{C}^{p \times p} \) of size \( p \) and split

\[
\tilde{J} = \lambda I + \tilde{A}, \quad \text{with} \quad \lambda \in \text{sp}(A), \quad \tilde{A}_{ij} = \delta_{j-i}^{p-1}.
\]

Note that \( \tilde{A} \) is a \( p \times p \) matrix, and for any integer \( j \geq 0 \) we have \( \|\tilde{A}^j\| = 1 \) for \( j < p \) and \( \|\tilde{A}^j\| = 0 \) for \( j \geq p \). Denoting by \( C^j_i \) the binomial coefficient \( \binom{n}{j-i-j} \), we compute

\[
\|\tilde{J}^i\| = \|(\lambda I + \tilde{A}^i)\| = \left\| \sum_{j=0}^i C^j_i \lambda^{i-j} \tilde{A}^j \right\| = \left\| \sum_{j=0}^{\min\{i, p-1\}} C^j_i \lambda^{i-j} \tilde{A}^j \right\|.
\]
For $p = 1$, we have $|\vec{f}^i| = \lambda|^i|$; for $p > 1$ we may assume $i$ is sufficiently large, i.e., $i \geq p$, and use the triangle inequality to estimate

$$
||\vec{f}^i|| \leq |\lambda|^{i-p+1} \sum_{j=0}^{p-1} C_j |\lambda|^{p-1-j} \leq |\lambda|^{i-p+1} p \max \{1, |\lambda|^{p-1}\}
$$

$$
\leq c_p \rho(S)^{i-p+1} i^{p-1}.
$$

Here we used the rough bound $\frac{1}{\rho(S)^{i-j}} \leq i^{p-1}$ for $i \geq p > 2, j < p$. This proves the desired estimate (1.15).

If all eigenvalues of the iteration matrix $S$ lie inside the unit circle, $\rho(S) < 1$, then (1.15) yields $||S^i|| \rightarrow 0$ for $i \rightarrow \infty$. Hence the identity (1.12) implies that the error of the method tends to zero for big enough $i$ (the method converges). Otherwise, if $\rho(S) \geq 1$, there exists an eigenvalue $|\lambda(S)| \geq 1$. Assume $x^0$ is such that the error vector $e^0$ is the corresponding eigenvector of $S$. For such choice of $x^0$, (1.12) yields $||e^i|| = |\lambda(S)|^i ||e^0|| \rightarrow 0$.

The proof of (1.15) remains valid for matrix norms different from the spectral norm, as long as $||\vec{f}^i|| \leq 1$. In particular, this is true for matrix norms introduced in Exercise 1.1.1. For an arbitrary matrix norm, estimate (1.15) holds with the constant $c_p$ replaced by some general constant $C$ dependent on matrix size.

Since the solution $x$ is usually not known a priori, one has no access to the error $e^i$. At the same time, if there exists a good estimate for $||S||$, then, thanks to the inequality $||e^i|| \leq ||S|| ||e^0||$, the value of $||e^0||$ can be estimated up to the value of the initial error $||e^0||$. Using exclusively information from the iteration process (1.11), one may judge the quality of an iterate $x^i$ by looking at the residual vector $r^i = b - Ax^i$. If $x^i$ is the exact solution, then $r^i = 0$. More generally, the equality

$$
A e^i = r^i
$$

implies $||x - x^i|| \leq ||A^{-1}|| ||b - Ax^i||$.

### 1.1.4 Positive definiteness, $A$-norm, and $A$-orthogonality

We shall often consider the important class of positive definite matrices. We write $A > 0$ (resp., $A \geq 0$), if $\text{Re} \langle Ax, x \rangle > 0$ (resp., $\text{Re} \langle Ax, x \rangle \geq 0$) for any $0 \neq x \in \mathbb{C}^n$. Also, $A > B$ if $A - B > 0$. A matrix $A > 0$ (resp., $A \geq 0$) will be called positive (nonnegative) definite. The definition yields $\text{Re} \lambda(A) > 0$ for eigenvalues of a positive definite matrix, $\lambda(A) \in \text{sp}(A)$. All eigenvalues of a positive (nonnegative) Hermitian matrix are real and positive (nonnegative). The positive definiteness of $A$ implies that the real parts of the main diagonal elements of $A$ are all positive (check this!). Off-diagonal elements of a positive definite matrix, however, can have positive, or negative, or zero real parts. The reverse implication is also not true; in general, the sign-definiteness of matrix elements does not necessarily result in $A > 0$ or $A \geq 0$.

A positive definite Hermitian matrix $A$ admits the unique Cholesky decomposition $A = LL^*$ with a lower triangle matrix $L$. One reason why positive definite Hermitian matrices form an important class of matrices is because, for a positive definite Hermitian matrix $A$, the following inner product and the norm are well defined:

$$
\langle x, y \rangle_A := \langle Ax, y \rangle \quad \forall x, y \in \mathbb{C}^n,
$$

$$
||x||_A := (\langle x, x \rangle_A)^{1/2} \quad \forall x \in \mathbb{C}^n.
$$

We leave it to the reader to check that $\langle Ax, y \rangle$ indeed defines an inner product in $\mathbb{C}^n$. 
1.1.5 A convergence result for Hermitian matrices

The norm $\|x\|_A$ is called the $A$-norm. The vectors $x$ and $y$ are called $A$-orthogonal if $\langle x, y \rangle_A = 0$. In this case, the Pythagorean theorem reads

$$\|x + y\|_A^2 = \|x\|_A^2 + \|y\|_A^2 \quad \text{if} \quad \langle x, y \rangle_A = 0,$$

and the best approximation property of the $A$-orthogonal projection can be formulated as follows: Let $V$ be a subspace of $\mathbb{C}^n$ and take $y \in \mathbb{C}^n$; then

$$\hat{y} = \arg\min\{\|x - y\|_A : x \in V\} \iff \langle \hat{y} - y, x \rangle_A = 0 \quad \forall x \in V. \quad (1.16)$$

In the equivalence (1.16), the implication ($\Rightarrow$) readily follows from the Pythagorean theorem above. To verify that the best approximation implies that $\hat{y} - y$ is $A$-orthogonal to $V$, it is sufficient to consider arbitrary $x \in V$ and the real-valued function $f(t) = \frac{1}{2}\|x - y + tx\|_A^2$, $t \in \mathbb{R}$. The best approximation property of $\hat{y}$ yields

$$f'(0) = 0 \quad \Rightarrow \quad f'(0) = \langle \hat{y} - y, x \rangle_A = 0.$$

### Theorem 1.3

Let $A = A^* > 0$ and $W > \frac{1}{2}A$, then the iterative method (1.11) geometrically converges, and

$$\|e^{i+1}\|_A \leq q \|e^i\|_A \quad \text{with} \quad q = (1 - 2\delta_1 \delta \|W\|^{2-2})^\frac{1}{2},$$

where $\delta = \lambda_{\min}(A)$, $\delta_1 = \lambda_{\min}(W_0 - \frac{1}{2}A)$, $W_0 = \frac{1}{2}(W + W^*)$.

Note that $\text{sp}(W_0 - \frac{1}{2}A) \subset \mathbb{R}$ and $\delta_1 > 0$ (see Exercise 1.1.4).

**Proof.** From the relation (1.12) and the definition of the $\| \cdot \|_A$ norm, one gets

$$\|e^{i+1}\|_A^2 = \langle ASe^i, Se^i \rangle = \langle A(I - W^{-1}A)e^i, (I - W^{-1}A)e^i \rangle.$$

Denote $v = W^{-1}Ae^i$. Using the identity (1.8) for $A = A^*$, we rewrite the latter relation as

$$\|e^{i+1}\|_A^2 = \|e^i\|_A^2 - 2\left(\langle W_0 - \frac{1}{2}A \rangle v, v \right). \quad (1.17)$$

From the Rayleigh equalities (1.50) it follows that

$$2\left(\langle W_0 - \frac{1}{2}A \rangle v, v \right) \geq 2\delta_1 \|v\|_2^2. \quad (1.18)$$

At the same time, it holds that

$$\|e^i\|_A^2 = \langle Ae^i, e^i \rangle = \langle Wv, A^{-1}Wv \rangle \leq \|A^{-1}\|_2 \|Wv\|_2^2$$

$$\leq \|A^{-1}\|_2 \|W\|_2^2 \|v\|_2^2 = \lambda_{\max}(A^{-1}) \|W\|_2^2 \|v\|_2^2$$

$$= \lambda_{\min}^{-1}(A) \|W\|_2^2 \|v\|_2^2 = \delta^{-1} \|W\|_2^2 \|v\|_2^2,$$

or simply that

$$\|v\|_2^2 \geq \delta \|W\|_2^{-2} \|e^i\|_A^2. \quad (1.19)$$
Estimates (1.18)–(1.19), together with (1.17), lead to
\[
\|e^{i+1}\|^2_A \leq (1 - 2\delta_1 \delta_2 \|W\|^2_2)\|e^i\|^2_A.
\]

1.1.6 Relaxation methods

Below we give several examples of basic iterative methods. These methods are often called relaxation methods.

**Example 1.4.** The simple iteration, or Richardson method, is defined by (1.11) with \(W = \omega^{-1}I\). A scalar parameter \(\omega \neq 0\) is often called the relaxation parameter.

In the next two examples, we assume that all diagonal elements of a matrix \(A\) are non-zero.

**Example 1.5.** The Jacobi method is the iteration (1.11) with \(W = \text{diag}(A)\). Here and further, \(\text{diag}(A)\) is the diagonal matrix obtained from \(A\) by setting all off-diagonal elements equal to zero.

**Example 1.6.** Consider the decomposition of \(A\) into \(A = D + L + R\), where \(D\) is a diagonal part of \(A\), and \(L(R)\) is a left (right) triangular part of \(A\) with zeros on the main diagonal. The Gauss–Seidel method is defined by setting \(W = L + D\) in (1.11).

Theorem 1.3 helps us to show the convergence of the Gauss–Seidel method for any \(A = A^* > 0\). Indeed, if \(A > 0\), then \(D > 0\). Thanks to the identity (see Exercise 1.1.4)
\[
\text{Re} \langle Wx, x \rangle = \langle W_0 x, x \rangle \quad \forall x \in \mathbb{C}^n,
\]
the assumption \(W > \frac{1}{2}A\) of Theorem 1.3 is equivalent to the inequality \(W_0 > \frac{1}{2}A\). For the Gauss–Seidel method, we have \(W = L + D\). Due to the conjugate symmetry of \(A\), it holds that \(R^* = L\) and \(W_0 = \frac{1}{2}(R + L + 2D)\). This implies the necessary assumption of Theorem 1.3:
\[
W_0 - \frac{1}{2}A = W_0 - \frac{1}{2}(D + L + R) = \frac{1}{2}D > 0.
\]

One iteration of the Jacobi method can be written as
\[
Dx^{i+1} + (L + R)x^i = b,
\]
and one iteration of the Gauss–Seidel method can be defined as
\[
(L + D)x^{i+1} + Rx^i = b.
\]

This form of the methods is more convenient for computations.

Even for a symmetric matrix \(A\), the iteration matrix of the Gauss–Seidel method is not symmetric in general, since the preconditioner \(W\) is not symmetric. Sometimes it is convenient to consider the symmetric variant of the Gauss–Seidel method.

**Example 1.7.** One iteration of the symmetric Gauss–Seidel method consists of two subsequent steps of the plain Gauss–Seidel method with forward and backward enumeration of the unknowns. This is equivalent to the alternating use of preconditioners constructed
from the diagonal and left-triangle parts of $A$ and from the diagonal and right-triangle parts of $A$: Given the iterate $x^i$, one finds $x^{i+1}$ in two steps from

$$(L + D)x^{i + \frac{1}{2}} + Rx^i = b,$$

$$(D + R)x^{i+1} + Lx^{i+\frac{1}{2}} = b,$$

where $x^{i+\frac{1}{2}}$ is an auxiliary vector.

For a Hermitian matrix $A$ it holds that $R = L^*$, $D = D^*$. In this case, the identity (1.52) from Exercise 1.1.9 shows that for the symmetric Gauss-Seidel method the preconditioning matrix $W$ from (1.11) satisfies $W = W^*$.

**Example 1.8.** There is a natural way to introduce a relaxation parameter in the Gauss-Seidel method. The resulting method is known as successive over relaxation (SOR). The preconditioner $W$ in the SOR method takes the form

$$W = \frac{1}{\omega}(D + \omega L),$$

where $\omega$ is the relaxation parameter.

Similarly to the case of the Gauss-Seidel method, Theorem 1.3 yields the following results: Assume $A = A^* > 0$; the SOR method converges with any $\omega \in (0, 2)$ for an arbitrary choice of the initial guess $x^0$.

**Definition.** A real-valued matrix $A$ is called an $M$-matrix if $a_{ij} \leq 0$ for all $i, j : i \neq j$, $A$ is nonsingular and all elements of $A^{-1}$ are nonnegative.

For this class of matrices some simple convergence results are at hand (see, e.g., [210]):

1. Assume that $W^{-1} \geq 0$ elementwise and $W - A \geq 0$ elementwise; then the method (1.11) converges.

2. Assume that $W$ is obtained from $A$ by replacing certain elements $a_{ij}, i \neq j$, by values $b_{ij}$ satisfying $a_{ij} \leq b_{ij} \leq 0$; then the method (1.11) converges.

In applications of numerical methods to partial differential equations, the property of a matrix $A$ to be an M-matrix inherits, in some sense, the monotonicity property of a numerical scheme. However, the monotonicity property does not hold for many problems. We will not exploit this matrix property further in this book.

### 1.1.7 Relaxation parameters and Chebyshev polynomials

To simplify the presentation in this section, we assume $W = I$ in (1.11). This is not less general, since (1.11) can be rewritten without a preconditioner by introducing the new matrix $\hat{A} = W^{-1}A$ and the right-hand side $\hat{b} = W^{-1}b$. Moreover, if $A = A^* > 0$ and $W = W^* > 0$, then $\hat{A}$ is self-adjoint and positive definite in the inner products $(\cdot, \cdot)_W$ and $(\cdot, \cdot)_A$. In particular, it holds that

$$(\hat{A}x, y)_A = (x, \hat{A}y)_A \quad \forall x, y \in \mathbb{C}^n, \quad \Re(\hat{A}x, x)_A > 0 \quad \forall 0 \neq x \in \mathbb{C}^n.$$ 

Note that Hermitian matrices can be considered as a particular case of self-adjoint matrices with respect to the Hermitian inner product. All spectral properties of the Hermitian
matrices used previously hold for self-adjoint matrices with respect to other scalar products as well. Hence for the important case of a Hermitian matrix \( A \), all results below are easily extended for iterative methods with a Hermitian preconditioner, but using another inner product.

Let us introduce in (1.11) relaxation parameters \( \alpha_i \in \mathbb{R} \), \( i = 0, 1, 2, \ldots \):

\[
x^{i+1} - x^i = \alpha_i (b - Ax^i), \quad i = 0, 1, 2, \ldots
\]  

(1.20)

Parameters \( \alpha_i \) are introduced to accelerate convergence. There are quite a few strategies for choosing \( \alpha_i \). Let us consider some of them.

First, assume \( A = A^* > 0 \), and suppose the following information on the eigenvalues of \( A \) is available: \( \text{sp}(A) \subseteq [m, M] \). If we use the same parameter \( \alpha_i = \alpha_o \) for all \( i > 0 \), then the “optimal” choice would be \( \alpha_o = 2(m + M)^{-1} \). Indeed, for the norm of the error vector we have

\[
\| e^{i+1} \|_2 \leq |I - \alpha A|_2 \| e^i \|_2.
\]

Since the goal is to minimize \( \| e^{i+1} \|_2 \), it is reasonable to minimize the first factor on the right-hand side by optimizing \( \alpha \). We already know that for the spectral norm it holds that

\[
\| I - \alpha A \|_2 = \max_{\lambda \in \text{sp}(I - \alpha A)} | \lambda | = \max_{\lambda \in \text{sp}(A)} | 1 - \alpha \lambda |.
\]

The only information on the eigenvalues available is that \( \text{sp}(A) \subseteq [m, M] \). Therefore, the following choice of \( \alpha \) looks reasonable:

\[
\alpha_o = \arg \min_{\alpha \in \mathbb{C}} \max_{\lambda \in [m, M]} | 1 - \alpha \lambda |
\]

One easily finds the solution

\[
\alpha_o = \frac{2}{m + M}.
\]

Moreover, for the spectral norm of the iteration matrix, we compute

\[
\| I - \alpha_o A \|_2 \leq q_o = \frac{M - m}{M + m}.
\]  

(1.21)

If we perform \( k \) steps with the method (1.20), then the optimal set of \( \alpha_i \), \( i = 0, 1, \ldots, k - 1 \), which intends to minimize the error \( \| e^k \|_2 \), also can be found. This set of \( \alpha \)'s depends on the value of \( k \). From the error evolution equation (1.12) one gets

\[
e^k = (I - \alpha_{k-1} A) \ldots (I - \alpha_0 A) e^0.
\]

Similarly to the case of constant \( \alpha \), we look for the solution to the optimization problem

\[
\{ \alpha_i \}_{i=0,\ldots,k-1} = \arg \min_{\alpha_i \in \mathbb{C}} \| (I - \alpha_{k-1} A) \ldots (I - \alpha_0 A) \|_2 = \arg \min_{\alpha_i \in \mathbb{C}} \left\{ \max_{\lambda \in \text{sp}(A)} \left| (1 - \alpha_{k-1} \lambda) \ldots (1 - \alpha_0 \lambda) \right| \right\}
\]

In general, the eigenvalues of \( A \) are unknown. However, we may use the information on the lower and upper bounds for \( \text{sp}(A) \). Therefore, instead of the previous optimization problem, we look for

\[
\{ \alpha_i \}_{i=0,\ldots,k-1} = \arg \min_{\alpha_i \in \mathbb{C}} \left\{ \max_{\lambda \in [m, M]} \left| (1 - \alpha_0 \lambda) \ldots (1 - \alpha_{k-1} \lambda) \right| \right\}.
\]  

(1.22)
1.1. Simple iterative methods

To find the solution to optimization problem (1.22), we need the Chebyshev polynomials $P_k(t), k \geq 0, \deg(P_k) = k$. The Chebyshev polynomials can be defined through the recursive relation

$$P_{k+1}(t) = 2tP_k(t) - P_{k-1}(t) \quad \text{for } k > 0, \quad P_0(t) = 1, \quad P_1(t) = t.$$  

From the recursive formula, one finds the following equivalent way of writing the Chebyshev polynomials:

$$P_k(t) = \cos(k \arccos t), \quad (1.23)$$

and

$$P_k(t) = \frac{1}{2} \left( t + \sqrt{t^2 - 1} \right)^k + \frac{1}{2} \left( t - \sqrt{t^2 - 1} \right)^k. \quad (1.24)$$

The trigonometric formula (1.23) readily provides the roots of $P_k$,

$$t_i = \cos \left( \frac{\pi (2i - 1)}{2k} \right), \quad i = 1, \ldots, k,$$

and $k+1$ extremum points on $[-1, 1]$,

$$\hat{t}_i = \cos \left( \frac{i \pi}{k} \right), \quad i = 0, \ldots, k.$$

Note that $P_k(\hat{t}_i) = (-1)^i$.

Let us parameterize $\lambda \in [m, M]$ as $\lambda = \frac{M+m}{2} + \frac{M-m}{2} t, t \in [-1, 1]$, and set

$$\tilde{P}_k(\lambda) = P_k \left( \frac{\lambda - (M+m)/2}{(M-m)/2} \right) / P_k \left( \frac{M+m}{M-m} \right).$$

The function $\tilde{P}_k(\lambda)$ is a polynomial of degree $k$ such that $\tilde{P}_k(0) = 1$. Further, we denote the set of all such polynomials by $\mathcal{F}_k$. Exploiting the fact that there are $k+1$ points $\hat{t}_i \in [m, M]$ such that $|\tilde{P}_k(\hat{t}_i)| = \max_{m \leq \lambda \leq M} |\tilde{P}_k(\lambda)|, i = 0, \ldots, k$, one shows that $\tilde{P}_k$ has minimal deviation from 0 on $[m, M]$ among all polynomials from $\mathcal{F}_k$ (try to prove this!), i.e., $\tilde{P}_k$ solves the following optimization problem:

$$\tilde{P}_k = \arg \min_{f \in \mathcal{F}_k} \max_{m \leq \lambda \leq M} |f(\lambda)|.$$

Comparing this to (1.22) we find that the optimal set of $\alpha$-s from (1.22) consists of the reciprocals to the roots of $\tilde{P}_k$. Using the parameterization of $[m, M]$ we get

$$\alpha_i^{-1} = \frac{1}{2} \left( m + M + (M-m) \cos \frac{\pi (2i - 1)}{2k} \right), \quad i = 1, \ldots, k. \quad (1.25)$$

The error can now be estimated:

$$||e^k||_2 = ||(I - \alpha_{k-1}A) \cdots (I - \alpha_0 A)||_2 ||e^0||_2$$

$$\leq \max_{\lambda \in \text{sp}(A)} ||(I - \alpha_{k-1} \lambda) \cdots (I - \alpha_0 \lambda)||_2 ||e^0||_2$$

$$\leq \max_{\lambda \in [m, M]} ||(I - \alpha_{k-1} \lambda) \cdots (I - \alpha_0 \lambda)||_2 ||e^0||_2$$

$$= \max_{m \leq \lambda \leq M} |\tilde{P}_k(\lambda)||e^0||_2 = \frac{1}{|P_k(-\frac{M+m}{M-m})|} ||e^0||_2.$$

For the last equality we used $\max_{t \in [-1, 1]} |P_k(t)| = 1$. 

To bound $|P_k(-\frac{M+m}{M-m})|$, set $t = -(1+\nu)/(1-\nu)\,\nu = m/M$ and compute

$$t^2 - 1 = \frac{(1+\nu)^2 - (1-\nu)^2}{(1-\nu)^2} = \frac{4\nu}{(1-\nu)^2}.$$ 

Hence,

$$\sigma \equiv |t| + \sqrt{t^2 - 1} = \frac{1+\nu + 2\sqrt{\nu}}{1-\nu} = \frac{1+\sqrt{\nu}}{1-\sqrt{\nu}}.$$ 

From (1.24) we have the bound $2|P_k(t)| \geq \sigma^k$. It yields the error estimate for the convergence of (1.20) with the Chebyshev parameters:

$$||e^k||_2 \leq 2 \left(1 - \frac{\sqrt{\nu}}{1+\sqrt{\nu}}\right)^k ||e^0||_2. \quad (1.26)$$

We remark that (1.25) gives the optimal set of parameters for a fixed $k$. In exact arithmetic the order in which the parameters $\alpha$-s are used does not matter. In practice, however, this can be important, and the "natural" ordering from (1.25) may lead to large amplification of roundoff error. Hence, several re-ordering methods leading to more stable calculations have been proposed in the literature; see [128, 167].

1.1.8 Simple bounds for non-Hermitian matrices

For a non-Hermitian matrix $A$, other quantities rather than eigenvalue bounds become more relevant for convergence analysis. Let us consider a simple example. Assume matrix $A$ is positive definite; then the inequalities

$$\gamma_1 ||x||_2^2 \leq \text{Re} \langle Ax, x \rangle \quad \forall \ x \in \mathbb{C}^n, \quad (1.27)$$

$$||x||_2^2 \leq \gamma_2 \text{Re} \langle A^{-1}x, x \rangle \quad \forall \ x \in \mathbb{C}^n \quad (1.28)$$

are valid with some $\gamma_1 > 0$ and $\gamma_2 > 0$ (explain why!). The second condition can also be rewritten as

$$||Ax||_2^2 \leq \gamma_2 \text{Re} \langle Ax, x \rangle = \gamma_2 \langle A_0 x, x \rangle \quad \text{for} \ A_0 = \frac{1}{2}(A + A^*).$$

To get a bound on the spectral norm of the iteration matrix $S = I - \alpha A$, assume $\alpha \in (0, 2/\gamma_2)$ and note the following relations:

$$||Sx||_2^2 = ||(I - \alpha A)x||_2^2 = ||x||_2^2 - 2\alpha \langle A_0 x, x \rangle + \alpha^2 ||Ax||_2^2$$

$$\leq ||x||_2^2 - 2\alpha \langle A_0 x, x \rangle + \alpha^2 \gamma_2 \langle A_0 x, x \rangle = ||x||_2^2 - (2\alpha - \alpha^2 \gamma_2)\langle A_0 x, x \rangle$$

$$\leq ||x||_2^2 - (2\alpha - \alpha^2 \gamma_2)\gamma_1 ||x||_2^2 = (1 - 2\alpha \gamma_1 + \alpha^2 \gamma_1^2) ||x||_2^2.$$ 

Hence for the choice $\alpha = \gamma_2^{-1}$ we obtain

$$||S||_2 \leq \sqrt{1 - \frac{\gamma_1}{\gamma_2}}. \quad (1.30)$$

The value on the right-hand side of (1.30) estimates the convergence factor of the method (1.11) with $W = I$ and an arbitrary nonsingular matrix $A$, satisfying assumptions (1.27) and (1.28).
1.1. Simple iterative methods

For a Hermitian $A$, conditions (1.27) and (1.28) are equivalent to the bounds on the eigenvalues of $A$: $\text{sp}(A) \in [\gamma_1, \gamma_2]$. For a non-Hermitian case, inequalities (1.27) and (1.28) state bounds on a field of values of matrix $A$ and its inverse. The field of values (also known as the numerical range) of a matrix $A$ is the subset of the complex plane

$$\Phi(A) = \{ z \in \mathbb{C} | z = \langle Ax, x \rangle \text{ for some } x \in \mathbb{C}^n, \|x\|_2 = 1 \}.$$ 

Thus, $\Phi(A)$ is the range of the Rayleigh quotient for $A$ and contains all eigenvalues of $A$. For a normal matrix (a matrix $A$ is called normal if $AA^* = A^*A$), $\Phi(A)$ is known to be the convex hull of its eigenvalues. However, in a general case, bounds on a field of values are more restrictive than eigenvalue estimates and more suitable for convergence analysis of various iterative methods for linear algebraic systems with non-Hermitian matrices, as we shall see in the next section.

1.1.9 Quadratic functionals and linear systems

It is important to realize connections between linear systems and optimization problems for certain quadratic functionals.

For a Hermitian matrix $A$, consider the functional

$$f(x) = \frac{1}{2} \langle Ax, x \rangle - \text{Re} \langle b, x \rangle, A = A^* \in \mathbb{C}^{n \times n}.$$ 

The functional $f$ is bounded from below iff the matrix $A$ is nonnegative definite. For the rest of this subsection, we assume that the matrix is positive definite, $A > 0$. In this case, the system $Ax = b$ has the unique solution. We denote this solution by $x$. One easily verifies that

$$f(x) - f(z) = \frac{1}{2} \langle A(x - z), x - z \rangle \equiv E(x) \quad \forall x \in \mathbb{C}^n.$$ 

Since $E(x) > 0$ for any $x \neq z$, we conclude that $z$ is the unique minimizer of $f(x)$, and any minimization method for the quadratic functional $f$ can be used as a method to solve the system of equations with a positive definite Hermitian coefficient matrix.

The functional $E(x) = \frac{1}{2} \|x - z\|^2_A$ is often called the error functional for the system $Ax = b$. It differs from the functional $f$ by a constant. Therefore, any method minimizing $f$ also minimizes $E$.

If $A$ is an arbitrary nonsingular matrix (not necessarily Hermitian), then solving the system $Ax = b$ is equivalent to minimizing the functional

$$R(x) = \|b - Ax\|_2.$$ 

We call it the residual functional for the system $Ax = b$.

1.1.10 Steepest descent and residual minimization

Tight spectra bounds or field of values estimates are not available for a given matrix $A$ in many practical cases. Yet it is possible to make the simple iterations (1.20) convergent and choose $\alpha_i$ even when lacking such information. Doing this, one exploits the minimization properties of the solution to $Ax = b$. Assume $x^i$ is given; then one may ask for such $\alpha_i$ that the error or the residual functional is minimal for $x^{i+1}$ defined by the linear iteration (1.20):

$$\alpha_i = \arg \min_{\alpha, x} F(x^{i+1}), \quad \text{where } F(x) = \begin{cases} E(x) & \text{if } A = A^* > 0, \\ R(x). & \end{cases}$$

(1.31)
Let us restrict ourselves for a moment to the real-valued matrices and vectors and assume \( A \in \mathbb{R}^{n \times n} \) is symmetric positive definite. The direct computations show that the gradient of the error functional at \( x \in \mathbb{R}^n \) is the residual vector \( \nabla F(x) = \nabla E(x) = Ax - b \). Hence the iteration (1.20) can be viewed as the method of minimizing \( F(x) \) by taking steps proportional to the negative gradient (sometimes called antigradient) at the current point \( x^i \):

\[
x^{i+1} = x^i - \alpha_i \nabla F(x^i).
\]

Since \( x^{i+1} \) minimizes \( F(x) \) for all \( x \in x^i + \text{span}\{\nabla F(x^i)\} \), the method is often called the 

steepest descent method.

It is easy to obtain a computable formula for \( \alpha_i \). For example, for the residual functional, one finds

\[
R(x^{i+1}) = ||Ax^{i+1} - b||_2^2 = ||A(x^i - \alpha_i r^i) - b||_2^2 = ||r^i - \alpha_i Ar^i||_2^2
\]

\[
= ||r^i||_2^2 - \alpha_i \langle (A + A^*)r^i, r^i \rangle + \alpha_i^2 ||Ar^i||_2^2.
\]

The minimum of the right-hand side is attained for

\[
\alpha_i = \frac{\text{Re} \langle Ar^i, r^i \rangle}{||Ar^i||_2^2}.
\]

(1.33)

Similar computations yield the expression for optimal \( \alpha_i \)'s for a Hermitian case and the error functional \( E(x) \):

\[
\alpha_i = \frac{||r^i||_2^2}{||Ar^i||_2^2}.
\]

(1.34)

Assume that matrix \( A \) is positive definite and thus satisfies conditions (1.27) and (1.28) with some (explicitly unknown) real constants \( \gamma_1 > 0, \gamma_2 > 0 \). Substituting \( \alpha_i \) from (1.33) into (1.32) and repeating arguments from (1.29), one gets

\[
R(x^{i+1}) \leq \left(1 - \frac{\gamma_1}{\gamma_2}\right)^{\frac{i}{2}} R(x^i).
\]

(1.35)

Hence, the method converges at least as fast as (1.20) with a priori given \( \alpha_{opt} = \gamma_2^{-1} \).

A similar conclusion holds for the steepest descent method. Indeed, assume \( A = A^* > 0 \), and consider optimal \( \alpha_{opt} = \frac{2}{\lambda_{max}(A) + \lambda_{min}(A)} \) and \( \hat{x}^{i+1} = x^i - \alpha_{opt}(Ax^i - b) \). Due to the optimization property (1.31), for the error of the steepest descent method we have

\[
E(x^{i+1}) \leq E(\hat{x}^{i+1}) = ||\hat{x}^{i+1} - z||_A = ||S(x^i - z)||_A \leq ||S||_A ||x^i - z||_A = ||S||_A E(x^i),
\]

(1.36)

where \( S \) is the iteration matrix \( S = (I - \alpha_{opt}A) \). It is easy to see from the definition of the matrix norm in (1.4) that the \( A \)-norm of \( S \) is equal to its spectral norm:

\[
||S||_A = \max_{x \in \mathbb{C}^n} \frac{||Sx||_A}{||x||_A} = \max_{x \in \mathbb{C}^n} \frac{||A^{\frac{1}{2}}SA^{-\frac{1}{2}}x||_2}{||x||_2} = ||A^{\frac{1}{2}}SA^{-\frac{1}{2}}||_2 = ||S||_2.
\]

For the spectral norm of \( S \), we already found the bound (1.21). Using it in (1.36), we have the convergence estimate for the steepest descent method,

\[
E(x^{i+1}) \leq \frac{\lambda_{max}(A) - \lambda_{min}(A)}{\lambda_{max}(A) + \lambda_{min}(A)} E(x^i).
\]
1.1. Simple iterative methods

No explicit knowledge of $\lambda_{\text{max}}(A)$, $\lambda_{\text{min}}(A)$ or their bounds is required by the steepest descent method to achieve this convergence rate.

Our final remark about steepest descent and the residual minimization concerns an implementation detail. The direct calculation using (1.20) and (1.33) require two matrix-vector multiplications on every iterative step. These are usually the most computationally time consuming operations. The method can be implemented with only one matrix-vector multiplication on every iteration. To see this, introduce the auxiliary vector $z^i = A r^i$ and rewrite formula (1.33) (similarly to (1.34)) and recursive relation (1.20) as

$$
\alpha_i = \frac{\text{Re} \langle z^i, r^i \rangle}{\| z^i \|^2_2}, \quad x^{i+1} = x^i + \alpha_i r^i, \quad r^{i+1} = r^i + \alpha_i z^i.
$$

In the implementation based on the above relations, the right-hand side $b$ of the system (1.1) is used only for the calculation of the initial residual: $r^0 = b - A x^0$.

1.1.11 Fixed point methods

In this book, we are mostly concentrated with solving systems of linear algebraic equations. However, we discuss below several basic approaches to more general nonlinear problems. Important ideas behind these approaches will emerge again later in our studies of methods for linear algebraic systems.

Consider a nonlinear mapping $f$ from a vector space $M$ into itself and the equation

$$
f(x) = 0. \tag{1.37}
$$

Of course, the linear system of algebraic equation (1.1) is the particular case of (1.37), with $f(x) = A x - b$, $M = \mathbb{C}^n$. The problem (1.37) can be written in the equivalent form of the fixed point problem:

$$
x = g(x). \tag{1.38}
$$

As an example, one may consider $g(x) = x - \alpha f(x)$, with $0 \neq \alpha \in \mathbb{C}$. Let us choose an initial guess $x_0$ and consider the fixed point method of solving (1.37) iteratively:

$$
x_{k+1} = g(x_k), \quad k = 0, 1, \ldots \tag{1.39}
$$

The solution $z$ to equation (1.38), $z = g(z)$, is called the fixed point of the mapping $g$.

**Theorem 1.9.** Let $M$ be a full metric space with the distance $\rho$. Assume that the mapping $g : M \to M$ is a contraction mapping, i.e.,

$$
\rho(g(x), g(y)) \leq q \rho(x, y) \quad \forall \ x, y \in M, \tag{1.40}
$$

where $0 < q < 1$ is independent of $x$ and $y$. Then the equation $x = g(x)$ always has the unique solution $z$, and for any initial guess $x_0$ the fixed point method (1.39) converges to $z$ as the geometric progression:

$$
\rho(x_k, z) \leq \frac{q^k}{1-q} \rho(x_1, x_0). \tag{1.41}
$$

**Proof.** For any $m \geq k$, we find

$$
\rho(x_m, x_k) \leq \sum_{i=k}^{m-1} \rho(x_{i+1}, x_i) \leq \sum_{i=k}^{m-1} q^i \rho(x_1, x_0) \leq \frac{q^k}{1-q} \rho(x_1, x_0).
$$
Therefore, \( x_\alpha \) is the Cauchy sequence. Since \( M \) is a full space, the sequence converges to some \( z \in M \). It is clear that \( g(z) = z \) (explain why). Passing to the limit for \( m \to \infty \), we get (1.41).

\[ \square \]

### 1.1.12 Convergence and divergence of fixed point methods

Assume \( g \in C^1[z - \delta, z + \delta] \), where \( z \) is the unique fixed point of \( g \). If \( |g'(z)| < 1 \), then for a sufficiently small \( \delta > 0 \), it holds that

\[ q \equiv \max_{|x-z| \leq \delta} |g'(x)| < 1 \Rightarrow |g(x) - g(y)| \leq q |x - y| \quad \forall x, y \in [z - \delta, z + \delta]. \]

In this example, \( g \) is the contraction mapping of the full metric space \( M = [z - \delta, z + \delta] \). Hence, the fixed point iteration converges for arbitrary initial guess \( x_0 \in M \).

If \( |g'(z)| > 1 \), then the fixed point method diverges for any \( x_0 \neq z \) (prove this!).

Similar conclusions about the properties of the method can be drawn in a multidimensional case. Let \( g : \mathbb{R}^n \to \mathbb{R}^n \) be a continuously differentiable mapping in a neighborhood of a unique fixed point \( z = g(z) \). We write

\[ g(x) = (q_1(x), \ldots, q_n(x))^T, \quad x = (x_1, \ldots, x_n)^T. \]

Consider the matrix

\[ g'(x) = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n(x)}{\partial x_1} & \cdots & \frac{\partial f_n(x)}{\partial x_n} \end{bmatrix}. \]

The matrix \( g'(x) \) is the Jacobian matrix of the mapping \( g \) at point \( x \).

Since \( g \) is a continuously differentiable mapping in \( x \), all elements of the Jacobian matrix exist and are continuous in point \( x \).

**Theorem 1.10.** Assume that the mapping \( g : \mathbb{R}^n \to \mathbb{R}^n \) has the unique fixed point \( z = g(z) \) and that \( g \) is continuously differentiable in a neighborhood of \( z \). If any operator norm of \( g'(x_\alpha) \) is less than 1 for all possible initial vectors \( x_\alpha \) from a neighborhood of \( z \), then the fixed point method converges to \( z \).

The proof is done by analogy with the one-dimensional case.

### 1.1.13 Newton’s method

There are many ways to write \( f(x) = 0 \), with a scalar function \( f \), in the equivalent form \( x = g(x) \). One way of doing this is

\[ g(x) = x - \alpha(x) f(x), \quad \text{where} \quad \alpha(x) \neq 0 \quad \forall x. \]

In particular, \( \alpha \) can be any nonzero constant. If \( z \) is an isolated root and \( f'(z) \neq 0 \), then \( \alpha \) can always be chosen in such a way that the sufficient condition of the fixed point iteration convergence is satisfied:

\[ |g'(z)| = |1 - \alpha f'(z)| < 1. \]
1.1. Simple iterative methods

To accelerate the convergence, one may try to minimize \(|g'(z)|\). The minimum (zero value) of \(|g'(z)|\) is achieved for \(z = 1/f''(z)\). However, one can be close to minimum by setting

\[
\alpha = \frac{1}{f'(x_k)} \approx \frac{1}{f(z)}.
\]

This leads us to the Newton method

\[
x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)},
\]

which is the fixed point iteration for the function

\[
g(x) = x - \frac{f(x)}{f'(x)}.
\]

The Newton method is easy to generalize for the multidimensional case. Consider the system of nonlinear equations:

\[
\begin{align*}
&f_1(x_1, \ldots, x_n) = 0, & f(x) = 0, \\
&f_n(x_1, \ldots, x_n) = 0.
\end{align*}
\]

For this system, \(1/f'(x_k)\) is replaced by the inverse Jacobian matrix of the mapping \(f\) in \(x_k\):

\[
x_{k+1} = x_k - [f'(x_k)]^{-1} f(x_k).
\]

**Theorem 1.11.** Denote by \(z\) a solution to the problem \(f(z) = 0\). Assume that the Jacobian of the mapping \(f : \mathbb{R}^n \rightarrow \mathbb{R}^n\) exists and is nonsingular in a closed ball, \(\{x : ||x-z||_\infty \leq \delta\}\). Also assume that for any \(x, y \) from the ball, it holds that

\[
||f'(x) - f'(y)||_\infty \leq c ||x-y||_\infty, \quad c > 0 \quad \text{(Lipschitz condition)}.
\]

Let \(y \equiv \epsilon \max_{||x||_\infty \leq \delta} ||[f'(x)]^{-1}||_\infty\) and \(0 < \epsilon < \min\{\delta, y^{-1}\}\). Then for any initial guess \(x_0 \in \{x : ||x-z||_\infty \leq \delta\}\), the Newton method converges and the errors \(e_k \equiv z - x_k\) satisfy the estimates

(a) \(||e_{k+1}||_\infty \leq \gamma ||e_k||_\infty^2\); \quad (b) \(||e_k||_\infty \leq \gamma^{-1} (\gamma^2 ||e_0||_\infty^2)^{\frac{1}{2}}\).

**Proof.** If the Jacobian is continuous on the interval connecting points \(x, y \in \mathbb{R}^n\), then according to the Lagrange equality there exist points \(\xi_i \in \mathbb{R}^n, i = 1, \ldots, n\), such that

\[
f(x) - f(y) = f(x-y), \quad J = \begin{bmatrix}
\frac{\partial f_1(\xi)}{\partial x_1} & \cdots & \frac{\partial f_1(\xi)}{\partial x_n} \\
\vdots & \ddots & \vdots \\
\frac{\partial f_n(\xi)}{\partial x_1} & \cdots & \frac{\partial f_n(\xi)}{\partial x_n}
\end{bmatrix}.
\]

Therefore, it holds that

\[
e_{k+1} = e_k - [f'(x_k)]^{-1} (f(z) - f(x_k)) \]
\[
= e_k - [f'(x_k)]^{-1} J_k e_k \]
\[
= [f'(x_k)]^{-1} (f'(x_k) - J_k) e_k,
\]

where \(J_k\) is the Jacobian matrix of \(f\) at \(x_k\).
and hence
\[ \|e_{k+1}\|_\infty \leq \|f'(x_k)^{-1}\|_\infty \|f''(x_k) - J_k\|_\infty \|e_k\|_\infty. \] (1.45)

Thanks to the Lipschitz condition, it holds that
\[ \|f'(x_k) - J_k\|_\infty \leq c \max_{1 \leq j \leq n} \|x_k - \xi_j\|_\infty \leq c \|x_k - z\|_\infty. \]

This, the definition of \(\gamma\), and (1.45) yield the inequality (a).

If \(\|z - x_k\|_\infty \leq \varepsilon\), then
\[ \|z - x_{k+1}\|_\infty \leq (\gamma \varepsilon) \varepsilon \leq \varepsilon. \]
Since the estimate (a) is valid for every \(k\), we immediately get estimate (b).

### 1.1.14 The Anderson acceleration

To accelerate the convergence of the fixed point method (1.39) for finding zero of \(f(x)\), one tries to exploit the information from the \(m\) steps preceding the current step \((k \rightarrow k + 1)\). For the fixed point method, \(m = 0\). Setting \(m > 0\) may help to find a better iterate \(x^{k+1}\). This can be done by a variety of (vector) extrapolation methods, which produce a better convergent sequence of vectors given \(x^k, \ldots, x^{k-m}\), see [47] and the references therein. There is hope for an even better result if one uses the information about the residuals of the current and the previous \(m\) iterates, \(f(x^k), \ldots, f(x^{k-m})\). A distinct algorithm from a class of multistep methods is the Anderson acceleration (also known as Anderson mixing) of fixed point iteration [5].

Assume we computed \(k\) iterates \(x^k, \ldots, x_{k-m}\). Denote by \(f_k = f(x^k)\) the residuals of the iterates. Now we look for the new iterate \(x^{k+1}\) as the linear combination
\[ x^{k+1} = \sum_{j=0}^{m} \alpha_j g(x^{k-j}), \] (1.46)
with a set of complex scalar parameters \(\alpha_j\), such that
\[ \sum_{j=0}^{m} \alpha_j = 1. \] (1.47)

Mapping \(g\) in (1.46) is the same as was used in (1.39). So far, the method (1.46), (1.47) is the straightforward generalization of the simple fixed point iteration (1.39), which corresponds to \(\alpha_0 = 1, \alpha_j = 0, j > 0\). In the Anderson acceleration method, the parameters \(\alpha_j\) are sought to minimize the functional of residuals \(f_k\):
\[
\min_{\sum_{j=0}^{m} \alpha_j = 1} \left\| \sum_{j=0}^{m} \alpha_j f_{k-j} \right\|_2. \] (1.48)

Obviously, the set of parameters solving the minimization problem (1.48) varies depending on \(k\). When it is important to stress this dependence, we index the solution of (1.48) with upper index \(k\):
\[ \alpha_0^{(k)}, \ldots, \alpha_m^{(k)}. \]

Regrouping the residual terms inside the norm in (1.48), we note that the constraint minimization problem (1.48) is equivalent to finding the minimum
\[
\min_{\beta_1, \ldots, \beta_m} \left\| f_k - \sum_{j=1}^{m} \beta_j (f_{k-j+1} - f_{k-j}) \right\|_2 \] (1.49)
without any constraint on scalars \( \beta_j \). The vector of coefficients \( \beta = (\beta_1, \ldots, \beta_m)^T \) is the solution to the linear algebraic system (check this!)

\[
F^T F \beta = F^T f_k, \quad \text{with matrix } F = [f_k - f_{k-1}, \ldots, f_{k-m+1} - f_{k-m}] \in \mathbb{C}^{1 \times m}.
\]

The rationality behind the Anderson acceleration method (1.46), (1.47) is seen from the following considerations [213]. Denote by \( f'(x) \) the Jacobian of \( f \) at \( x \). Assume that the iterates \( x^k, \ldots, x^{k-m+1} \) are sufficiently close to each other and that the defects \( \|r_k\|_2 = \|x^k - g(x^k)\|_2 \) are sufficiently small; thus the following linear analysis makes sense:

\[
 f(x^{k+1}) = f\left( \sum_{j=0}^{m} \alpha_j g(x^{k-j}) \right) = f\left( \sum_{j=0}^{m} \alpha_j x^{k-j} - \sum_{j=0}^{m} \alpha_j r_{k-j} \right) \\
 \approx f\left( \sum_{j=0}^{m} \alpha_j x^{k-j} \right) = f\left( x^k - \sum_{j=1}^{m} \beta_j (x^{k-j+1} - x^{k-j}) \right) \\
 \approx f(x^k) - \sum_{j=1}^{m} \beta_j (f(x^{k-j+1}) - f(x^{k-j})) \\
 = \sum_{j=0}^{m} \alpha_j f_{k-j}.
\]

Thus, solving (1.48), i.e., minimizing the absolute value of the right-hand side of the above chain of relations, one may hope to minimize the residual for the iterate \( x^{k+1} \), i.e., \( \|f(x^{k+1})\|_2 \).

---

**Exercises**

1.1.1. Denote by \( \|x\|_2 \) the Euclidean norm of a vector \( x = (x_1, \ldots, x_n) \) and consider also the norms

\[
\|x\|_1 = \sum_{i=1}^{n} |x_i|, \quad \|x\|_\infty = \max_{1 \leq i \leq n} |x_i|.
\]

Show that the corresponding matrix norms are

(a) \( \|A\|_2 = \sqrt{\rho(AA^T)} \),
(b) \( \|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{n} |a_{ij}| \),
(c) \( \|A\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}| \).

1.1.2. For a Hermitian matrix \( A \) show the following Rayleigh equalities:

\[
\min_{\lambda \in \text{sp}(A)} \lambda = \min_{\|x\|_2 \neq 0} \frac{(Ax, x)}{(x, x)}, \quad \max_{\lambda \in \text{sp}(A)} \lambda = \max_{\|x\|_2 \neq 0} \frac{(Ax, x)}{(x, x)}. \tag{1.50}
\]

1.1.3. For a matrix \( A \in \mathbb{C}^{n \times n} \) prove the Hölder inequality

\[
\|A\|_1^2 \leq \|A\|_\infty \|A\|_1. \tag{1.51}
\]
1.1.4. For an arbitrary matrix $A$ check that
(a) the matrix $AA^*$ is Hermitian nonnegative definite;
(b) for a square matrix $A$, $A = A_0 + \beta A_n$, the matrices $A_0 = \frac{1}{2}(A + A^*)$ and $A_n = -\frac{3}{2}(A - A^*)$ are Hermitian; furthermore check the following equalities:

$$\text{Re} \langle Ax, x \rangle = \langle A_0 x, x \rangle, \quad \text{Im} \langle Ax, x \rangle = \langle A_n x, x \rangle \quad \forall x \in \mathbb{C}^n.$$

1.1.5. Assume that $\text{sp}(A) \subset (0, +\infty)$. Prove that the Richardson method converges with any initial guess $x_0$ if $w = \|A\|^{-1}$ for some norm $\|\cdot\|$.

1.1.6. Assume that $A$ has diagonal dominance:

$$|a_{ii}| > \sum_{j=1, j \neq i}^{n} |a_{ij}|, \quad i = 1, \ldots, n;$$

then show that the Jacobi method converges.

1.1.7. Let

$$A = \begin{pmatrix} \alpha & \beta & 0 \\ \beta & \alpha & \beta \\ 0 & \beta & \alpha \end{pmatrix}.$$

For which values of $\alpha$ and $\beta$
(a) does the Jacobi method converge?
(b) does the Gauss–Seidel method converge?

1.1.8. Assume that $A$ is strictly diagonal dominant:

$$q|a_{ii}| > \sum_{j=1, j \neq i}^{n} |a_{ij}|, \quad i = 1, \ldots, n,$$

with a coefficient $q \in (0, 1)$. Prove that the iteration matrix $S = (L + D)^{-1}R$ of the Gauss–Seidel method satisfies the estimate

$$\|S\|_\infty < q.$$

Therefore, the Gauss–Seidel method converges with the convergence factor $q$.

1.1.9. Prove that the symmetric Gauss–Seidel method can be written as the basis iteration (1.11) with the preconditioner

$$W = (L + D)D^{-1}(D + R). \quad (1.52)$$

1.1.10. To solve system $Ax = b$, with some nonsingular $A \in \mathbb{C}^{n \times n}$, consider the equivalent system $A^T Ax = A^T b$ and apply the steepest descent method. Assuming $A$ is positive definite, how does this approach compare to the residual minimization algorithm (1.20), i.e., $\alpha_i$ is defined in (1.33), in terms of convergence estimates and computational complexity?

### 1.2 Subspaces and iterative methods

We recall that solving the linear algebraic system $Az = b$ is equivalent to minimizing a functional $f(x)$ on $\mathbb{C}^n$ (see Section 1.1.9). A natural functional of choice can be the residual functional $f(x) = \|b - Ax\|_2$, and in the Hermitian positive definite case one may prefer the error functional. Two simple gradient methods were deduced from this
1.2. Subspaces and iterative methods

 observation in Section 1.1.10. Both gradient methods can be observed as the minimization of $f(x)$ on every iteration over a one-dimensional affine space: New iterate $x^{i+1}$ satisfies

$$x^{i+1} = \arg\min \{ f(x) : x \in x^i + \text{span}\{r^i\} \}, \quad r^i = b - Ax^i.$$ 

A natural idea is to try minimizing $f(x)$ over affine spaces of higher dimensions. This idea will be developed further in this chapter. It will lead us to Krylov subspace iterative methods.

For the analysis, it is often convenient to reformulate optimization problems in a way such that a minimum is sought over a vector space rather than over an affine space. For one-dimensional optimization, the equivalent form is

$$y = \arg\min \{ \tilde{f}(x) : x \in \text{span}\{r^i\} \}, \quad r^i = b - Ax^i,$$

$$x^{i+1} = x^i + y,$$

where $\tilde{f}(x)$ is the error or residual functional for the system $Ax = r^i$. This simple observation will be exploited also for minimization over spaces of higher dimensions.

1.2.1 Minimization and projection

Many iterative algorithms for finding a minimum of a functional $f$ solve a local minimization problem on every iteration. The local minimization can be done on a subspace of fixed (same for any iteration) dimension, e.g., of dimension 1 for the steepest descent method. However, by increasing the dimension of subspaces on every iteration we may obtain a better approximation to the solution $z$.

Assume we build a sequence of subspaces

$$L_1 \subset L_2 \subset \cdots \subset L_k \subset \mathbb{C}^n, \quad \dim L_i = i, \quad i = 1, \ldots, k,$$

and on iteration $k$ we find $x_k = \arg\min \{ f(x) : x \in L_k \}$. For a functional $f$ bounded from below, this approach leads to the solution $z$ in at most $n$ iterations.

Extending the idea, one may consider various projection methods. With the goal of solving the system $Ax = b$, we introduce projectors $Q_k, P_k$ of rank $k$ and assume that the projected system

$$(Q_k A P_k) y = Q_k b, \quad y_k = P_k y,$$

has the unique solution $y_k$. One may hope that $y_k$ would be a good approximation to $z$.

Under certain assumptions, solving the projected system is equivalent to minimizing $f$ on a subspace $L_k = \text{im} P_k$.

1.2.2 Krylov subspaces

The subspaces

$$\mathcal{K}_i \equiv \mathcal{K}_i(b, A) \equiv \text{span}\{b, Ab, \ldots, A^{i-1} b\}, \quad i = 1, 2, \ldots,$$

are known as the Krylov subspaces.

If we wish to solve the system $Ax = b$, then the idea of minimizing the residual functional on a system of subspaces takes the form

$$x_i = \arg\min \{ ||b - Ax||_2 : x \in \mathcal{K}_i \}, \quad i = 1, 2, \ldots.$$  \hspace{1cm} (1.53)
For a nonsingular matrix $A$, the vectors $x_i$ are uniquely defined (check this!). A method that produces the sequence $x_i$ is natural to call the **minimal residual method**. Another widely used name for this method is the **generalized minimal residual method** (GMRES). However, our approach to this method is pretty straightforward and does not appear as a generalization of another technique. Thus, in the book we mostly call the method **minimal residual**, while GMRES will refer to a particular algorithmic implementation of the method presented in subsection 1.2.5.

One may ask: Why does the minimal residual method necessarily give the solution to the system $Ax = b$? To see this, note that embeddings $\ldots \mathcal{X}_{i-1} \subset \mathcal{X}_i \subset \mathcal{X}_{i+1} \ldots$ imply that for some $i$ we get $\mathcal{X}_i = \mathcal{X}_{i+1}$. By the construction of the Krylov subspaces, $\mathcal{X}_i = \mathcal{X}_{i+1}$ implies $A \mathcal{X}_i \subset \mathcal{X}_i$. Since the matrix $A$ is nonsingular, it holds that $\dim(A \mathcal{X}_i) = \dim(\mathcal{X}_i)$ and thus $A \mathcal{X}_i = \mathcal{X}_i$. For $b \in \mathcal{X}_i$, the identity $A \mathcal{X}_i = \mathcal{X}_i$ implies $Ax = b$ for some $x \in \mathcal{X}_i$, and hence $x_i$ from (1.53) solves the system.

An algorithmic realization of the minimal residual method will be discussed later in this section.

### 1.2.3 Optimal subspaces

Suppose that we wish to solve the system $Ax = b$ with a nonsingular matrix $A$ by minimizing the residual on a sequence of subspaces. How should we choose the subspaces?

To address this question, consider a generic algorithm $\Phi$ which recursively produces a sequence of embedded subspaces $L_i$:

$$ L_{i+1} = \Phi(b, L_i, AL_i). $$

(1.54)

Consider a basis of $L_i$, $L_i = \text{span}\{p_1, \ldots, p_i\}$.

Since $L_i \subset L_{i+1}$, the new subspace is defined by a vector

$$ p_{i+1} = \phi_{i+1}(b, p_1, \ldots, p_i, Ap_1, \ldots, Ap_i), $$

and $\phi_{i+1}$ defines a method for computing $p_{i+1}$.

Therefore, for building a new subspace, it should be sufficient to perform one matrix-vector multiplication with the matrix $A$ and vector $p_{i+1}$. In addition, the result of this multiplication can be used to minimize the residual on $L_{i+1}$. To assess the quality of a sequence of subspaces generated by $\Phi$, fix some $\varepsilon > 0$; we define the **index** of an algorithm $\Phi$ as follows:

$$ m(\Phi, A, b) \equiv \min \{ i : \min_{y \in L_i} \| b - Ay \|_2 \leq \varepsilon \}. $$

A poor algorithm $\Phi$ may have $m(\Phi, A, b) = +\infty$.

For any given matrix and a right-hand side, the best algorithm is the one with the minimal index. However, any particular algorithm may be applied to different matrices, and its optimality should be judged with respect to a class of matrices. The index of an algorithm $\Phi$ on a class of matrices $\mathcal{A}$ is defined as its index in the worst case:

$$ m(\Phi, b) \equiv \sup_{A \in \mathcal{A}} m(\Phi, A, b). $$

An algorithm $\hat{\Phi}$ is called **optimal** on a class of matrices $\mathcal{A}$ if

$$ m(\hat{\Phi}, b) \leq m(\Phi, b) \quad \forall \Phi, \forall b. $$

In the late 1970s, Nemirovskii and Judin figured out that the information about a system of linear equations contained in the Krylov subspaces is optimal (in a certain sense) regardless of how one wishes to use this information. Here we use this information in the following particular way: Given a Krylov subspace we minimize the residual, $Ax - b$, ...
over this subspace. The next subsection shows a particular optimality result for Krylov subspaces.

1.2.4 The optimality of the Krylov subspaces

A class of matrices $\mathcal{A}$ is said to be unitary invariant if for any matrix from the class it contains all unitary similar matrices: $A \in \mathcal{A} \Rightarrow Q^*AQ \in \mathcal{A}$, with any unitary matrix $Q$. An algorithm $\Phi_{X}$, which generates Krylov subspaces, is almost optimal on any unitary invariant class of matrices $\mathcal{A}$ in the following sense:

$$m(\Phi, b) \leq 2m(\Phi, b) + 1 \quad \forall \Phi, \forall b. \quad (1.55)$$

This estimate on the efficiency index of $\Phi_{X}$ immediately follows from the theorem below.

**Theorem 1.12.** For any nonsingular matrix $A$, any vector $b$, and any subspaces generating algorithm $\Phi$ given by (1.54) there exists a unitary matrix $Q$ such that

$$m(\Phi_{X}, A, b) \leq 2m(\Phi, Q^*AQ, b) + 1. \quad (1.56)$$

**Proof.** First we formulate the idea of the proof. Consider an arbitrary unitary matrix $Q$ such that $Qb = b$. Since a unitary matrix preserves the Euclidean norm of a vector, $\|Qx\|_2 = \|Q^*x\|_2 = \|x\|_2$, for any $r = 1, 2, \ldots$, it holds that

$$\min_{v \in \mathcal{K}_r} \|b - Av\|_2 = \min_{u \in Q^*\mathcal{K}_r} \|b - (Q^*AQ)u\|_2. \quad (1.57)$$

Thus the minimization over $\mathcal{K}_r$ for $A$ is as efficient as the minimization for matrix $S := Q^*AQ$ over $Q^*\mathcal{K}_r$. To prove the theorem, one builds a unitary matrix $Q$ such that either $m(\Phi, S, b) = \infty$ or for some index $r$ it holds that

$$Q^*\mathcal{K}_r = \text{span} \{b, L_m, SL_m\}, \quad (1.58)$$

where $m := m(\Phi, S, b)$ and $L_m$ is the subspace produced by $\Phi$ in (1.54) for the matrix $S$ and vector $b$.

Assume such a unitary matrix $Q$ is shown to exist. If $m(\Phi, S, b) = \infty$, then the theorem is proved. Otherwise, by the definition of the index $m(\Phi, S, b)$, it holds that

$$\min_{y \in \mathcal{K}_m} \|b - Sy\|_2 \leq \varepsilon, \quad (1.59)$$

and so equalities (1.58) and (1.57) imply

$$\min_{v \in \mathcal{K}_r} \|b - Av\|_2 = \min_{u \in \text{span} \{b, L_m, SL_m\}} \|b - Su\|_2 \leq \min_{u \in L_m} \|b - Su\|_2 \leq \varepsilon. \quad (1.60)$$

Hence one proves the desired result:

$$m(\Phi_{X}, A, b) \leq \dim \mathcal{K}_r \leq \text{dim span} \{b, L_m, SL_m\} \leq 2m + 1. \quad (1.61)$$

The proof below develops the idea sketched above and uses some arguments from [60]. As said, to prove (1.56), we build a unitary matrix $Q$ such that for $S = Q^*AQ$ one has either $m(\Phi, S, b) = +\infty$ or

$$m(\Phi_{X}, A, b) \leq \dim \text{span} \{b, L_m, SL_m\}, \quad (1.60)$$

where $L_m$ is a first subspace produced by the algorithm $\Phi$ for $S$ such that (1.59) holds.

We shall build a sequence of unitary matrices $Q_i$, $i = 1, \ldots$, such that the following relations are satisfied:

$$Q_i b = b, \quad (1.61)$$

$$S_i = Q_i^*AQ_i, \quad (1.62)$$

$$Q_i \mathcal{M}_i = \mathcal{K}_{(i)}, \quad \mathcal{M}_i := \text{span} \{b, L_i, S_{i-1}L_{i-1}\}, \quad (1.63)$$
for a certain index $r(i)$. Here $L_i$ is a subspace produced by the algorithm $\Phi$ on step $i$ applied to the matrix $S_i$ and the vector $b$. If, for some $i$, the bound (1.59) holds with $m = i$, $S = Q_i$, then the $i$th matrix in the sequence will be the desired unitary matrix, $Q := Q_i$. Otherwise, if there is no index $i$ such that the $\varepsilon$-accuracy for $S_i$ is attained on the step $i$, then for a certain $i$ we would get $Q_i = Q_{i+1} = \ldots$. Therefore, for $S_i$ the $\varepsilon$-accuracy is not attained on any step. Hence, $m(\Phi, S_m, b) = +\infty$.

Thus, we need to build the sequence of unitary matrices $Q_i$ satisfying (1.61)–(1.63). To this end, as long as $\dim \mathcal{X}_i = i$, consider an orthonormal basis $v_1, \ldots, v_i$ of $\mathcal{X}_i(A, b)$ such that

$$\mathcal{X}_i(A, b) = \text{span} \{v_1, \ldots, v_i\}, \quad v_i = b.$$  

In parallel to defining $Q_i$, $i = 1, \ldots, m$, we build an auxiliary orthonormal sequence of vectors $u_1, \ldots, u_{r(m)}$ such that

$$\mathcal{M}_i = \text{span} \{u_1, \ldots, u_{r(i)}\}. \quad (1.64)$$

With the help of induction, we define below a sequence of $S_i$ such that

$$S_i L_j = S_j L_j \quad \text{for} \quad 1 \leq j \leq i. \quad (1.65)$$

Thanks to the property (1.65), the algorithm $\Phi$ applied to $S_j$ and to $S_i$ produces the same subspace $L_j = \text{span} \{p_1, \ldots, p_j\}$ for $1 \leq j \leq i$.

First, set $r(1) = 1$, $Q_1 = I$, and $u_1 = v_1$, $L_0 := 0$. Suppose that we already defined $S_j = Q_j^* A Q_j$, $1 \leq j \leq i$, satisfying the desired property (1.65). To build $S_{i+1}$, we note that (1.65) allows us to write

$$\mathcal{M}_{i+1} = \text{span} \{b, L_{i+1}, S_i L_i\} = \text{span} \{b, L_i, p_{i+1}, S_{i-1} L_{i-1}, S_i p_1\} = \text{span} \{M_i, p_{i+1}, S_i p_1\}, \quad (1.66)$$

where the vector $p_{i+1}$ is generated by the algorithm $\Phi$ applied to $S_i$. Further, we can choose a basis in $\text{span} \{p_{i+1}, S_i p_1\}$ orthogonal to $\mathcal{M}_i$. These (at most two) basis vectors are the new $u$’s added to the sequence in (1.64). The index $r$ is increased accordingly. Now, the new matrix $Q_{i+1}$ is defined as an arbitrary unitary matrix, satisfying

$$Q_{i+1} u_j = v_j, \quad 1 \leq j \leq r(i+1). \quad (1.67)$$

Such a $Q_{i+1}$ exists, since both sets of vectors are orthonormal. We set $S_{i+1} := Q_{i+1}^* A Q_{i+1}$. To verify the property (1.65), we note that (1.67) yields $Q_{i+1} L_j = Q_j L_j$ and $Q_{i+1} S_j L_j = Q_j S_j L_j$ for $1 \leq j \leq i$. Hence

$$S_j L_j = Q_{i+1}^* Q_j S_j L_j = Q_{i+1}^* A Q_j L_j = Q_{i+1}^* A Q_{i+1} L_j = S_{i+1} L_j.$$  

We have finished the induction step of building the sequence of $Q_i$ and $S_i$ satisfying (1.65). The desired relations (1.61)–(1.63) are satisfied by the construction of $\mathcal{M}_i$ and $Q_i$ in (1.66) and (1.67).

### 1.2.5 The minimal residual method, GMRES

To solve the system $Ax = b$, with a nonsingular matrix $A$, we choose an arbitrary initial vector $x_0$ and consider the equivalent problem

$$Ay = r_0, \quad r_0 = b - Ax_0, \quad x = x_0 + y.$$
Further we treat the modified system $Ay = r_0$. Krylov subspaces are also built for the modified system (assuming $r_0 \neq 0$). Therefore, the minimization problem (1.53) takes the following form: Find $y \in \mathcal{X}_i$ such that $\|r_0 - Ay\|_2 \rightarrow \min$, and set $x_i = x_0 + y$. This is equivalent to solving for

$$x_i = \arg\min \{ \|b - Ax\|_2 : x = x_0 + y, y \in \mathcal{X}_i \}, \quad i = 1, 2, \ldots ,$$

(1.68)

By the best approximation property (1.16) with $A = I$ and $V = A\mathcal{X}_i$, the length of the vector $r_i = r_0 - Ay$ is minimal only if

$$r_i \perp A\mathcal{X}_i.$$

Thus, to make the $i$th step, one needs to drop a perpendicular from the vector $r_0$ on the subspace $A\mathcal{X}_i$. Doing this is easy if one knows an orthogonal basis of $A\mathcal{X}_i$. In this form the minimal residual method was presented in the book [141] as a multistep iteration that minimizes the residual on a fixed dimension Krylov subspace defined by the current residual vector. An implementation of the same method was earlier reported in [112].

The geometric approach for the minimal residual method consists of building a sequence of vectors $q_1, q_2, \ldots$ such that $q_1, \ldots, q_i$ form a basis in the Krylov subspace $\mathcal{X}_i$ and the vectors $p_1 = Aq_1, \ldots, p_i = Aq_i$ form an orthogonal basis in the subspace $A\mathcal{X}_i$. A vector $q_{i+1}$ from this sequence should have the following properties:

$$q_{i+1} \notin \mathcal{X}_i, \quad q_{i+1} \in \mathcal{X}_{i+1}, \quad p_{i+1} = Aq_{i+1} \perp A\mathcal{X}_i.$$

It is clear that we can find such a vector by applying the orthogonalization procedure to the vector $p = Aq$, where $q = Aq_i$. In this geometric approach one needs to store two sequences of vectors $q_1, \ldots, q_i$ and $p_1, \ldots, p_i$.

The algebraic approach for the minimal residual method uses only one sequence of vectors $q_1, \ldots, q_i$, which now form an orthogonal basis in the subspaces $\mathcal{X}_i$. This approach was suggested in [174]. From this paper the method earned the well-known acronym GMRES (generalized minimal residual). We explain this approach below.

Let $q_1 = r_0/\|r_0\|_2$. In order to build an orthonormal basis $q_1, \ldots, q_{i+1}$ in $\mathcal{X}_{i+1} = \mathcal{X}_i + (r_0, A)$, we orthogonalize the vector $Aq_i$ to the vectors $q_1, \ldots, q_i$.

We compute

$$\tilde{q}_{i+1} = Aq_i - \langle Aq_i, q_1 \rangle q_1 - \cdots - \langle Aq_i, q_i \rangle q_i.$$

If $\tilde{q}_{i+1} = 0$, then $Aq_i \in \mathcal{X}_i \Rightarrow \mathcal{X}_{i+1} = \mathcal{X}_i$ (explain this implication!) and thus $r_i = 0$. Otherwise, set

$$q_{i+1} = \tilde{q}_{i+1}/\|\tilde{q}_{i+1}\|_2.$$

Denoting

$$b_{i+1} = \langle Aq_i, q_{i+1} \rangle, \quad \hat{H}_{i+1} = \|q_{i+1}\|_2,$$

we rewrite the above equality in the matrix-vector notation

$$Aq_i = [q_1 \ldots q_{i+1}] \begin{pmatrix} h_{1i} \\ \vdots \\ h_{i+1 \cdot} \end{pmatrix}.$$

We note that we use the notation $[q_1 \ldots q_{i+1}]$ for a matrix with $i+1$ columns of vectors $q_i$. Thus, the first $i$ steps of the orthogonalization process are written in the following matrix form:

$$AQ_i = Q_{i+1} \hat{H}_i, \quad \hat{H}_i = \begin{bmatrix} H_i & 0 \cdots 0 \\ 0 & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix},$$

(1.69)

where $Q_i \equiv [q_1 \ldots q_i] \in \mathbb{C}^{n \times i}$, $H_i \in \mathbb{C}^{i \times i}$ is the upper Hessenberg matrix, and $\hat{H}_i \in \mathbb{C}^{(i+1) \times i}$ is built by adding to $H_i$ one bottom row of all zeros except the last entry in the row, $b_{i+1}$. 
Consider the QR-decomposition of the rectangle matrix $\hat{H}_i$:

$$\hat{H}_i = U_i R_i,$$

where $U_i \in \mathbb{C}^{i \times i}$ and $R_i \in \mathbb{C}^{(i+1) \times i}$.

Since the columns of $Q_i$ form the basis of $\mathcal{X}_i$, the matrix $Q_i$ defines a mapping from $\mathbb{C}^n$ onto $\mathcal{X}_i$. Hence, $\|r_0 - Ay_1\|_2 \to \min$ over $\mathcal{X}_i$ is equivalent to minimizing $\|r_0 - AQ_i y\|_2$ over $\mathbb{C}^n$. Further, with the help of decomposition (1.70), one shows that the minimum of $\|r_0 - AQ_i y\|_2$ over all $y \in \mathbb{C}^n$ is obtained for $y = y_i$, solving the equation (prove this; see Exercise 1.1.4)

$$R_i y_i = z_i \equiv \|r_0\|_2 U_i^* e_1,$$

where $e_1 = [10 \ldots 0]^T$.

The matrix $R_i$ is nonsingular. Therefore, one computes the $i$th iterate of GMRES from

$$x_i = x_0 + Q_i y_i = x_0 + Q_i R_i^{-1} z_i.$$

Note that the Hessenberg matrices $H_i$ are the leading submatrices of the “largest” Hessenberg matrix corresponding to the last iteration. The main computational costs on an iteration $i$ are related to the orthogonalization procedure, i.e., finding the $i$th column in $\hat{H}_i$. Apart from this, and since the matrices $H_i$ are Hessenberg matrices, the computation of vectors $z_i$ and $y_i$ can be performed in $O(i^3)$ arithmetic operations (find the way of doing this computation in $O(i^2)$ operations!).

From the algorithmic viewpoint, the Gram-Schmidt orthogonalization procedure used to build the Hessenberg matrices $\hat{H}_i$ can be replaced by a more stable modified Gram-Schmidt method [173]. To find the QR-decomposition of Hessenberg matrices in (1.70), a good choice is the use of rotation matrices; see, e.g., [173]. Finally, for implementing a stopping criterion, it is helpful to note that $r_0 = \|r_0\|_2 q_1 = \|r_0\|_2 Q_{i+1} e_1$, and using (1.71) we compute

$$\|r_i\|_2^2 = \|r_0 - AQ_i y\|_2^2 = \|r_0 - Q_{i+1} H_i y_i\|_2^2 = \|r_0 - Q_{i+1} U_i R_i y_i\|_2^2 = \|r_0\|_2^2 \|U_i (I - U_i^* U_i) e_1\|_2^2$$

$$= \|r_0\|_2^2 \|I - U_i U_i^*\| e_1\|_2^2 + \|U_i U_i^* e_1\|_2^2 = \|r_0\|_2^2 (1 - 2 \|U_i e_1\|_2^2 + \|U_i^* e_1\|_2^2).$$

Therefore, evaluating $\|r_i\|_2/\|r_0\|_2$ reduces to computing the Euclidean norm of the first row of $U_i$.

We summarize the method in the form of the following algorithm.

**Algorithm 1.1.** GMRES method.

Input: Nonsingular matrix $A \in \mathbb{C}^{n \times n}$, right-hand side vector $b \in \mathbb{C}^n$, initial approximation $x_0 \in \mathbb{C}^n$, maximum number of iterations $N_{\text{max}}$, stopping criterion.

Set $r_0 := b - Ax_0$, $q_1 := r_0/\|r_0\|$, $\hat{H}_0 = \emptyset$, $Q_1 = [q_1]$ (Initialization);

For $i = 1, \ldots, N_{\text{max}}$ or until stopping criterion is met, do

1. $v := Aq_i$, $b_i := Q_i^* v$,
2. $q_{i+1} := v - Q_i b_i$ (Orthogonalization is completed),
3. $H_i := [\hat{H}_{i-1}, b_i]$ (New upper Hessenberg matrix is ready),
4. \( \tilde{H}_i := \begin{bmatrix} H_i & \|q_{i+1}\|_2 e_i^T \end{bmatrix} \),

5. \( q_{i+1} := q_{i+1}/\|q_{i+1}\|_2 \),

6. \( Q_{i+1} := [Q_i, q_{i+1}] \) (Orthonormal basis for \( \mathcal{K}_{i+1} \) is ready),

7. \( U_i R_i := \tilde{H}_i \) (Compute QR-factorization of \( \tilde{H}_i \)),

8. \( y_i := ||r_0||_2 R_i^{-1} U_i^* e_1 \) (Find \( y_i \) from a system with upper triangle \( R_i \)),

9. \( x_i = x_0 + Q_i y_i \).

End

Output: New approximation \( x_i \) to the solution of \( Ax = b \) after iterations are stopped.

The minimal residual method applied to Hermitian matrices is often called the (non-generalized) minimal residual method, and has the commonly used acronym MINRES. In this case, the matrices \( H_i \) appear to be tridiagonal (Exercise 1.1.6). Therefore, the orthogonalization procedure and finding \( z_i, y_i \) can be computed by short-term recurrent relations. Thus, in the Hermitian case, the computation work does not increase for larger \( i \). In a general case, and if the number of iterations is large, then computational and memory costs for storing all vectors \( q_i \) may become too high. In such a situation, it is reasonable to use the minimal residual method with restarts: One sets a limit for the maximal possible dimension of the Krylov subspace and, if the desired accuracy is not obtained, then the final iterate \( x_i \) is taken as the initial guess for the new sequence of the Krylov subspaces.

1.2.6 The conjugate gradient method

Let \( A = A^* > 0 \). Denote by \( z \) the solution to \( Ax = b \) and choose an arbitrary initial guess \( x_0 \). Similarly to the minimal residual method from the previous section, we consider the equivalent system

\[
A u_0 = r_0, \quad \text{with} \quad r_0 = b - Ax_0, \quad z = x_0 + u_0.
\]

The Krylov subspaces are build for the modified system (assuming \( r_0 \neq 0 \)). The conjugate gradient (CG) method minimizes the \( A \)-norm of the error, \( ||y - u_0||_A \), over all \( y \) from the Krylov subspaces \( \mathcal{K}_i (r_0, A) \). One can rewrite this minimization problem as

\[
x_i = \text{argmin} \{ ||x - z||_A : x = x_0 + y, \ y \in \mathcal{K}_i \}.
\]

Applying the best approximation property (1.16) with \( V = \mathcal{K}_i \), we find that the solution of this minimization problem satisfies

\[
( x_i, y )_A = ( z, y )_A \quad \forall \ y \in \mathcal{K}_i.
\]  

(1.72)

By the definition of \( A \)-orthogonality we get from (1.72):

\[
0 = (Ax_i, y) - (Az, y) = (Ax_i - b, y) \quad \forall \ y \in \mathcal{K}_i.
\]

Hence, equation (1.72) is equivalent to the orthogonality of the residual \( r_i \) to \( i \)th Krylov subspace:

\[
r_i \equiv b - Ax_i \perp \mathcal{K}_i.
\]  

(1.73)

This important orthogonality property of the CG method will be exploited below.
Consider an $A$-orthogonal basis $p_1, \ldots, p_i$ in $\mathcal{K}_i$. Let $y_i$ be the $A$-orthogonal projection of the fixed vector $z-x_0$ to $\mathcal{K}_i$. Then coefficients $\alpha_j$, $1 \leq j \leq i$, in the decomposition

$$y_i = \alpha_1 p_1 + \cdots + \alpha_i p_i$$

are independent of $i$. Therefore, we have

$$x_i = x_0 + y_i \Rightarrow x_i = x_{i-1} + \alpha_i p_i.$$  \hfill (1.75)

To find $\alpha_j$, we multiply the last equality with $A$ and get

$$r_i = r_{i-1} - \alpha_i A p_i.$$ \hfill (1.76)

Thanks to $r_i \perp \mathcal{K}_i$, (1.76) yields

$$\alpha_i = \frac{\langle r_{i-1}, p_i \rangle}{\langle A p_i, p_i \rangle}.$$ \hfill (1.77)

Assume $r_i \neq 0$. Since $r_0 \in \mathcal{K}_1$, using an induction argument and noting $A p_i \in \mathcal{K}_{i+1}$, we conclude from (1.76) that $r_i \in \mathcal{K}_{i+1}$. Because of this and $r_i \perp \mathcal{K}_i$ we may write $p_{i+1}$ in the form

$$p_{i+1} = r_i + \beta_{i+1} p_1 + \cdots + \beta_{ii} p_i.$$ \hfill (1.78)

The orthogonality $r_i \perp \mathcal{K}_i$ implies $\beta_{ij} = 0$ for $j < i$ (prove this by multiplying (1.78) from the left by $(A p_j)^T$ and using the definition of Krylov subspaces!). Instead of $\beta_{ii}$ we will write $\beta_i$. We have

$$p_{i+1} = r_i + \beta_i p_i, \quad \beta_i = -\frac{\langle r_i, A p_i \rangle}{\langle A p_i, p_i \rangle}.$$ \hfill (1.79)

Thanks to $p_i = r_{i-1} + \beta_{i-1} p_{i-1}$ and (1.73), we find $\alpha_i = \frac{\langle r_{i-1}, r_{i-1} \rangle}{\langle A p_i, p_i \rangle}$. If $r_i \neq 0$, then $\alpha_i \neq 0$. Therefore, using (1.76) we get

$$\langle r_{i+1}, A p_i \rangle = \frac{\langle r_i, r_{i-1} \rangle}{\alpha_i} \Rightarrow \beta_i = \frac{\langle r_{i+1}, r_i \rangle}{\langle r_{i-1}, r_{i-1} \rangle}.$$ \hfill (1.80)

Collecting (1.75), (1.76), (1.77), and (1.80), we obtain the following iterative method, known as the conjugate gradient (CG) method.

**Algorithm 1.2. CG method.**

Input: Matrix $A \in \mathbb{C}^{n \times n}$, $A^* A > 0$, right-hand side vector $b \in \mathbb{C}^n$, initial approximation $x_0 \in \mathbb{C}^n$, maximum number of iterations $N_{max}$, stopping criterion.

Set $r_0 := b - A x_0$, $p_1 := r_0$ (Initialization);

For $i = 1, \ldots, N_{max}$ or until stopping criterion is met, do

1. $\alpha_i := \langle r_{i-1}, r_{i-1} \rangle / \langle A p_i, p_i \rangle$;
2. $x_i := x_{i-1} + \alpha_i p_i$;
3. $r_i := r_{i-1} - \alpha_i A p_i$,
4. \( \beta_i := (r_i, r_i)/(r_{i-1}, r_{i-1}) \),

5. \( p_{i+1} := r_i + \beta_i p_i \).

End

Output: New approximation \( x_i \) to the solution of \( Ax = b \) after iterations are stopped.

It is remarkable that the recursive relations turned out to be “short”: to perform the minimization over the subspace \( \mathcal{K}_i \), we do not need to know a complete basis of \( \mathcal{K}_i \!\)!

### 1.2.7 Krylov subspaces and matrix decompositions

The CG method can also be deduced from the Cholesky matrix decomposition. Let \( A = A^* > 0 \). Consider the unitary matrix \( Q_i = [q_1, \ldots, q_i] \), where \( q_1, \ldots, q_i \) form an orthogonal basis in \( \mathcal{K}_i \). From (1.69) we know that \( A_i = Q_i^* A Q_i \) is a Hermitian tridiagonal matrix (Exercise 1.2.6). Moreover, it holds that \( A_i > 0 \) (explain why!). Consider the Cholesky decomposition: \( A_i = R_i^* R_i \). Thanks to the fact that \( A_i \) is tridiagonal, the upper triangle matrix \( R_i \) is a bidiagonal one:

\[
R_i = \begin{pmatrix}
\gamma_1 & \delta_1 & \delta_2 & & \\
\gamma_2 & \gamma_2 & \delta_3 & & \\
& \ddots & \ddots & \ddots & \\
& & \gamma_{i-1} & \delta_{i-1} & \\
& & & \gamma_i &
\end{pmatrix}
\]

Assume a nonsingular diagonal matrix \( D_i \) (arbitrary at this moment) and set

\[
P_i = [p_1 \ldots p_i] = Q_i R_i^{-1} D_i. \tag{1.81}
\]

From \( A_i = Q_i^* A Q_i \) and \( A_i = R_i^* R_i \), we obtain

\[P_i^* A P_i = D_i^* D_i.\]

Thus, the columns of \( P_i \) are \( A \)-orthogonal, and from (1.81) it follows that

\[
p_{i+1} \frac{\gamma_{i+1}}{d_{i+1}} + p_i \frac{\delta_i}{d_i} = q_{i+1} \Rightarrow p_{i+1} = \frac{d_{i+1}}{\gamma_{i+1} q_{i+1}} + \frac{\delta_i}{d_i} p_i.
\]

Therefore, we obtained recurrent relations for finding factor matrices \( P_i \), which give a sequence of Cholesky decompositions of the matrix \( A \) “on subspaces.” Now let us use these factor matrices for finding an approximate solution of the system \( Ax = b \). To do this, we look for \( x_i \) in the form \( x_i = x_0 + P_i y \) such that the residual \( r_i \) is orthogonal to \( \text{Im} P_i \). Note that \( \text{Im} P_i = \text{span} \{ p_1, \ldots, p_i \} \) and, as we checked in Section 1.2.6, this is equivalent to the minimization of the \( A \)-norm of the error. Repeating arguments from the previous section, we get \( x_i = x_{i-1} + \alpha_i p_i \), and so \( r_i = r_{i-1} - \alpha_i A p_i \). Moreover, if \( r_i \neq 0 \), then we know that the Krylov subspace basis vector \( q_{i+1} \) is parallel to \( r_i \). If we build vectors \( q_1, \ldots, q_{i+1} \) starting from \( q_1 = r_0/\|r_0\|_2 \), then we are allowed to choose \( q_{i+1} = r_i/\|r_i\|_2 \) and define \( D_j \) in such a way that

\[d_1 = \|r_0\|_2; \quad d_{i+1} = \gamma_{i+1}/\|r_{i+1}\|_2, \quad j = 0, 1, \ldots.
\]
Recursive relations for $p_{i+1}$ now take the same form as in (1.79):

$$p_{i+1} = r_i + \beta_i p_i.$$  

Therefore, we see that the CG method, given in Algorithm 1.2, can be directly deduced from constructing a sequence of Cholesky decompositions without resorting to the Krylov subspaces. Moreover, the Cholesky decompositions implicitly provide a stable basis of the Krylov subspaces. However, taking such a point of view on the CG method, it is hard to deduce its minimization and convergence properties.

### 1.2.8 Formal scalar product

It is remarkable that, in the case of Hermitian matrices, the CG and minimal residual methods can be written in “short” recursive relations, which allow us to store only a small number of vectors, independent of the matrix size or iteration number. If instead of a local minimization problem, one considers solving a certain projected system, then it is possible to obtain short recursive relations for non-Hermitian matrices as well.

To introduce the required projected system of equations, it is convenient to consider more general bilinear forms instead of inner products. With a matrix $D \in \mathbb{C}^{n \times n}$, we introduce a formal scalar product in one of two ways:

$$\langle x, y \rangle = y^T D x, \quad x, y \in \mathbb{C}^n,$$  \hspace{1cm} (1.82)

or

$$\langle x, y \rangle = y^* D x, \quad x, y \in \mathbb{C}^n.$$    \hspace{1cm} (1.83)

For a matrix $A \in \mathbb{C}^{n \times n}$ denote by $A'$ its conjugate matrix, defined through the relation

$$\langle Ax, y \rangle = \langle x, A'y \rangle \quad \forall \, x, y \in \mathbb{C}^n.$$  

If $\alpha$ is a scalar, then $\alpha'$ denotes its conjugate value, which is defined by

$$\langle \alpha x, y \rangle = \langle x, \alpha'y \rangle \quad \forall \, x, y \in \mathbb{C}^n.$$  

In the next subsection, we write $x \perp y$ if $\langle x, y \rangle = 0$.

### 1.2.9 The biconjugate gradient method

To solve the system $Ax = b$, with a nonsingular matrix $A \in \mathbb{C}^{n \times n}$, we choose an initial guess $x_0$ and try to approximate $x$ by a vector of the form

$$x_i = x_0 + y_i, \quad \text{where } y_i \in \mathcal{K}_i(r_0, A) = \text{span} \{ p_1, \ldots, p_i \}, \quad r_0 = b - Ax_0.$$  

However, instead of the residual minimization property or, equivalently, the orthogonality condition (1.73), we enforce

$$r_i \equiv b - Ax_i \perp \mathcal{K}_i(r_0', A') = \text{span} \{ p_1', \ldots, p_i' \},$$ \hspace{1cm} (1.84)

where $r_0'$ is chosen at the beginning of the recursion in such a way that $\langle r_0, r_0' \rangle \neq 0$.

Assume that the vectors $p_1, \ldots, p_i$ and $p_1', \ldots, p_i'$ are formally biconjugate in the sense that $\langle Ap_j, p_k' \rangle = 0$ for all $j \neq k$ and $\langle Ap_j, p_k' \rangle \neq 0$ for $j = k$. Then the residual relation

$$r_i = r_0 - Ax = r_0 - \sum_{j=1}^i \alpha_j A p_j$$  

and the orthogonality property (1.84) yield

$$\alpha_{ji} = \alpha_j = \langle r_0, p_j' \rangle / \langle Ap_j, p_j' \rangle.$$
Therefore, as in the CG method we see that coefficients in decomposition (1.74) are independent of $i$. This yields $x_i = x_{i-1} + a_i p_i$ and $r_i = r_{i-1} - a_i A p_i$. The orthogonality assumption (1.84) implies

$$a_i = \frac{\langle r_{i-1}, p'_i \rangle}{\langle A p_i, p'_i \rangle} \quad (1.85)$$

Let us additionally introduce the vectors $r'_j = r'_j - \bar\alpha' A' p'_i$, satisfying orthogonality conditions “dual” to (1.84):

$$\mathcal{H}(r_0, A) \perp r'_j \quad \forall j \leq i. \quad (1.86)$$

If conditions (1.86) are already assured for any $j < i$, then to satisfy them in the next step, we need to choose $\bar\alpha_i = \langle p_i, r'_{i-1} \rangle / \langle A p_i, p'_i \rangle$.

Assume that $\alpha_j, \bar\alpha_j \neq 0$ for all $j \leq i$. Then we can write

$$p_{i+1} = r_i + \beta_i p_i, \quad p'_{i+1} = r'_i + \bar\beta'_i p'_i,$$

and hence $\alpha_i = \bar\alpha_i = \langle r_{i-1}, r'_{i-1} \rangle / \langle A p_i, p'_i \rangle$. Since $\langle A r_i, p'_i \rangle = \langle r_i, A' p'_i \rangle = 0$ for $j < i$, the biorthogonal property of $p$ and $p'$ vectors is preserved for the choice

$$\beta_i = \frac{\langle A r_i, p'_i \rangle}{\langle A p_i, p'_i \rangle}.$$

Furthermore, it holds that

$$\langle A r_i, p'_i \rangle = \langle r_i, A' p'_i \rangle = \frac{\langle r_i, r'_{i-1} \rangle - \langle r_i, r'_i \rangle}{\alpha_i} \quad \Rightarrow \quad \beta_i = \frac{\langle r_i, r'_{i-1} \rangle}{\langle r_{i-1}, r'_{i-1} \rangle}.$$

It is easy to check that $\bar\beta'_i = \beta_i$. Note that $\alpha'_i = \alpha_i, \beta'_i = \beta_i$ for the formal scalar product of the form (1.82) and $\alpha'_i = \bar\alpha_i, \beta'_i = \bar\beta'_i$ (adjoint complex numbers) in the case (1.83).

As a result, we obtain the following algorithm known as the bi conjugate gradient (Bi-CG) method.

**Algorithm 1.3. Bi-CG method.**

Input: Matrix $A \in \mathbb{C}^{n \times n}$, $\det(A) \neq 0$, right-hand side vector $b \in \mathbb{C}^n$, initial approximation $x_0 \in \mathbb{C}^n$, maximum number of iterations $N_{\text{max}}$, stopping criterion.

Set $r_0 := b - A x_0$, $p_1 := r_0$; choose $r'_0 \in \mathbb{C}^n$, set $p'_0 = r'_0$; (Initialization);

For $i = 1, \ldots, N_{\text{max}}$ or until stopping criterion is met, do

1. $\alpha_i := \langle r_{i-1}, r'_{i-1} \rangle / \langle A p_i, p'_i \rangle$,
2. $x_i := x_{i-1} + \alpha_i p_i$,
3. $r_i := r_{i-1} - \alpha_i A p_i$,
4. $r'_i := r'_{i-1} - \alpha'_i A p'_i$,
5. $\bar\beta_i := \langle r_{i-1}, r'_i \rangle / \langle r_{i-1}, r'_{i-1} \rangle$,
6. $p_{i+1} := r_i + \bar\beta_i p_i$,
7. $p'_{i+1} := r'_i + \bar\beta'_i p'_i$.

End

Output: New approximation $x_i$ to the solution of $Ax = b$ after iterations are stopped.
In practice, the bi-CG algorithm is often unstable. It admits two types of breakdown case:

- \( \langle Ap_i, p_i' \rangle = 0 \) for some \( i \);
- \( \langle r_{i-1}, r_{i-1}' \rangle = 0 \Rightarrow \alpha_i = 0, \) with \( r_i \neq 0. \)

In either of these cases, it is not possible to continue the iterations in the above form. To resolve or avoid this emergency case, it is possible to use some block versions of the bi-CG method or versions making use of local residual minimization [67, 183]. Another way to improve the bi-CG method is discussed in the next subsection.

### 1.2.10 The quasi-minimal residual method

In the non-Hermitian case, the price we paid for keeping recurrence relations short is that the residual (or any other appropriate functional) is no longer minimized. As a result, in the bi-CG method the residual can decrease or grow from one step to another in a chaotic way. It is sometimes hard to decide when to stop the iterative process. It is natural to look for vectors \( \hat{x}_i \) instead of \( x_i \) such that for \( \hat{x}_i \), the residual is decreasing monotonically or at least in a more regular way.

Consider an iterative process such that the iterates \( x_0, x_1, \ldots, x_i \) satisfy relations

\[
x_j = x_{j-1} + \alpha_j p_j, \quad \alpha_j \neq 0, \quad 1 \leq j \leq i.
\]

For residuals, this yields \( r_j = r_{j-1} - \alpha_j Ap_j \). Let \( P_i = [p_1, \ldots, p_i], \rho_i = \| r_i \|_2, \) and \( R_i = [\rho_0/\rho_0, \ldots, r_i/\rho_i] \). Then in matrix notation, we can write

\[
[r_0, Ap_1, \ldots, Ap_i] = R_i Z_i, \quad Z_i = \begin{pmatrix}
\rho_0 & \rho_0/\alpha_1 & \ldots & \ldots & \rho_1/\alpha_2 & \ldots & \rho_i/\alpha_i \\
-\rho_1/\alpha_1 & \rho_1/\alpha_1 & \ldots & \ldots & \rho_i/\alpha_i \\
& & \ddots & \ddots \\
& & & \rho_i/\alpha_i & \rho_{i-1}/\alpha_i & \ldots & \ldots & \rho_i/\alpha_i \\
& & & & \rho_i/\alpha_i & \rho_{i-1}/\alpha_i & \ldots & \ldots & \rho_i/\alpha_i
\end{pmatrix}.
\]

According to this equality, if \( \hat{x}_i = x_0 + P_i y \) for some \( y \in \mathbb{C}^n \), then

\[
\hat{r}_i = r(y) \equiv r_0 - AP_i y = R_i Z_i \begin{pmatrix} \rho_0 \\ \rho_0/\alpha_1 \\ \rho_1/\alpha_2 \\ \ldots \\ \rho_i/\alpha_i \\ -1 \end{pmatrix}.
\]

This implies \( \hat{r}_i = R_i v \), where the vector \( v = v(y) \) has the form

\[
v = [\rho_0(1-\xi_1), \rho_1(\xi_1-\xi_2), \ldots, \rho_i(\xi_{i-1}-\xi_i), \rho_i \xi_i]^T,
\]

\[
\xi_j = y_j/\alpha_j, \quad 1 \leq j \leq i.
\]

Let us choose \( \xi_1, \ldots, \xi_i \) such that they minimize the norm \( \| v \|_2 \). Since \( \| R_i \|_2 \leq \sqrt{i+1} \) (explain why this holds, recalling \( \| R_i \|_2 \leq \| R_i \|_F \)), we get

\[
\| \hat{r}_i \|_2 \leq \sqrt{i+1} \min_y \| v(y) \|_2.
\]

This inequality can be considered as the quasi-minimization property, since, instead of minimizing the norm of the residual \( r(y) \), we minimize the norm of the vector \( v(y) \), which is uniquely defined by the relation \( r(y) = R_i v(y) \). One may consider \( \hat{x}_i \) as approximations to the solution of the system and expect more monotone convergence of \( \hat{x}_i \) towards the solution, compared to the convergence of \( x_i \) from the bi-CG method.
Now we deduce recurrent relations for computing \( \hat{x}_i \). Let \( \eta_0 = 1 - \xi_1, \eta_1 = \xi_1 - \xi_2, \ldots, \eta_{i-1} = \xi_{i-1} - \xi_i, \eta_i = \xi_i \). We should minimize the functional \( f = \| \hat{e} \|^2 = \rho_0^2 \eta_0^2 + \cdots + \rho_{i}^2 \eta_i^2 \) subject to the constraint \( \eta_0 + \cdots + \eta_i = 1 \). Introducing Lagrange multiplies [231], we get

\[
\eta_j = \frac{\sigma_j}{s_j},
\]

where

\[
\sigma_j = \frac{1}{\rho_j^2}, \quad s_i = \sum_{j=1}^{i} \frac{1}{\rho_j^2}.
\] (1.88)

Recalling that columns of \( R_j \) are the normalized residuals generated by an iterative process, we get from (1.87)

\[
\hat{r}_i = \sum_{j=0}^{i} \frac{\sigma_j}{s_j} r_j \Rightarrow \hat{x}_i = \sum_{j=0}^{i} \frac{\sigma_j}{s_j} x_j.
\]

From these identities we deduce the following relations:

\[
\hat{x}_i = \left( 1 - \frac{\sigma_i}{s_i} \right) \hat{x}_{i-1} + \frac{\sigma_i}{s_i} x_i, \quad \hat{r}_i = \left( 1 - \frac{\sigma_i}{s_i} \right) \hat{r}_{i-1} + \frac{\sigma_i}{s_i} r_i.
\] (1.89)

Consider the bi-CG method as the general iterative process. Enhancing it with the ideas of the minimal residual method, it is natural to make the adaptation of the bi-CG method to solve nonlinear equations. Applying the ideas of the minimal residual method based on the history of iterates already appeared in this chapter as the Anderson acceleration of a fixed point method in subsection 1.1.14. It turns out that, applied to linear algebraic systems, the Anderson acceleration is equivalent in a certain sense to the minimal residual method. This suggests that the Anderson acceleration is one reasonable extension of Krylov subspace methods for nonlinear problems. Below we state more precisely the equivalence result following arguments from [212].

### 1.2.11 A minimal residual method for nonlinear problems

Assume we are interested in solving a system of nonlinear equations, as in (1.43). Consider the fixed point method (1.39), with \( g(x) = x - f(x) \). Applied to the linear problem, \( f(x) = Ax - \hat{b}, \) the method in (1.39) is the simple iteration (1.11). Thus, it is natural to question whether it is possible to accelerate convergence of a fixed point method following the ideas of the minimal residual method. It is not immediately clear how to extend the definition of Krylov subspaces for nonlinear operators. However, for a nonlinear problem, the idea of the residual minimization based on the history of iterates already appeared in this chapter as the Anderson acceleration of a fixed point method in subsection 1.1.14. It turns out that, applied to linear algebraic systems, the Anderson acceleration is equivalent in a certain sense to the minimal residual method. This suggests that the Anderson acceleration is one reasonable extension of Krylov subspace methods for nonlinear problems. Below we state more precisely the equivalence result following arguments from [212].

Consider the Anderson acceleration of a fixed point method (1.46)–(1.48) with \( f(x) = Ax - \hat{b} \) and \( g(x) = x - f(x) \), where \( A \in \mathbb{C}^{n \times n} \) is nonsingular. Denote by \( x_i \) and \( r_i \) the iterates and residuals produced by the method, starting with some initial guess \( x_0 \). Let \( \hat{x}_i \) and \( \hat{r}_i \) be the iterates and residuals of the minimal residual method applied to solve \( Ax = \hat{b} \), with the same initial guess.

**Theorem 1.13.** Let the parameter \( m \) in (1.46)–(1.48) be such that \( i < m \), and assume \( ||\hat{r}_{j-1}||_2 > ||\hat{r}_j||_2 > 0 \) for \( 0 < j < i \). Then for linear problems the method is equivalent to

\[\text{QMR}\] method differ from our formulas above, but in the exact arithmetic the iterates \( \hat{x}_i \) coincide.

\[\text{QMR}\] method differ from our formulas above, but in the exact arithmetic the iterates \( \hat{x}_i \) coincide.
minimal residual method (GMRES) in the following sense. It holds that
\[
\tilde{x}_i = \sum_{j=0}^{i} \alpha_j^{(i)} x_j \quad \text{and} \quad x_{i+1} = \tilde{x}_i + \tilde{r}_i,
\]
(1.90)
where \( \{\alpha_j^{(i)}\}_{j=0}^{i} \) is the set of parameters minimizing (1.48).

Proof. Using \( f(x) = Ax - b \), we compute for \( j = 0, \ldots, i \),
\[
f(x_j) = Ax_0 - b + A(x_j - x_0) = \tilde{r}_0 + A(x_j - x_0).
\]
We sum up these equalities with coefficients \( \alpha_j^{(i)} \) and, thanks to (1.47), we get
\[
\sum_{j=0}^{i} \alpha_j^{(i)} f(x_j) = \tilde{r}_0 + A \sum_{j=0}^{i} \alpha_j^{(i)} (x_j - x_0) = \tilde{r}_0 + A \sum_{j=1}^{i} \alpha_j^{(i)} (x_j - x_0).
\]
(1.91)
We recall that GMRES minimizes \( \tilde{r} \) over all possible iterates \( \tilde{x}_i = x_0 + y_i \in \mathcal{H}_i \), which is equivalent to finding \( \|\tilde{r}\|_2 \to \min \) over all \( \tilde{x}_i = \tilde{r}_0 + A y_i \in \mathcal{H}_i \). Therefore, we conclude from (1.91) that the Anderson minimization step (1.48) is equivalent to the GMRES minimization problem if the vectors \( y_j = x_j - x_0 \), \( j = 1, \ldots, i \), span the whole space \( \mathcal{H}_i \). If this is the case, then the iterates satisfy
\[
\sum_{j=0}^{i} \alpha_j^{(i)} x_j = \sum_{j=0}^{i} \alpha_j^{(i)} x_0 + \sum_{j=1}^{i} \alpha_j^{(i)} y_j = x_0 + \sum_{j=1}^{i} \alpha_j^{(i)} y_j = \tilde{x}_i.
\]
(1.92)
This would prove the first identity of (1.90). The second immediately follows from (1.92) and (1.46):
\[
x_{i+1} = (I - A) \left( \sum_{j=0}^{i} \alpha_j^{(i)} x_j \right) + b = (I - A) \tilde{x}_i + b = \tilde{x}_i + \tilde{r}_i.
\]
(1.93)
Thus, it remains to show that the vectors \( y_j = x_j - x_0 \), \( j = 1, \ldots, i \), form a basis in \( \mathcal{H}_i \). This is proved by an induction argument. For \( j = 1 \), it holds that \( y_1 = x_1 - x_0 = g(x_0) - x_0 = \tilde{r}_0 \). By the theorem assumption, \( \tilde{r}_0 \neq 0 \), and so \( \mathcal{H}_1 = \text{span} \{\tilde{r}_0\} = \text{span} \{y_1\} \). Assume that \( \mathcal{H}_i = \text{span} \{y_1, \ldots, y_i\} \) is valid for some \( i \geq 1 \), and so (1.92) and (1.93) holds. Therefore, we have
\[
y_{i+1} = x_{i+1} - x_0 = \tilde{x}_i + \tilde{r}_i - x_0 = \sum_{j=0}^{i} \alpha_j^{(i)} x_j + \tilde{r}_i - x_0 = \tilde{r}_i + \sum_{j=1}^{i} \alpha_j^{(i)} y_j.
\]
Since \( \tilde{r}_i \in \mathcal{H}_{i+1} \) and (by induction hypothesis) \( \sum_{j=1}^{i} \alpha_j^{(i)} y_j \in \mathcal{H}_i \), it follows that \( y_{i+1} \in \mathcal{H}_{i+1} \). On the other hand, the assumption \( \|\tilde{r}\|_2 > \|\tilde{r}_{i+1}\|_2 > 0 \) implies \( \tilde{r}_{i+1} \notin \mathcal{H}_i \) (explain why!). Therefore, \( y_{i+1} \notin \mathcal{H}_i \), and we conclude that \( \mathcal{H}_{i+1} = \text{span} \{y_1, \ldots, y_{i+1}\} \). \( \square \)

1.2.12 Finite precision computations

The effect of roundoff errors on the convergence of CG and GMRES is commonly characterized as the delay of convergence. This point of view on finite precision CG originates
from the seminal works by Lanczos [123] and Hestens and Stiefel [104]. It is commonly accepted now, due to Paige [162] and Greenbaum [88], that the finite precision CG method behaves like the CG method in exact arithmetic applied to a matrix with its eigenvalues replaced with tight clusters; see also [70].

A rigorous exposition of this important topic is beyond the scope of our book. However, we find it important to inform the reader of some general observations. Both the CG and GMRES methods construct some orthogonal systems of vectors and hence involve an orthogonalization procedure that can be viewed as a version of the Gram-Schmidt method. A well-known phenomenon recognized in the machine implementations of the latter method is the loss of orthogonality. A remedy is also well known and called re-orthogonalization. Doing the reorthogonalization allows one to diminish to next to zero the effects of roundoff error, but it is too costly, especially when compared to the computationally cheap short recurrence relations of CG. It is also observed that the orthogonality is lost only in some particular directions: those that are determined by the converged Ritz vectors. For certain implementations of GMRES, a connection is also known between the loss of orthogonality and the decrease of the current residual. The maximal attainable accuracy exists in the presence of roundoff errors, and it is important to stop iterations when it is reached. If the accuracy is still not satisfactory, it makes sense (and has been recommended in earlier papers) to make a restart of the method.

We refer the reader to a nice overview of the state-of-the-art analysis of finite precision implementations of the Krylov subspace methods in [131].

---

**Exercises**

1.2.1. Let Az = b and \( f(x) = \frac{1}{2} \langle Ax, x \rangle - \text{Re} \langle b, x \rangle, \ A = A^* \in \mathbb{C}^{n \times n}, \ b \in \mathbb{C}^n \). Prove that \( f(x) - f(z) = \frac{1}{2} \langle (A - A^*) (x - z), x - z \rangle \).

1.2.2. Let \( A = A^* \in \mathbb{C}^{n \times n} \) and \( b \in \mathbb{C}^n \). Prove that the functional \( f(x) = \frac{1}{2} \langle Ax, x \rangle - \text{Re} \langle b, x \rangle \) is bounded from below iff the matrix \( A \) is nonnegative definite.

1.2.3. A functional \( f(x) \) is said to be strongly convex if the functional \( f(x) - \delta \|x\|^2 \) is convex with some \( \delta > 0 \). Show that for a nonsingular matrix \( A \in \mathbb{C}^{n \times n} \), the residual functional \( f(x) = \|b - Ax\|^2 \) is strongly convex. Make a conclusion on the existence of a unique \( x \), solving the minimization problem (1.53).

1.2.4. Let \( A \in \mathbb{C}^{n \times m}, \ n \geq m \). Show that (i) a minimizer of \( f(x) = \|b - Ax\|_2 \) over \( \mathbb{C}^n \) solves the system \( A^T Ax = A^T b \); (ii) if \( A \) is of a full rank, then the minimizer is unique. Using this and \( Q_i e_1 = \|r_0\|^{-1} r_0 \) (in notation of subsection 1.2.5) show that (1.71) holds.

1.2.5. Describe all vectors which are orthogonal and \( A \)-orthogonal at the same time.

1.2.6. Show that the Hessenberg matrix \( H_i \) in the matrix form of the orthogonalization process (1.69) is tridiagonal if \( A = A^* \).

1.2.7. Show that in the CG method the \( i \)th residual is \( A \)-orthogonal to any \( j \)th residual if \( |i - j| > 1 \).

1.2.8. Explain why, after the restart of the minimal residual method, one typically observes the increase of the residual.

1.2.9. Is it possible that the residual in the QMR method is increasing?
1.3 • Analysis of the minimal residual method

1.3.1 • Convergence

The (generalized) minimal residual method is the Krylov subspace method and gives the exact solution of $Ax = b$ in a finite number of steps. For this reason, it is the direct method for solving a system of algebraic equations. However, more often one considers the minimal residual method as an iterative method, and it is natural to look for convergence estimates.

We want to make clear from the very beginning that obtaining reasonable convergence estimates requires additional assumptions on the coefficient matrix $A$. Indeed, in a general case of a system of order $n$, it may happen that after $n - 1$ steps the residual norm remains equal to the norm of the initial residual. Here is an example of the system $Ax = b$:

$$
\begin{bmatrix}
0 & 1 \\
0 & 1 \\
\vdots & \vdots \\
1 & 0 \\
\end{bmatrix}
\begin{bmatrix}
1 \\
0 \\
\vdots \\
0 \\
\end{bmatrix}
= 
\begin{bmatrix}
0 \\
0 \\
\vdots \\
1 \\
\end{bmatrix}.
$$

If $x_0 = 0$, then $r_0 = b$. One gets the Krylov subspaces

$$K_i = \text{span} \{e_i, e_{i-1}, \ldots, e_{n-r+1}\}, \quad 1 \leq i \leq n.$$

Here $e_j$ denotes the $j$th column of the identity matrix. The exact solution to the system takes the form $x = e_1$, and the iterates in the minimal residual method are (check this!)

$$x_0 = x_1 = \cdots = x_{n-1} = 0, \quad x_n = e_1.$$

1.3.2 • Positive definite matrices

For a system with a positive definite matrix, a convergence estimate for the minimal residual method is simple to show. First, note that the positive definiteness of a matrix, $\text{Re} \langle Ax, x \rangle > 0$ for all $x \neq 0$, implies (explain why!) the ellipticity condition (coercivity),

$$\text{Re} \langle Ax, x \rangle \geq \tau \langle x, x \rangle \quad \forall \ x \in \mathbb{C}^n,$$

where $\tau > 0$ is a constant independent of $x$. A matrix $A$ satisfying the ellipticity condition is also called elliptic (coercive).

The $i$th iterate of the minimal residual method, vector $x_i$, minimizes a residual on the set of all vectors of the form $x_0 + y$, $y \in K_i$. Since $r_{i-1} \in K_i$, this implies

$$\|r_i\|^2 \leq \min_{\alpha} \|r_{i-1} - \alpha Ar_{i-1}\|^2$$

$$= \min_{\alpha} \{\|\alpha\|^2\|Ar_{i-1}\|^2 - 2\text{Re} \left( \langle \alpha Ar_{i-1}, r_{i-1} \rangle \right) + \|r_{i-1}\|^2\}. \quad (1.94)$$

One can take

$$\alpha = \frac{\text{Re} \langle Ar_{i-1}, r_{i-1} \rangle}{\|Ar_{i-1}\|^2}.$$

Substituting this in (1.94) yields

$$\|r_i\| \leq \sqrt{1 - \frac{(\text{Re} \langle Ar_{i-1}, r_{i-1} \rangle/\|r_{i-1}\|^2)^2}{\|Ar_{i-1}\|^2/\|r_{i-1}\|^2 \|r_{i-1}\|^2}} \|r_{i-1}\|^2.$$
Therefore, if the matrix \( A \) is elliptic, the following estimate holds: \(^3\)

\[
\|r_i\|_2 \leq \sqrt{1 - \frac{\tau^2}{\|A\|_2^2}} \|r_{i-1}\|_2. \quad (1.95)
\]

We note that the estimate (1.95) relies on the observation that the norm of the residual on the iteration \( i \) of the minimal residual method does not exceed the norm of residual for any vector (new iterate) of the form \( x_{i-1} + \alpha r_{i-1} \), since \( r_{i-1} \in \mathcal{X}_i \). Thus, in essence, (1.95) is the estimate for a method which solves a one-dimensional residual minimization problem on each iteration, but does not account for the fact that the norm of the residual is minimal over the whole Krylov subspace of dimension \( i \).

### 1.3.3 Convergence of the restarted method

Consider the minimal residual method with restarts after every \( p \) steps. If the initial residual is \( r_0 \), then after \( p \) steps it becomes

\[
r_p = \left( I - \sum_{i=1}^{p} y_i A^i \right) r_0,
\]

where the coefficients \( y_i \) are such that

\[
\left\| \left( I - \sum_{i=1}^{p} y_i A^i \right) r_0 \right\|_2 = \min_{\beta_1, \ldots, \beta_p} \left\| \left( I - \sum_{i=1}^{p} \beta_i A^i \right) r_0 \right\|_2.
\]

This has a clear geometrical meaning: \( r_p \) is the perpendicular dropped from \( r_0 \) onto the linear span of the vectors \( A r_0, A^2 r_0, \ldots, A^p r_0 \).

The minimal residual method with restarts at every \( p \) steps produces a sequence of residuals \( r_0, r_p, r_{2p}, \ldots \), where

\[
r_{(k+1)p} = \left( I - \sum_{i=1}^{p} y_{ik} A^i \right) r_{kp}, \quad k = 0, 1, \ldots,
\]

and the coefficients are chosen so that the norm of \( r_{(k+1)p} \) is minimal possible:

\[
\left\| \left( I - \sum_{i=1}^{p} y_{ik} A^i \right) r_{kp} \right\|_2 = \min_{\beta_1, \ldots, \beta_p} \left\| \left( I - \sum_{i=1}^{p} \beta_i A^i \right) r_{kp} \right\|_2. \quad (1.97)
\]

In other words, \( r_{(k+1)p} \) is the perpendicular dropped from \( r_{kp} \) onto the linear span of the vectors \( A r_{kp}, A^2 r_{kp}, \ldots, A^p r_{kp} \).

Of course, we are no longer able to claim that the exact solution is obtained in finitely many steps. Instead, we say that the method is convergent if for an arbitrary vector \( r_0 \) it holds that \( r_p \to 0 \) as \( k \to \infty \). Obviously, \( r_0 = b - A x_0 \) is completely determined by the initial guess \( x_0 \). However, we assume that \( A \) is a nonsingular matrix, and therefore \( x_0 \) is uniquely determined by \( r_0 \) as well.

\(^3\)This estimate coincides with the one proved by Krasnoselski and Krein [114] for the gradient decent method (1.20), (1.31). In 1982 Elman noted that the same estimate remains true for the minimal residual method and general elliptic matrices.
Theorem 1.14 (Kuznetsov). Let \( A \) be a nonsingular matrix. Then the method of minimal residuals with restarts at every \( p \) steps is convergent iff

\[
\tau := \min_{\|\xi\|_2=1} \max_{1 \leq i \leq p} |\langle A^i \xi, \xi \rangle| > 0. \tag{1.98}
\]

Moreover, in the case of convergence there exists a constant \( 0 < q < 1 \) such that

\[
\|r_{(k+1)p}\|_2 \leq q\|r_k\|_2, \quad k = 0, 1, \ldots. \tag{1.99}
\]

Proof. Assume first that \( \tau = 0 \). It stems from the fact that there is a vector \( \xi \) on the unit sphere and orthogonal to the linear span of the vectors \( A\xi, A^2\xi, \ldots, A^p\xi \). Let the initial guess \( x_0 \) be chosen so that \( b - Ax_0 = \xi \). Thus, \( r_0 = \xi \), and, thanks to the choice of \( \xi \), we obtain \( r_p = 0 \). Consequently, \( r_{kp} = \xi \) for all \( k \), and the method does not converge.

Now assume that \( \tau > 0 \). Suppose that \( r_{kp} \neq 0 \) and set \( \xi := r_{kp}/\|r_{kp}\|_2 \). By the assumption (1.98), for some \( 1 \leq i \leq p \) we have \( |\langle A^i \xi, \xi \rangle| \geq \tau \), and hence,

\[
|\langle A^i r_{kp}, r_{kp} \rangle| \geq \tau\|r_{kp}\|_2^2.
\]

From the minimal residual property (1.97),

\[
\|r_{(k+1)p}\|_2^2 \leq \min_\beta \|r_{kp} - \beta A^i r_{kp}\|_2^2,
\]

and the minimum takes place when the vector \( r_{kp} - \beta A^i r_{kp} \) is orthogonal to \( A^i r_{kp} \). Consequently,

\[
\beta = \frac{\langle r_{kp}, A^i r_{kp} \rangle}{\|A^i r_{kp}\|_2^2},
\]

and, according to the Pythagorean theorem,

\[
\|r_{(k+1)p}\|_2^2 \leq \|r_{kp} - \beta A^i r_{kp}\|_2^2 = (1 - |\beta|^2)\|r_{kp}\|_2^2 \leq \left(1 - \frac{\tau^2}{\|A^i\|_2^2}\right)\|r_{kp}\|_2^2.
\]

Now, the inequality (1.99) is guaranteed for any \( r_{kp} \neq 0 \) if we take

\[
q := \sqrt{1 - \frac{\tau^2}{\theta^2}}, \quad \text{with } \theta := \max_{1 \leq i \leq p} \|A^i\|_2. \tag{1.100}
\]

The estimate (1.99), together with the specification of \( q \) in (1.100), can be regarded as a generalization of the result of previous subsection.

Corollary 1.15. If \( A^i \) is positive definite and \( p \geq i \), then the method of minimal residuals with restarts after every \( p \) steps is convergent.

Note that Theorem 1.14 is a particular case of the original result by Kuznetsov [118, 140]; see also an expanded presentation in [141]. The analysis from [141] can be applied to consistent linear systems with possibly singular matrices and allows a special kind of preconditioning and minimization in a more general \( B \)-norm defined by a Hermitian positive definite matrix \( B \).
1.3. Analysis of the minimal residual method

1.3.4 Polynomials and estimates

The starting point for getting more accurate convergence estimates is the following observation. The inclusion $r_i \in r_0 + A \mathcal{X}_i$ and the definition of $\mathcal{X}_i$ implies

$$r_i = f_i(A)r_0,$$  \hspace{1cm} (1.101)

where $f_i(\zeta)$ is a polynomial of degree as most $i$, with the zero order term, $f_i(0) = 1$. Denote by $\mathcal{F}_i$ the set of all such polynomials. Any vector $f_i(A)r_0$, with arbitrary $f_i \in \mathcal{F}_i$, belongs to $r_0 + A \mathcal{X}_i$. Since the minimal residual method minimizes residuals over all vectors from $r_0 + A \mathcal{X}_i$, it holds that

$$\|r_i\|_2 = \min_{r \in A \mathcal{X}_i} \|r_0 + A \mathcal{X}_i\|_2 = \min_{f_i \in \mathcal{F}_i} \|f_i(A)r_0\|_2 \leq \left( \min_{f_i \in \mathcal{F}_i} \|f_i(A)\|_2 \right) \|r_0\|_2. \hspace{1cm} (1.102)$$

Obviously, if $A$ is diagonalizable, i.e., for the Jordan normal form, $A = CJC^{-1}$, the matrix $J$ is diagonal, estimate (1.102) yields

$$\|r_i\|_2 \leq \text{cond}_2(C) \left( \min_{f_i \in \mathcal{F}_i} \max_{\lambda \in \sigma(A)} |f_i(\lambda)| \right) \|r_0\|_2.$$

1.3.5 Polynomials and resolvent

Let $f(z)$ be an arbitrary function of complex variable $z$. Assume $f(z)$ is analytic in an open simply connected domain $\Omega$ with the piecewise smooth boundary $\Gamma$. In this case, $f(z)$ can be expressed in terms of the Cauchy integral:4

$$f(z) = \frac{1}{2\pi i} \int_{\Gamma} \frac{f(\zeta)}{z - \zeta} d\zeta, \hspace{1cm} z \in \Omega.$$  

Let $f(z) = a_0 + a_1 z + \cdots + a_n z^n$ and $f(A) = a_0 I + a_1 A + \cdots + a_n A^n$ be an arbitrary polynomial and the corresponding matrix polynomial for $A \in \mathbb{C}^{m \times n}$. The matrix $(A - \zeta I)^{-1}$ is called the resolvent of the matrix $A$. Obviously, it is well-defined for all $\zeta$ not equal to eigenvalues of $A$.

**Lemma 1.16.** If all eigenvalues of $A$ belong to a bounded domain $\Omega$ with the boundary $\Gamma$, then for any polynomial $f(z)$ it holds that

$$f(A) = \frac{1}{2\pi i} \int_{\Gamma} (A - \zeta I)^{-1} f(\zeta) d\zeta.$$  

**Proof.** The curve $\Gamma$ separates two domains in the complex plane: the bounded domain $\Omega$ and the unbounded complement domain $\Omega' := \mathbb{C} \setminus (\Omega \cup \Gamma)$. For sufficiently large $R > 0$, the elements of matrix $(A - \zeta I)^{-1}$ are analytic functions in the “curved ring” $\Omega_R, \Omega_R = \{\zeta \in \Omega' : |\zeta| < R\}$. For an analytic function $g$ on $\Omega_R$ it holds that $\int_{\Omega_R} g(\zeta) d\zeta \equiv 0$. Therefore,

$$\frac{1}{2\pi i} \int_{\Gamma} (A - \zeta I)^{-1} f(\zeta) d\zeta = \frac{1}{2\pi i} \int_{|\zeta| = R} (A - \zeta I)^{-1} f(\zeta) d\zeta.$$  

4One assumes a counterclockwise parameterization of $\Gamma$. 

---

---
For sufficiently large $|\zeta| = R$, the matrix series $\sum_{k=0}^{\infty} \zeta^{-k-1} A^k$ is convergence, and hence we find

$$(A - \zeta I)^{-1} = -\sum_{k=0}^{\infty} \zeta^{-k-1} A^k.$$  

The direct computation gives

$$\frac{1}{2\pi i} \int_{|\zeta|=1} A^k \zeta^{j-k-1} d\zeta = \begin{cases} A^i, & k = j, \\ 0, & k \neq j. \end{cases}$$

Therefore, it holds that

$$\frac{1}{2\pi i} \int_{|\zeta|=R} \sum_{k=0}^{\infty} A^k \zeta^{-k-1} \sum_{j=0}^{n} a_j \zeta^j d\zeta = \sum_{j=0}^{n} a_j A^j = f(A). \quad \square$$

### 1.3.6 Asymptotic convergence factor

The estimate (1.102) and Lemma 1.16 immediately yield the following theorem.

**Theorem 1.17.** If all eigenvalues of the matrix $A$ belong to a bounded domain $\Omega$, with the boundary $\Gamma$, then the residuals $r_i$ of the minimal residual method satisfy the estimates

$$\frac{||r_i||_2}{||r_0||_2} \leq \frac{||\Gamma|| \max_{\zeta \in \Gamma} ||(A - \zeta I)^{-1}||_2}{2\pi} \min_{f \in \mathcal{F}} \max_{\zeta \in \Gamma} |f(\zeta)|, \quad (1.103)$$

where $||\Gamma||$ denotes the length of $\Gamma$, and $\mathcal{F}$, is the set of all polynomials $f(\zeta)$ of the degree at most $i$, with $f(0) = 1$.

Therefore, proving convergence estimates for the minimal residual method comes down to estimating

$$T_i(\Gamma) \equiv \min_{f \in \mathcal{F}} \max_{\zeta \in \Gamma} |f(\zeta)|. \quad (1.104)$$

Denote by $\mathcal{A}$ the set of all matrices $A$ such that all eigenvalues of every matrix $A$ belong to $\Omega$ and for any matrix $A$ the norm of its resolvent on $\Gamma$ does not exceed a constant $R(\mathcal{A}, \Gamma)$. Further, let $r_i = r_i(A)$ denote the $i$th residual of the minimal residual method applied to a system with arbitrarily chosen matrix $A$ from $\mathcal{A}$. Then

$$\frac{||r_i||_2}{||r_0||_2} \leq \frac{||\Gamma|| R(\mathcal{A}, \Gamma)}{2\pi} T_i(\Gamma) \Rightarrow \lim_{i \to \infty} \left( \frac{||r_i||_2}{||r_0||_2} \right)^{1/i} \leq q,$$

where

$$q = q(\Gamma) \equiv \lim_{i \to \infty} (T_i(\Gamma))^{1/i}. \quad (1.105)$$

The quantity $q = q(\Gamma)$ is called the *asymptotic convergence factor* of the minimal residual method on the set $\mathcal{A}$. This notion is reasonable, since it appears to be possible to construct a sequence of matrices $\{A_n\}, A \in \mathcal{A}, n = 1, 2, \ldots$, such that

$$\lim_{n \to \infty, i \to \infty} \left( \frac{||r_i(A_n)||_2}{||r_0(A_n)||_2} \right)^{1/i} = q. \quad (1.106)$$

We note that the sizes of matrices $A_n$ may increase as $n \to \infty$. 

1.3. Analysis of the minimal residual method

1.3.7 Numerical range

The norm of the resolvent on $\Gamma$ defines a factor in the convergence estimate (1.103), which is independent of the iteration number $i$. However, for some matrices $A$ of size $n$, this coefficient can be large (for example, it can be of order $q^{-n}$) and then the estimate (1.103) is useless. Fortunately, for a particular choice of the curve $\Gamma$, the norm of the resolvent can be easily estimated independently of $n$.

This can be done if the domain $\Omega$ contains the numerical range of matrix $A$. Recall that the numerical range of a matrix $A$ is the set of complex values defined by

$$\Phi(A) = \{ z \in \mathbb{C} | z = \langle Ax, x \rangle \text{ for some } x \in \mathbb{C}^n, \|x\| = 1 \}.$$  

The set $\Phi(A)$ is closed (prove this). Moreover, one can prove the following result.

**Theorem 1.18 (The Toeplitz–Hausdorff theorem).** The numerical range of a matrix $A$ is convex.

**Proof.** The set $\Phi(A)$ contains an interval connecting two points $z_1$ and $z_2$ in the complex plane iff the set $a\Phi(A) + b \equiv \{az + b, z \in \Phi(A)\}$ contains an interval connecting points $az_1 + b$ and $az_2 + b$ for any complex $a, b$ such that $a \neq 0$. If $\|x_1\|_2 = \|x_2\|_2 = 1$ and $\langle Ax_1, x_1 \rangle \neq \langle Ax_2, x_2 \rangle$, then we consider a matrix $B = aA + bI$, and we can always choose $a$ and $b$ in such a way that $\langle Bx_1, x_1 \rangle = 0$ and $\langle Bx_2, x_2 \rangle = 1$. It holds that $\Phi(B) = a\Phi(A) + b$. Furthermore, multiplying $x_1$ by $e^{i\phi}$, with some real $\phi$ we can always assume that the quantity $\langle Bx_1, x_2 \rangle + \langle Bx_2, x_1 \rangle$ is real (note that multiplying $x_1$ with $e^{i\phi}$ does not change corresponding value $z_1 \in \Phi(A)$). It is sufficient to prove that $\Phi(B)$ contains all points of the interval $[0, 1]$. The vectors $x_1$ and $x_2$ are linearly independent (explain why!). Therefore, for $0 \leq t \leq 1$ the function

$$\phi(t) = \frac{\langle B(tx_1 + (1-t)x_2), tx_1 + (1-t)x_2 \rangle}{\|tx_1 + (1-t)x_2\|_2^2}$$

in continuous and real-valued. Moreover, it holds that $\phi(0) = 1$ and $\phi(1) = 0$. Thus $\phi(t)$ attains all values between 0 and 1. [Box]

The numerical range $\Phi(A)$ contains all the eigenvalues of $A$ (prove this!).

1.3.8 An estimate for resolvent

**Theorem 1.19.** Assume that $\zeta \in \mathbb{C}$ does not belong to $\Phi(A)$. Then

$$\|(A - \zeta I)^{-1}\|_2 \leq \frac{1}{d(\zeta, \Phi(A))},$$

where $d(\zeta, \Phi(A)) = \min_{\xi \in \Phi(A)} \|\zeta - \xi\|_2$ is a distance from $\zeta$ to $\Phi(A)$.

**Proof.** Consider the singular value decomposition $(A - \zeta I)^{-1} = U D Q$. Since $U$ and $Q$ are unitary and the diagonal matrix $D$ is such that $\|D\|_2 = \|(A - \zeta I)^{-1}\|_2$, there exist two vectors $x, y$, $\|x\|_2 = \|y\|_2 = 1$ such that

$$(A - \zeta I)^{-1} y = \|(A - \zeta I)^{-1}\|_2 x.$$  

We have

$$\|(Ax, x) - \zeta(x, x)\| = \|(A - \zeta I)x, x\| = \frac{\|y, x\|}{\|(A - \zeta I)^{-1}\|_2} \leq \frac{1}{\|(A - \zeta I)^{-1}\|_2}. $$
Hence, it holds that

$$
\|(A - \zeta I)^{-1}\|_2 \leq \frac{1}{\| (Ax, x) - \zeta \|} \leq \frac{1}{d(\zeta, \Phi(A))}.
$$

\[ \square \]

Theorems 1.17 and 1.19 yield the following corollary.

**Corollary 1.20.** Assume that a distance from any point on the curve \( \Gamma \) to the numerical range \( \Phi(A) \) is at least \( d \). Then

$$
\max_{\zeta \in \Gamma} \|(A - \zeta I)^{-1}\|_2 \leq \frac{1}{d}.
$$

The residuals \( r_i \) of the minimal residual method satisfy the estimates

$$
\frac{\| r_i \|_2}{\| r_0 \|_2} \leq \frac{|\Gamma|}{2\pi d} \min_{f \in \Phi, \zeta \in \Gamma} |f(\zeta)|.
$$

By virtue of the Toeplitz–Hausdorff theorem, the convergence estimate from the corollary may be practical for large enough \( i \) only if the numerical range of \( A \) does not contain \( z = 0 \).

### 1.3.9 Pseudospectra

There is another direct way to relate an estimate for the resolvent to the matrix spectral properties or, more precisely, to the stability properties of the matrix eigenvalue problem. To this end, one defines the **pseudospectra** \( \text{sp}_\varepsilon(A) \) of \( A \in \mathbb{C}^{n \times n} \) as the set of eigenvalues of all possible \( \varepsilon \)-perturbations of the matrix \( A \):

$$
\text{sp}_\varepsilon(A) := \{ z \in \mathbb{C} : z \in \text{sp}(A + E) \text{ for some } E \in \mathbb{C}^{n \times n}, \| E \|_2 < \varepsilon \}.
$$

The following elementary result links the notion of the pseudospectra to the resolvent.

**Theorem 1.21.** For \( A \in \mathbb{C}^{n \times n} \) and \( z \in \mathbb{C} \), it holds that

$$
z \in \text{sp}_\varepsilon(A) \iff \{ z \in \text{sp}(A) \text{ or } \|(A - zI)^{-1}\|_2 > \varepsilon^{-1} \}.
$$

**Proof.** For \( z \in \text{sp}(A) \) the assertion is trivial, so we may assume \( z \notin \text{sp}(A) \). If \( \|(A - zI)^{-1}\|_2 \leq \varepsilon^{-1} \) and \( \| E \|_2 < \varepsilon \), then \( \| E(A - zI)^{-1} \|_2 < 1 \), and the matrix \( I + E(A - zI)^{-1} \) is nonsingular. Therefore, from the identity

$$
A + E - zI = (I + E(A - zI)^{-1})(A - zI)
$$

it follows that \( A + E - zI \) is the product of two nonsingular matrices, and hence it is nonsingular as well. We conclude that \( z \notin \text{sp}_\varepsilon(A) \). Thus, by contradiction, \( z \in \text{sp}_\varepsilon(A) \) yields \( \|(A - zI)^{-1}\|_2 > \varepsilon^{-1} \).

To show the implication \( \|(A - zI)^{-1}\|_2 > \varepsilon^{-1} \Rightarrow z \in \text{sp}_\varepsilon(A) \), consider a vector \( y \in \mathbb{C}^n \) such that \( \|(A - zI)^{-1}y\|_2 > \varepsilon^{-1}\|y\|_2 \). Then for \( v = (A - zI)^{-1}y/\|(A - zI)^{-1}y\|_2 \) we have \( \|(A - zI)v\|_2 < \varepsilon \), \( \|v\|_2 = 1 \). Define a matrix \( E \in \mathbb{C}^{n \times n} \) through the action of an operator on arbitrary \( x \in \mathbb{C}^n \),

$$
E x := -\langle v, x \rangle (A - zI)v \text{ for } x \in \mathbb{C}^n.
$$

It is easy to check that \( \| E \|_2 < \varepsilon \) and \( (A + E - zI)v = 0 \). Hence, \( z \in \text{sp}_\varepsilon(A) \). \[ \square \]
1.3. Analysis of the minimal residual method

Theorems 1.17 and 1.21 yield the following estimate of convergence of the minimal residual method in terms of pseudospectra.

**Corollary 1.22.** Assume \( \text{sp}_\epsilon(A) \subset \Omega \) for a bounded domain \( \Omega \), with the boundary \( \Gamma \). Then the residuals \( r_i \) of the minimal residual method satisfy the estimates

\[
\frac{\|r_i\|_2}{\|r_0\|_2} \leq \frac{\|\Gamma\|}{2\pi \epsilon} \min_{f \in \mathcal{F}_{\text{cal}}} \max_{\zeta \in \Omega \cup \Gamma} |f(\zeta)|.
\]

Note that \( \| (A-zI)^{-1} \|_2 \) is the largest singular value of \( (A-zI)^{-1} \), which is the inverse of the smallest singular value of \( A-zI \). Thus, the pseudospectra of \( A \) can also be characterized as all \( z \in \mathbb{C} \), where the smallest singular value of \( A-zI \) is less than \( \epsilon \).

The curve \( \Gamma \) in Corollary 1.22 can be taken as the \( \epsilon \) level set of the function \( f(z) := \lambda_{\min}((A-zI)(A-zI)^*) \):

\[
\Gamma = \{ z \in \mathbb{C} : \lambda_{\min}((A-zI)(A-zI)^*) = \epsilon \}.
\]

In practice, computing \( f(z) \) on a set of nodes in \( \mathbb{C} \) and reconstructing numerically its level sets is one way of visualizing and measuring a pseudospectra. We refer the reader to the monograph [194] for a comprehensive discussion of pseudospectra properties and various applications.

1.3.10 • An estimate for normal matrices

A matrix \( A \) is called *normal* if there exists an orthonormal basis of its eigenvectors, and hence

\[
A = Q \Lambda Q^{-1},
\]

where \( Q \) is a unitary matrix, and \( \Lambda \) is the diagonal matrix with eigenvalues of \( A \) standing on the main diagonal. If \( f \) is a polynomial, then due to the unitary invariance of the spectral norm, it holds that

\[
\|f(A)\|_2 = \|Q f(\Lambda) Q^{-1}\|_2 = \|f(\Lambda)\|_2.
\]

Assume all eigenvalues of \( A \) lie in a domain \( \Omega \) with the boundary \( \Gamma \). Then from (1.102) it follows that

\[
\frac{\|r_i\|_2}{\|r_0\|_2} \leq \min_{f \in \mathcal{F}_{\text{cal}}} \max_{\zeta \in \mathbb{C} \setminus \Gamma} |f(\zeta)| = \min_{f \in \mathcal{F}_{\text{cal}}} \max_{\zeta \in \mathbb{C} \setminus \Gamma} |f(\zeta)|. \tag{1.108}
\]

The last equality in (1.108) is valid, since \( f \) is an analytic function.

When \( \Gamma \) is an ellipse on a real part of a complex plane, the right-hand side in (1.108) can be bounded with the help of Chebyshev polynomials. We leave the proof of the following theorem as an exercise.

**Theorem 1.23.** Assume that all eigenvalues of \( A \) belong to a domain \( \Omega \) and that the boundary \( \Gamma \) is an ellipse with the center at point \( p \) on the real axis. The major axis of the ellipse is equal to \( \alpha \), and the distance between two foci is equal to \( 2\beta \). Moreover, assume \( \Omega \notin \Omega \cup \Gamma \). Then for the minimal residual method applied to solve a system with matrix \( A \), the following estimate is valid:

\[
\|r_i\|_2 \leq \frac{P_i(\alpha/\beta)}{P_i'(\alpha/\beta)} \|r_0\|_2,
\]

where \( P_i \) is the Chebyshev polynomial of order \( i \).
1.3.11 Minimal residuals and the Laplace equation

For certain curves $\Gamma$ (ellipse was one example), the asymptotic convergence rate for the minimal residual method can be computed explicitly. In the case of a general smooth (and even piecewise smooth) curve, the rate can be estimated. This can be done in a relatively fast and simple way with the help of the solution to the following external boundary value problem for the Laplace equation:

$$\Delta g(z) = 0, \quad z \in \Omega',
\quad g(z) = 0, \quad z \in \Gamma,
\quad g(z) = \ln |z| + \gamma + o(1), \quad z \to \infty.$$  \hfill (1.109)

Here $g(z)$ is a function of $z = x + i y$, where $x$ and $y$ are real variables; $\Delta \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ is the Laplace operator, $\Omega' = \mathbb{C} \setminus \overline{\Gamma}$, and $\gamma$ is a constant known as the Robin constant, which is found (as part of the solution to the problem) together with $g(z)$.$^5$

It is known that the problem (1.109) has a unique solution. This solution provides the exact expression for the asymptotic rate of the minimal residual method. To simplify the explanation, we assume further in this section that $\Gamma$ is a $C^1$ smooth curve. We shall prove the following theorem.

**Theorem 1.24.** Assume $0 \notin \Omega$ and $g(z)$ is a solution to the problem (1.109). Then

$$q(\Gamma) = e^{-g(0)},$$  \hfill (1.110)

where $q(\Gamma)$ is the asymptotic convergence factor from (1.105).

1.3.12 The method of logarithmic potential

The proof of Theorem 1.24 and a convenient numerical method for solving the problem (1.109) make use of logarithmic potentials. The single layer logarithmic potential is the function

$$v(z) = \int_{\Gamma} \ln |z - \zeta| \sigma(\zeta) |d\zeta|,$$  \hfill (1.111)

where $\sigma(\zeta)$ is a real-valued function, the potential density.

It is easy to check that the logarithmic potential, with an arbitrary density function, satisfies

$$\Delta v(z) = 0 \quad \text{for} \quad z \in \mathcal{F}' \quad \text{and} \quad z \in \Omega.$$

In the theory of logarithmic potentials (see, e.g., [175]), it is proved that the function $v(z)$ is well-defined and continuous for all $z$.

Consider the following system of integral and scalar equations for the potential density $\sigma(\zeta)$ and the scalar quantity $\gamma$:

$$\begin{cases}
\int_{\Gamma} \ln |z - \zeta| |\sigma(\zeta) |d\zeta| = -\gamma & \forall z \in \Gamma, \\
\int_{\Gamma} \sigma(\zeta) |d\zeta| = 1.
\end{cases}$$  \hfill (1.112)

$^5$The quantity $e^{-\gamma}$ is known as the logarithmic capacity of $\Gamma$. 
1.3. Analysis of the minimal residual method

Lemma 1.25. Let \( \sigma(\zeta) \) and \( \gamma \) satisfy the system (1.112). Then the function \( g(z) = v(z) + \gamma \) is the solution to the external boundary value problem (1.109).

**Proof.** The first two relations from (1.109) are satisfied by the construction of \( g(z) \) and the first equality in (1.112). The condition for \( |z| \to \infty \) follows from

\[
v(z) = \ln|z| \int_{\Gamma} \sigma(\zeta) |d\zeta| + \int_{\Gamma} \ln|1 - \zeta/z| \sigma(\zeta) |d\zeta| = \ln|z| + o(1).
\]

To compute \( \sigma(\zeta) \) and \( \gamma \), satisfying (1.112), one can take the following steps: First, find the solution to the integral equation

\[
\int_{\Gamma} \ln|z - \zeta| \phi(\zeta) |d\zeta| = -1.
\]

Further, compute

\[
s = \int_{\Gamma} \phi(\zeta) |d\zeta|,
\]

and finally set

\[
\sigma(\zeta) = \phi(\zeta)/s, \quad \gamma = 1/s.
\]

If a solution to the integral equation above does not exist, then one needs to find a non-trivial solution \( \phi(\zeta) \) of a similar integral equation with zero right-hand side. In this case, \( \gamma = 0 \), and \( \sigma(\zeta) \) is found by scaling \( \phi(\zeta) \) as above.

The next subsection is intended for those readers who are interested in the proof of Theorem 1.24.

1.3.13 Further analysis of the method of logarithmic potential

Lemma 1.26. If \( g(z) \) is the solution to the problem (1.109), then \( g(z) > 0 \) for any point \( z \in \Omega'. \)

**Proof.** Thanks to the maximum principle, \( g(z) \) is nonnegative in \( \Omega' \). Denote by \( M \subset \Omega' \) a set consisting of all points \( z \) such that \( g(z) = 0 \). This set is closed. Assume \( M \) is not empty. If it is also bounded, then it is compact and we can pick a point \( z \in M \) with a maximum modulus \( |z| \). Hence, for any circle \( \Theta \), with its center in \( z \), and belonging to the exterior domain \( \Omega' \), it holds that

\[
\int_{\Theta} g(\zeta) |d\zeta| = 0 \quad \Rightarrow \quad g(\zeta) = 0 \quad \forall \ z \in \Theta.
\]

This is in contradiction to the choice of \( z \) which maximizes \( |z| \). Thus, the set \( M \) cannot be bounded. Therefore, \( g(z) \) does not tend to \( \infty \) for \( z \to \infty \). This contradicts (1.109). We conclude that the set \( M \) is empty.

Lemma 1.27. Assume \( \Gamma \) is \( C^1 \) smooth. Then there exists a continuous potential density \( \sigma(\zeta) \) and some \( \gamma \), satisfying the system (1.112). Furthermore, it holds that \( \sigma(\zeta) \geq 0 \) and \( \int_{\Gamma} \sigma(\zeta) |d\zeta| > 0 \) for any \( \tilde{\Gamma} \subset \Gamma \), with the measure \( |\tilde{\Gamma}| > 0 \).

**Proof of Theorem 1.24.** Here we adapt arguments given in the book [87] for a similar (though formally different) problem. Fix an arbitrary sufficiently small \( \varepsilon > 0 \) and consider the curve \( \Gamma_\varepsilon \) defined by the equation \( g(z) = \varepsilon \). \( \Gamma \) and \( \Gamma_\varepsilon \) are two-level sets of a
continuous function \( g(z) \), and hence \( \Gamma \subset \Omega \) and a distance between \( \Gamma \) and \( \Gamma \) is positive.

We choose \( n \) different points \( \zeta_1, \ldots, \zeta_n \in \Gamma \) and define a polynomial

\[
p_n(z) = \prod_{i=1}^{n} (z - \zeta_i).
\]

Below we show that \( \zeta_1, \ldots, \zeta_n \in \Gamma \) can be taken such that

\[
|\ln | p_n(z) |^{1/n} - v(z) | \leq \varepsilon \quad \forall \ z \in \Gamma,
\]  
(1.113)

where \( v(z) \) is the logarithmic potential defined in subsection 1.3.12, and \( g(z) = v(z) + \gamma \). Thanks to Lemma 1.27, we can divide the curve \( \Gamma \) on \( n \) connected pieces \( \gamma_1, \ldots, \gamma_n \) such that

\[
\int_{\gamma_i} \sigma(\zeta)|d\zeta| = \frac{1}{n}.
\]  
(1.114)

Moreover, there exists a constant \( N = N(\varepsilon) \) such that for any \( n \geq N \) and all \( 1 \leq i \leq n \) the following inequality is valid:

\[
|\ln | z - \zeta | - \ln | z - \eta | | \leq \varepsilon \quad \forall \ \zeta, \eta \in \gamma_i.
\]  
(1.115)

With the help of inequalities (1.114), (1.115) and recalling \( \sigma(\zeta) \geq 0 \) on \( \Gamma \) (Lemma 1.27) we get

\[
\left| \frac{1}{n} \sum_{i=1}^{n} \ln | z - \zeta_i | - v(z) \right| = \left| \sum_{i=1}^{n} \left( \frac{1}{n} \ln | z - \zeta_i | - \int_{\gamma_i} \ln | z - \zeta | \sigma(\zeta) |d\zeta| \right) \right|
\]

\[
= \left| \sum_{i=1}^{n} \int_{\gamma_i} \left( \ln | z - \zeta_i | - \ln | z - \zeta | \right) \sigma(\zeta) |d\zeta| \right| \leq \varepsilon
\]

for any \( z \in \Gamma \) and \( \zeta_i \in \gamma_i \). In the definition of the polynomial \( p_n(z) \) we take any \( \zeta_i \in \gamma_i \). With this choice, the last estimate immediately yields (1.113). Now we recall that \( v(z) = g(z) - \gamma = \varepsilon + \gamma \) for \( z \in \Gamma \). Estimate (1.113) implies

\[
-\gamma \leq \ln | p_n(z) |^{1/n} \leq 2\varepsilon - \gamma \quad \iff \quad e^{-\gamma} \leq | p_n(z) |^{1/n} \leq e^{2\varepsilon - \gamma}.
\]  
(1.116)

One can assume that \( \varepsilon \) is sufficiently small, and so \( 0 \) lies in an external domain with the boundary \( \Gamma \). Then (1.113) yields

\[
|\ln | p_n(0) |^{1/n} - v(0) | \leq \varepsilon \quad \Rightarrow \quad |\ln | p_n(0) |^{1/n} \geq g(0) - \gamma - \varepsilon.
\]

This and (1.116) imply

\[
\max_{z \in \Gamma} \left| \frac{p_n(z)}{p_n(0)} \right|^{1/n} \leq \max_{z \in \Gamma} \left| \frac{p_n(z)}{p_n(0)} \right|^{1/n} \leq \frac{e^{2\varepsilon - \gamma}}{e^{g(0) - \gamma - \varepsilon}}.
\]

We get

\[
\max_{z \in \Gamma} \left| \frac{p_n(z)}{p_n(0)} \right|^{1/n} \leq e^{3\varepsilon} e^{-g(0)}.
\]
Due to an arbitrary choice of \( \epsilon \), we estimate the asymptotic convergence factor

\[
q(\Gamma) = \lim_{n \to \infty} (T_n(\Gamma))^{1/n} \leq e^{-g(0)}.
\]

It remains to prove that the above inequality turns out to be the equality. Denote

\[
m_n(\Gamma) = \max_{z \in \Gamma} |p_n(z)|
\]

for a polynomial \( p_n \), \( \deg(p_n) = n \). Consider a sequence of polynomials \( p_n \) such that

\[
(T_n(\Gamma))^{1/n} = \left( \frac{m_n(\Gamma)}{|p_n(0)|} \right)^{1/n} + o(1) \to q_0.
\]

Without loss of generality, we may assume that the roots of the polynomials \( p_n(z) \) do not lie on \( \Gamma \). Consider a circle of radius \( \delta > 0 \) around any root in \( \Omega' \). For sufficiently small \( \delta \) all the circles belong to \( \Omega' \). Denote by \( \Omega'(\delta) \) the part of \( \Omega' \) without these circles.

For sufficiently small \( \delta \) the function

\[
u(z) = g(z) - \frac{1}{n} \ln \frac{|p_n(z)|}{m_n(\Gamma)}
\]

satisfies the Laplace equation \( \Delta \nu(z) = 0 \) for all \( z \in \Omega'(\delta) \); it is nonnegative on \( \Gamma \) and on all considered circles of radius \( \delta \); moreover, it has a finite limit for \( z \to \infty \). Therefore, we conclude that \( \nu(z) \geq 0 \) for all \( z \in \Omega'(\delta) \). We assume that \( \delta \) is so small that \( 0 \in \Omega'(\delta) \), yielding \( \nu(0) \geq 0 \). This implies

\[
\left( \frac{m_n(\Gamma)}{|p_n(0)|} \right)^{1/n} \geq e^{-g(0)} \Rightarrow q_0 \geq e^{-g(0)} \Rightarrow q(\Gamma) \geq e^{-g(0)}.
\]

We have proved the theorem.

For \( T_n(\Gamma) \) defined in (1.104) we conclude the following result.

**Corollary 1.28.** The sequence \( (T_n(\Gamma))^{1/n} \) is convergent and

\[
q(\Gamma) = \lim_{n \to \infty} (T_n(\Gamma))^{1/n}.
\]

\[
(1.117)
\]

**Exercises**

1.3.1. Suppose that the minimal residual method is applied to solve the system \( Ax = b \) and the matrix \( A \) is such that for a complex number \( \zeta \), with \( |\zeta| = 1 \), the matrix \( \zeta A \) is elliptic. Prove that the estimate (1.95) remains true.

1.3.2. Let all eigenvalues of matrix \( A \) belong to the sector \( \text{Re} \lambda(A) \geq \tau > 0 \), \( |\lambda(A)| \leq \rho \). Prove that the estimate (1.95) is valid if the matrix \( A \) is normal and may fail to be true otherwise (give an example).

1.3.3. For some \( \tau > 0 \) and all vectors \( x \) it holds that \( |\text{Re} \langle Ax, x \rangle| \geq \tau \langle x, x \rangle \) and there exists a vector \( x_0 \) such that \( \text{Re} \langle Ax_0, x_0 \rangle \geq \tau \langle x_0, x_0 \rangle \). Prove the bound \( \text{Re} \langle Ax, x \rangle \geq \tau \langle x, x \rangle \) for all \( x \).
1.3.4. Let $x$ be an exact solution to the system $Ax = b$ with a nonsingular Hermitian matrix $A$. Consider the sequence $x_0, x_1, \ldots$ of approximate solutions produced by the minimal residual method. Prove that $x_{i-1} \neq x$ implies $\|x_i - x\|_2 < \|x_{i-1} - x\|_2$.

1.3.5. Prove that $\text{sp}(A) \subset \Phi(A)$ and that for a normal matrix its numerical range coincides with the convex envelope of the set of matrix eigenvalues.

1.3.6. Let $0 < r < a$ and let $\Gamma$ be a circle of radius $r$ with a center in point $a$ on the real axis. Let $g(z)$ be the solution to the external boundary value problem (1.109). Prove the identity

$$g(0) = \ln \frac{a}{r}. \quad (1.118)$$

1.3.7. The curve $\Gamma$ is an ellipse with the center in point $a$ on the real axis, $r_1$ and $r_2$ are the half axes of the ellipse, and $0 < r_1 < a$. Let $g(z)$ be the solution to the external boundary value problem (1.109). Prove the identity

$$g(0) = \ln \frac{\sqrt{a^2 - r_1^2 + r_2^2} + a}{r_1 + r_2}. \quad (1.119)$$

1.3.8. Prove Theorem 1.23.

### 1.4 • Analysis of the conjugate gradient method

#### 1.4.1 • Convergence of the conjugate gradient method

Consider the decomposition of the residual $r_0 = \sum_{i=1}^n \xi_i z_i$, using the orthonormal system $\{z_i\}$ of the eigenvectors of a matrix $A = A^* > 0$ of size $n$. If only $k$ coefficients $\xi_i$ are nonzero, then the CG method gives the solution to the system $Ax = b$ in at most $k$ steps (in exact arithmetic, of course). If the eigenvalues corresponding to these $k$ eigenvectors are all different, then $\mathcal{K}_{k+1}(A, r_0) = \mathcal{K}_k(A, r_0)$ (show this!) and the solution will be obtained exactly at the step $k$.

It is remarkable that sufficiently accurate approximations to the solution are often achieved in less than $k$ iterations (steps). Similarly to the minimal residual method, the key for studying convergence estimates is the identity $x_i = x_0 + \phi_{i-1}(A) r_0$ with a polynomial $\phi_{i-1}(\lambda)$ of degree $i - 1$. Therefore, for the residual vectors we obtain $r_i = \phi_i(A) r_0$, where $\phi_i(\lambda) = 1 - \lambda \phi_{i-1}(\lambda)$ is the uniquely defined polynomial of degree $i$ subject to condition $\phi_i(0) = 1$. The set of all such polynomials already appeared in the analysis of the minimal residual method and was denoted by $\mathcal{F}_i$.

In step $i$, the CG method minimizes the $A$-norm of the error for all possible vectors $x_0 + y$, where $y \in \mathcal{K}_i$. Since $y \in \mathcal{K}_i$, the residual for arbitrary $x_0 + y$ can be written as $r = Ax_0 - Ay - b = r_0 - Ay = f(A) r_0$, with some $f \in \mathcal{F}_i$. The minimal $A$-norm of the error in step $i$, $e_i = A^{-1} r_i$, is attained for $f(\lambda) = \phi_i(\lambda)$. Therefore, we can estimate

$$\|e_i\|_A^2 = (A e_i, e_i) = (r_i, A^{-1} r_i) = \sum_{j=1}^k \frac{\phi_j^2(\lambda_j)}{\lambda_j} \xi_j^2 \leq \min_{f \in \mathcal{F}_i} \sum_{j=1}^k \frac{\phi_j^2(\lambda_j)}{\lambda_j} \xi_j^2 \leq \min_{f \in \mathcal{F}_i} \left( \max_{0 \leq i \leq k} |f(\lambda_i)| \right)^2 \|e_0\|_A^2.$$ 

If $\lambda(A) \subset [m, M]$, then it clearly holds that

$$\|e_i\|_A \leq \min_{f \in \mathcal{F}_i} \max_{m \leq \lambda \leq M} |f(\lambda)| \|e_0\|_A. \quad (1.120)$$
1.4. Analysis of the conjugate gradient method

1.4.2 The classic and some sharper estimates

In subsection 1.1.7 we found that the optimal \( f \in \mathcal{F}_i \) in (1.120) is the rescaled Chebyshev polynomial on \([m, M]\) and for this optimal \( f \) we have the bound

\[
\max_{m \leq \lambda \leq M} |f(\lambda)| = \frac{1}{|P_i\left(-\frac{M+m}{M-m}\right)|} \leq \left(\frac{1-\sqrt{\frac{m}{M}}}{1+\sqrt{\frac{m}{M}}}\right)^i.
\]

This leads to the classic convergence estimate for the CG method:

\[
\|e_i\|_A \leq 2 \left(\frac{1-\sqrt{\frac{m}{M}}}{1+\sqrt{\frac{m}{M}}}\right)^i \|e_0\|_A. \tag{1.121}
\]

The estimate (1.121) looks similar to the bound (1.26) on the error of the Richardson iteration with the Chebyshev set of parameters (note that the error is measured in different norms). The obvious improvement of CG over (1.20) with Chebyshev parameters is that no explicit knowledge of spectral bounds is required for running the method. Moreover, some sharper convergence estimates can be effective for the CG method, as we discuss below.

The estimate (1.121) suggests that the convergence deteriorates for poorly conditioned matrices. However, in practice, it may remain quite fast even for \( \frac{M}{m} \gg 1 \). Let us explain why this may happen [10].

Let \( \lambda_1 \geq \cdots \geq \lambda_n \) be the eigenvalues of a matrix \( A = A^* > 0 \), and assume that \( \lambda_1 \gg \lambda_2 \).

In this case, the CG method converges as fast as if it applied for a matrix with condition number equal to \( \lambda_2/\lambda_n \).

Indeed, consider the bound (1.120) and take the polynomial \( f \in \mathcal{F}_i \) as

\[
f(\lambda) = \frac{P_{i-1}\left(\frac{2\lambda-\lambda_n-\lambda_2}{\lambda_n-\lambda_2}\right)}{P_{i-1}\left(\frac{-\lambda+\lambda_n+\lambda_2}{\lambda_2-\lambda_n}\right)} \left(1 - \frac{\lambda}{\lambda_1}\right).
\]

This leads to the estimate

\[
\|e_i\|_A \leq 2 \left(\frac{1-\sqrt{\frac{\lambda_n}{\lambda_2}}}{1+\sqrt{\frac{\lambda_n}{\lambda_2}}}\right)^i \|e_0\|_A. \tag{1.122}
\]

If \( \lambda_n \ll \lambda_{n-1} \), then the convergence pattern is similar: the method tends to “ignore” the smallest eigenvalue and behaves asymptotically as if the condition number is equal to \( \lambda_1/\lambda_{n-1} \). However, in this case, an additional factor independent of \( i \) arises in the convergence estimate:

\[
\|e_i\|_A \leq 2 \frac{\lambda_1}{\lambda_n} \left(\frac{1-\sqrt{\frac{\lambda_{n-1}}{\lambda_1}}}{1+\sqrt{\frac{\lambda_{n-1}}{\lambda_1}}}\right)^i \|e_0\|_A. \tag{1.123}
\]

In practice, the appearance of the extra factor corresponds to periods of subasymptotic convergence (so-called stagnation periods) in the convergence history of the CG method.
1.4.3 The Arnoldi and Lanczos methods

To better understand properties of the CG method, it is helpful to consider its close connection to the Lanczos method of computing eigenvalues and eigenvectors of Hermitian matrices. In the non-Hermitian case, the minimal residual method is related to the Arnoldi method of computing eigenvalues and eigenvectors of non-Hermitian matrices.

We start by explaining the Arnoldi method. Let us fix a vector \( r_0 \neq 0 \) and assume that \( r_0, Ar_0, \ldots, A^{i-1}r_0 \) are linear independent, but \( A^i r_0 \) can be written as a linear combinations of \( r_0, Ar_0, \ldots, A^{i-1}r_0 \). This yields \( \mathcal{K}_i = \mathcal{K}_i(r_0, A) = \mathcal{K}_{i+1}(r_0, A) \).

In the Krylov subspace \( \mathcal{K}_i \), consider an orthonormal basis \( q_1, \ldots, q_i \) such that \( \text{span} \{ q_1, \ldots, q_i \} = \mathcal{K}_j \) for all \( 1 \leq j \leq i \). The next vector \( q_{i+1} \) can be constructed from \( Aq_j \) by orthogonalizing it to all already found vectors \( q_1, \ldots, q_i \). Let us set \( Q_j = [q_1 \ldots q_i] \).

The matrix

\[
A_j = Q_j^* A Q_j
\]

is known as the restriction of \( A \) on \( \mathcal{K}_j \). For \( 1 \leq j \leq i \) the matrix \( A_j \) is the leading submatrix of \( A_j \). By definition, a leading submatrix of \( A \) is any square submatrix in the left upper corner of \( A \).

In a general case, the identity (1.69) implies that \( A_j \) is an upper-Hessenberg matrix. If \( A \) is Hermitian, then \( A_j \) is the Hermitian tridiagonal matrix. The process of building the sequences of \( A_j \) and \( Q_j \) is commonly attributed to Arnoldi in a general case and to Lanczos in the Hermitian case.

We recall that the assumption \( \mathcal{K}_i = \mathcal{K}_{i+1} \) implies that \( \mathcal{K}_i \) is an invariant subspace for \( A \). Therefore, \( \text{sp}(A_j) \subset \text{sp}(A) \). Moreover, the eigenvalues of the restricted matrices \( A_j \) for \( j < i \) may happen to be good approximations of the eigenvalues of \( A \). The eigenvalues of upper-Hessenberg or tridiagonal matrices \( A_j \) can be numerically computed by different fast methods (see, e.g., [86]). The method described above for computing eigenvalues is called the Arnoldi method in a general case and the Lanczos method in the Hermitian case.

In both the Arnoldi and Lanczos methods, one builds an orthonormal basis in Krylov subspaces. The same basis can be obtained with the help of well-known matrix decompositions and without any knowledge of Krylov subspaces. Consider the unitary reduction of a matrix \( A \in \mathbb{C}^{n \times n} \) to an upper-Hessenberg matrix \( H \):

\[
A[q_1 \ldots q_n] = [q_1 \ldots q_n] H.
\]

Choose \( q_1 \) arbitrary, but satisfying \( ||q_1||_2 = 1 \), and equalize the first columns on the left- and right-hand sides:

\[
Aq_1 = q_1 b_{11} + q_2 b_{21}.
\]

From \( q_1 \perp q_2 \) it follows that \( b_{11} = \langle Aq_1, q_2 \rangle \). The vector \( q_2 \) is obtained by normalizing the vector \( Aq_1 - q_1 b_{11} \), if it does not equal \( 0 \). After that we equalize the second columns and find \( q_3 \), and so on. The process is halted when we reach the \( i \)th column such that a candidate for \( q_{i+1} \) is found to be zero. In this case,

\[
A[q_1 \ldots q_i] = [q_1 \ldots q_i] H_i,
\]

where \( H_i \) is the leading \( i \times i \) submatrix in \( H \).

Thus, we have a natural way of generating the subspaces

\[
L_j = \text{span} \{ q_1, \ldots, q_j \}.
\]
1.4. Analysis of the conjugate gradient method

This gives us the foundation for developing different projection methods. Formally, we managed to build \( L_j \) without introducing Krylov subspaces. If we recall them, we would, of course, find the equality \( L_j = \mathcal{K}_j(q_1, A) \).

### 1.4.4 Ritz vectors and Ritz values

Assume \( A = A^* > 0 \). Let \( r_0, \ldots, r_{i-1} \) be nonzero residuals from \( i - 1 \) steps of the CG method. Since \( r_j \perp \mathcal{K}_j \) and \( r_j \in r_0 + A\mathcal{K}_j \subset K_{i+1} \), the nonzero residuals form an orthogonal system, and iterations are stopped when the computed residual is zero. The restriction of \( A \) on \( \mathcal{K}_i(r_0, A) \) was defined as

\[
A_i = Q_i^* A Q_i, \quad Q_i = [q_1, \ldots, q_i] = \begin{bmatrix} r_0/\|r_0\|_2, \ldots, r_{i-1}/\|r_{i-1}\|_2 \end{bmatrix}.
\]

(1.124)

Denote by \( \theta_1 \geq \cdots \geq \theta_i \) the eigenvalues of the matrix \( A_i \). These values are also known as Ritz values. If \( A_i v_j = \theta_j v_j, \|v_j\|_2 = 1 \), then the vector \( y_j = Q_i v_j \) is the Ritz vector, corresponding to \( \theta_j \).

We find

\[
0 = Q_i^* (A Q_i v_j - \theta_j Q_i v_j) = Q_i^* (A y_j - \theta_j y_j).
\]

(1.125)

Since the rows of \( Q_i^* \) are the orthonormal basis vectors of \( \mathcal{K}_i \), we obtain from (1.125)

\[
A y_j - \theta_j y_j \perp \mathcal{K}_i, \quad 1 \leq j \leq i.
\]

Recall that the matrix \( A_i \) is tridiagonal. Therefore, its eigenvalues can be easily computed with well-known methods (especially if \( i \ll n \)).

Further in this section we address the following question: Why do Ritz values approximate the eigenvalues of \( A \)? For the sake of presentation, we consider the first Ritz value: Why \( \theta_1 \approx \lambda_1 \)? Let us consider the representation

\[
\theta_1 = \max_{\|x\|_2 = \|Q_i x\|_2} \frac{\langle Q_i^* A Q_i x, x \rangle}{\langle x, x \rangle},
\]

(\( K_i = Q_i \mathbb{C}^n \))

\[
= \max_{\|x\|_2 = \|Q_i x\|_2} \frac{\langle A Q_i x, Q_i x \rangle}{\langle Q_i x, Q_i x \rangle} = \max_{\|y\|_2 = \|y \|_{\mathcal{K}_i}} \frac{\langle A y, y \rangle}{\langle y, y \rangle} = \max_{\phi_{i-1}} \frac{\langle A \phi_{i-1} (A) r_0, \phi_{i-1} (A) r_0 \rangle}{\langle \phi_{i-1} (A) r_0, \phi_{i-1} (A) r_0 \rangle},
\]

where the maximum is sought over all polynomials \( \phi_{i-1} \) of the degree not larger than \( i - 1 \), satisfying \( \phi_{i-1} (A) r_0 \neq 0 \).

The Rayleigh quotient for the maximum eigenvalue of \( A \) and \( \mathcal{K}_i \subset \mathbb{R}^n \) readily gives \( \theta_1 \leq \lambda_1 \). Using the decomposition of the residual \( r_0 = \sum_{j=1}^n \xi_j z_j \) into the linear combination of eigenvectors of the matrix \( A \), we get

\[
\lambda_1 - \theta_1 \leq \lambda_1 \frac{\langle A \phi_{i-1} (A) r_0, \phi_{i-1} (A) r_0 \rangle}{\langle \phi_{i-1} (A) r_0, \phi_{i-1} (A) r_0 \rangle} = \lambda_1 - \frac{\sum_{k=1}^n \lambda_k |\phi_{i-1} (\lambda_k)\|^2 |\xi_k|^2}{\sum_{k=1}^n |\phi_{i-1} (\lambda_k)\|^2 |\xi_k|^2}.
\]

\[
= \frac{\sum_{k=2}^n (\lambda_1 - \lambda_k) |\phi_{i-1} (\lambda_k)\|^2 |\xi_k|^2}{|\phi_{i-1} (\lambda_1)\|^2 |\xi_1|^2 + \sum_{k=2}^n |\phi_{i-1} (\lambda_k)\|^2 |\xi_k|^2} \leq (\lambda_1 - \lambda_n) \frac{\max_{k \in \mathbb{N}} |\phi_{i-1} (\lambda_k)\|^2}{|\phi_{i-1} (\lambda_1)\|^2} \gamma,
\]

where \( \gamma = |\phi_{i-1} (\lambda_1)\|^2 |\xi_1|^2 + \sum_{k=2}^n |\phi_{i-1} (\lambda_k)\|^2 |\xi_k|^2 \).


where
\[
\gamma = \gamma(\tau_0) = \frac{\sum_{k=2}^{n} |\xi_k|^2}{|\xi_1|^2}.
\]

The resulting inequality is valid for any polynomial \(\phi_{i-1}\) of degree not greater than \(i - 1\). Now, the idea is to look for a polynomial such that its value in the point \(\lambda_1\) is much bigger than its value in the points \(\lambda_2, \ldots, \lambda_n\). This property can be attributed to the Chebyshev polynomial for the interval \([\lambda_n, \lambda_2]\). The choice
\[
\lambda_1 - \theta_1 = \frac{(\lambda_1 - \lambda_n)}{P_{i-1}(1 + 2 \mu)} \gamma(\tau_0), \quad \mu = \frac{\lambda_1 - \lambda_2}{\lambda_2 - \lambda_n}.
\]
leads to the following Kaniel–Page estimate (check it!):
\[
\lambda_1 - \theta_1 \leq \frac{(\lambda_1 - \lambda_n)}{P_{i-1}(1 + 2 \mu)} \gamma(\tau_0), \quad \mu = \frac{\lambda_1 - \lambda_2}{\lambda_2 - \lambda_n}.
\]

From the formula (1.24) for \(P_i\), we conclude that \(\theta_1\) converges to \(\lambda_1\) geometrically when \(\lambda_1 \neq \lambda_2\). The convergence is faster for \(\lambda_1\) well separated from \(\lambda_2\). For \(\lambda_1 = \lambda_2\) the estimate (1.126) is useless.

### 1.4.5 Recurrences and polynomials

**Lemma 1.29.** Consider \(A = A^* \in \mathbb{C}^{n \times n}\) and a unitary matrix \(Q = [q_1 \ldots q_n]\) such that

\[
A_n = Q^* A Q = \begin{bmatrix}
\alpha_1 & \beta_1^* \\
\beta_1 & \alpha_2 & \beta_2^* \\
& \ddots & \ddots & \ddots \\
& & \ddots & \ddots & \ddots \\
& & & \beta_{n-2} & \alpha_{n-1} & \beta_{n-1}^* \\
& & & & \beta_{n-1} & \alpha_n
\end{bmatrix}
\]

is Hermitian tridiagonal.

Assume \(\beta_j \neq 0\) for \(1 \leq j \leq i\); then \(q_{i+1} = \pi_i(A) q_i\) for some polynomial \(\pi_i\) of degree \(i\). Moreover, up to a normalizing factor, \(\pi_i\) coincides with the characteristic polynomial of \(A_i\).

\(A_i \in \mathbb{C}^{i \times i}\) is the leading submatrix of the matrix \(A_n\).

**Proof.** It is easy to see that the first \(i + 1\) rows of the matrix equation (1.127) can be written in the form
\[
A[q_1 \ldots q_i] = [q_1 \ldots q_i] A_i + \beta_i q_{i+1} e_i^T, \quad e_i^T = (0, \ldots, 0, 1).
\]

From (1.128) we conclude that there exist polynomials \(\pi_i\) of degree \(i\) such that \(q_{i+1} = \pi_i(A) q_i\) for \(1 \leq i \leq n\). Indeed, for \(i = 1\), \(\pi_1\) explicitly follows from (1.128) and (1.127); for \(i > 1\) one shows the existence of \(\pi_i\) with the help of (1.128) and an induction argument. Also, let us set \(\pi_0(\lambda) = 1\). Substituting \(q_{j+1} = \pi_j(A) q_j\), \(j = 0, \ldots, i - 1\) to (1.128), we obtain
\[
\lambda(\pi_0(\lambda), \ldots, \pi_{i-1}(\lambda)) = (\pi_0(\lambda), \ldots, \pi_{i-1}(\lambda)) A_i + \beta_i \pi_i(\lambda) e_i^T.
\]
This is equivalent to
\[
(\pi_0(\lambda), \ldots, \pi_{i-1}(\lambda))(A_i - \lambda I) = -\beta_i \pi_i(\lambda)e_i^T. \tag{1.129}
\]
We conclude that any root of \(\pi_i\) is the eigenvalue of \(A_i\). Thus, to show that \(\pi_i\) is (up to a factor) the characteristic polynomial of \(A_i\), it is sufficient to prove that all roots of \(\pi_i\) are of degree 1.

We prove that all roots of \(\pi_i\) are of degree 1 by contradiction. Thus, assume that \(\lambda\) is a root of \(\pi_i\) and \(\text{deg}(\lambda) > 1\). Differentiating the equality (1.129) gives
\[
(\pi'_0(\lambda), \ldots, \pi'_{i-1}(\lambda))(A_i - \lambda I) - (\pi_0(\lambda), \ldots, \pi_{i-1}(\lambda)) = -\beta_i \pi'_i(\lambda)e_i^T. \tag{1.130}
\]
Since \(\text{deg}(\lambda) > 1\), it holds that \(\pi'_i(\lambda) = 0\) and (1.129), together with (1.130), yield
\[
(\pi'_0(\lambda), \ldots, \pi'_{i-1}(\lambda))(A_i - \lambda I)^2 = 0.
\]
For a Hermitian matrix, we have \(\ker(A_i - \lambda I)^2 = \ker(A_i - \lambda I)\). Hence, we obtain
\[
(\pi'_0(\lambda), \ldots, \pi'_{i-1}(\lambda))(A_i - \lambda I) = 0. \tag{1.131}
\]
By the assumption of the lemma, \(\beta_j \neq 0\) for \(j = 1, \ldots, i\). Therefore, all eigenspaces of \(A_i\) have dimension one. This and (1.129), (1.131) imply that there exists \(\alpha\) such that
\[
(\pi'_0(\lambda), \ldots, \pi'_{i-1}(\lambda)) = \alpha(\pi_0(\lambda), \ldots, \pi_{i-1}(\lambda)).
\]
From \(\pi'_0(\lambda) = 0, \pi_0(\lambda) = 1\) we conclude that \(\alpha = 0\), and so \((\pi'_0(\lambda), \ldots, \pi'_{i-1}(\lambda)) = 0\). Relation (1.130) and \(\pi'_i(\lambda) = 0\) yield
\[
(\pi_0(\lambda), \ldots, \pi_{i-1}(\lambda)) = 0.
\]
This, however, contradicts \(\pi_0(\lambda) = 1\). \(\square\)

If \(r_i = \phi_i(A)r_0 \neq 0\), then the roots of \(\phi_i(\lambda)\) coincide with the eigenvalues of the matrix \(A_i\), which is the restriction of matrix \(A\) on \(\mathcal{K}_i(r_0, A)\).

1.4.6 • “Superlinear convergence” and “disappearing” eigenvalues

Strictly speaking, the notion of superlinear convergence does not directly apply to a process with a finite number of iterations. Nevertheless, the CG method possesses certain similar properties: The quantity \(\omega_i \equiv ||e_i||_A/||e_{i-1}||_A\) typically (not monotonically) decreases for larger \(i\), while for a method with linear convergence, one has \(\omega_i \approx \text{const}\).

It turns out that in exact arithmetic, the CG method at some stage starts converging as one would expect if the extreme eigenvalues \(\lambda_1\) or \(\lambda_n\) (or both) are not present in the matrix spectrum; later, the method converges as if \(\lambda_2\) or \(\lambda_{n-1}\) disappears from the spectrum as well, and so on [207].

An eigenvalue does not affect the convergence once it is well approximated by a Ritz value, i.e., by an eigenvalue of the restricted matrix constructed (only virtually!) by the Lanczos method started from the initial vector \(q_1 = r_0/||r_0||_2\).

Let \(r_i\) be a residual at iteration \(i\) of the CG method. Consider its decomposition
\[
r_i = \sum_{j=1}^{n} \xi_j z_j.
\]
in the orthonormal basis of the eigenvectors of matrix $A$. We continue the iterations, and in parallel we virtually define another iteration process starting with the residual
\[
\tilde{r}_0 = \sum_{j=2}^{n} \tilde{\xi}_j z_j = r_j - \tilde{\xi}_1 z_1.
\]

Therefore, along with the iterates $x_{i+j}$, we obtain another sequence of iterates $\tilde{x}_j$ from the second iterative process. Let us denote two error vectors
\[
e_{i+j} = x_{i+j} - x, \quad \tilde{e}_j = \tilde{x}_j - x, \quad j = 0, 1, \ldots,
\]
where $x$ is the solution to the system $Ax = b$.

**Theorem 1.30 (van der Sluis–van der Vorst).** Assume $\lambda_1 \geq \cdots \geq \lambda_n$ are the eigenvalues of the matrix $A = A^* > 0$ and $\theta_1$ is the leading eigenvalue of the restriction $A_1$. Then for all $j = 0, 1, \ldots$ the following inequality holds:
\[
\|e_{i+j}\|_A \leq c_i \|\tilde{e}_j\|_A,
\]
where
\[
c_i = \max_{2 \leq k \leq n} \left| \frac{(\lambda_1 - \lambda_k) \theta_1}{(\theta_1 - \lambda_k) \lambda_1} \right|.
\]

**Proof.** Let $r_0 = \sum_{k=1}^{n} \xi_k z_k$. We have
\[
r_i = \sum_{k=1}^{n} \psi_i(\lambda_k) \xi_k z_k, \quad r_{i+j} = \sum_{k=1}^{n} \psi_{i+j}(\lambda_k) \xi_k z_k,
\]
\[
\tilde{r}_0 = \sum_{k=2}^{n} \psi_i(\lambda_k) \xi_k z_k, \quad \tilde{r}_j = \sum_{k=2}^{n} \psi_i(\lambda_k) \tilde{\psi}_j(\lambda_k) \xi_k z_k,
\]
where $\tilde{\psi}_j(\lambda)$ is the polynomial of degree at most $j$.

According to Lemma 1.29, $\theta_1$ is the root of the polynomial $\psi_i(\lambda)$, and hence the expression
\[
\Psi_i(\lambda) = \frac{1 - \frac{\lambda}{\lambda_1}}{1 - \frac{\lambda}{\theta_1}} \psi_i(\lambda)
\]
is the polynomial of degree at most $i$, satisfying conditions $\Psi_i(0) = 1$ and $\Psi_i(\lambda_1) = 0$. Since the CG method minimizes the $A$-norm of the error, we have
\[
\|e_{i+j}\|_A^2 = \sum_{k=1}^{n} \frac{|\psi_{i+j}(\lambda_k)|^2}{\lambda_k} |\xi_k|^2 \leq \sum_{k=2}^{n} \frac{1}{\lambda_k} |\Psi_i(\lambda_k)|^2 |\tilde{\psi}_j(\lambda_k)|^2 |\xi_k|^2
\]
\[
\leq \max_{2 \leq k \leq n} \left| \frac{1 - \frac{\lambda}{\lambda_k}}{1 - \frac{\lambda}{\theta_1}} \right| \sum_{k=2}^{n} \frac{|\psi_i(\lambda_k)|^2}{|\lambda_k|} |\tilde{\psi}_j(\lambda_k)|^2 |\xi_k|^2 = c_i^2 \|\tilde{e}_j\|_A^2.
\]

We see that $c_i$ tends to 1 as soon as $\theta_1$ approaches $\lambda_1$. If $c_i$ is close to 1, then the iterative process behaves as if the initial residual spans over all eigenvectors except the one corresponding to $\lambda_1$. 

1.4.7 Inexact iterations

Consider an iterative method for solving the system of algebraic equations

\[ Ax = b, \quad (1.134) \]

with a Hermitian \( n \times n \) matrix \( A \). It may happen that the matrix-vector product with \( A \) is expensive to compute exactly. For example, when multiplication with \( A \) involves a solution of an auxiliary linear system, as is the case with Schur complement equations (see subsection 5.2.2), or when the matrix \( A \) is very large or dense. Instead, a “cheaper” approximation of \( A \) may be available, and the product with matrix \( A \) is computed inexactly. Another obvious source of inexactness is roundoff errors. Thus, instead of exact matrix-vector multiplication \( Av \), the product

\[ (A + E)v \quad (1.135) \]

is computed, where \( E \) is a perturbation matrix that may be different every time the matrix-vector product is evaluated. Krylov subspace methods with approximate evaluation of matrix-vector products are often called inexact Krylov subspace methods. It is natural to ask how the properties of iterative methods depend on such perturbations. If the original (exact) method applied to solve (1.134) converges and all perturbation matrices satisfy \( \|E\| \leq \epsilon \), one may expect that after a certain number of steps the iterates \( x_k \) of its inexact counterpart would be in an \( \epsilon \)-neighborhood of \( x \). In practice, it turns out that inexact iterative methods show even stronger convergence properties: The accuracy of matrix-vector multiplication with \( A \) can be relaxed in the progress of iterations without apparent convergence degradation. We shall make this statement more precise further in this section. Note that such a situation may be counterintuitive, since other inexact iterative strategies, such as the inexact Newton method, require the inner tolerance to be tightened as the iterates converge to the solution [66]. To understand this convergence phenomena of inexact Krylov subspace methods, we shall consider the inexact Richardson and CG iterations.

The convergence of inexact Krylov methods, with the perturbation norm \( \|E\| \) growing when iterations progress, was observed empirically in [29, 30]. A more rigorous explanation of such convergence behavior was given later in [206, 182].

1.4.8 An inexact Richardson method

Following [206], we start with the Richardson iteration, which is straightforward to analyze. Thus, assume \( A \in \mathbb{R}^{n \times n} \), \( A = A^T > 0 \), and consider the simple iteration (1.20) with the optimal relaxation parameter \( \alpha = 2(\lambda_{\text{max}}(A) + \lambda_{\text{min}}(A))^{-1} \). It is convenient to express the iterations in the form of recurrent relations for iterates \( x_i \) and residuals \( r_i = b - Ax_i \):

\[ r_{i+1} = r_i - \alpha x_i r_i, \quad x_{i+1} = x_i + \alpha r_i, \quad i = 0, 1, \ldots, \quad (1.136) \]

with \( r_0 = b \) and \( x_0 = 0 \). In the inexact Richardson method, however, instead of (1.136) one computes

\[ \tilde{r}_{i+1} = \tilde{r}_i - \alpha_0 (A + E_i) \tilde{r}_i, \quad \tilde{x}_{i+1} = \tilde{x}_i + \alpha_0 \tilde{r}_i, \quad i = 0, 1, \ldots, \quad (1.137) \]

with \( \tilde{r}_0 = b \) and \( \tilde{x}_0 = 0 \). Here \( E_i \) are some perturbation matrices. Two immediate questions are: (i) What can be said about the convergence of \( \tilde{x}_i \)? and (ii) How can one judge the accuracy of \( \tilde{x}_i \) in terms of recursive residuals \( \tilde{r}_i \), which are now differ from the real residuals \( b - Ax_i \)? Answers to both questions are easy to give when inspecting recurrent
relations for the differences \( \tilde{d}_i = \tilde{r}_i - (b - A\tilde{x}_i) \) and \( \delta d_i = \tilde{r}_i - r_i \). Indeed, one checks the equalities
\[
\begin{align*}
\tilde{d}_{i+1} &= \tilde{d}_i - \alpha_i E_i \tilde{r}_i, \\
\delta d_{i+1} &= \delta d_i - \alpha_i (AD_i + E) \tilde{r}_i.
\end{align*}
\] (1.138)

(1.139)

Taking the norm of both sides of (1.138), applying the triangle inequality, and summing up all inequalities for \( i = 0, \ldots, m - 1 \) lead to the bound \( (d_0 = 0) \)
\[
\|d_m\|_2 \leq \alpha_0 \sum_{i=0}^{m-1} \|E_i\|_2 \|\tilde{r}_i\|_2 \leq \alpha_0 m \max_i \|E_i\|_2 \|\tilde{r}_i\|_2.
\] (1.140)

For (1.139) similar arguments give
\[
\|\delta d_m\|_2 \leq \alpha_0 \sum_{i=0}^{m-1} q_{m-n-1} \|E_i\|_2 \|\tilde{r}_i\|_2 \leq \frac{\alpha_0}{1 - q} \max_i \|E_i\|_2 \|\tilde{r}_i\|_2
\] (1.141)

with \( q = \|I - \alpha_0 A\|_2 = \frac{\lambda_{\max}(A) - \lambda_{\min}(A)}{\lambda_{\max}(A) + \lambda_{\min}(A)} \).

Estimates (1.140) and (1.141) show that the norms of the defects \( d_i \) and \( \delta d_i \) are controlled by the maximum of \( \|E_i\|_2 \|\tilde{r}_i\|_2 \) over all iterations. Now, assume that the “inexactness” in the matrix-vector product is bounded as
\[
\|E_i\|_2 \leq \epsilon \|A\|_2 \|\tilde{r}_i\|_2^{-1} \quad \forall \ i = 0, 1, \ldots,
\] (1.142)

with some \( \epsilon > 0 \). Then (1.141) implies that the recursive residuals \( \tilde{r}_i \) satisfy
\[
\|\tilde{r}_i - r_i\|_2 \leq \epsilon \ \text{cond}(A).
\] (1.143)

Since we know that residuals of exact method tend to zero, \( \|r_i\|_2 \to 0 \), inequality (1.143) implies that \( \tilde{r}_i \) converge to an \( \epsilon \)-neighborhood of 0. Also (1.142) and (1.140) yield
\[
\|\tilde{r}_i - (b - A\tilde{x}_i)\|_2 \leq 2\epsilon i.
\] (1.144)

Inequalities (1.141) and (1.143) answer questions (i) and (ii) about the convergence of inexact iterations and accuracy of recursive residuals \( \tilde{r}_i \). Indeed, combining (1.141) and (1.143), one may show the estimate on the error of the inexact Richardson method:
\[
\|x - \tilde{x}_i\|_2 \leq \|A^{-1}\|_2 \|b - A\tilde{x}_i\|_2 \leq \|A^{-1}\|_2 (\|\tilde{r}_i - (b - A\tilde{x}_i)\|_2 + \|\tilde{r}_i - r_i\|_2 + \|r_i\|_2)
\leq \|A^{-1}\|_2 (\epsilon (2i + \text{cond}(A)) + q\|\tilde{b}\|_2).
\]

This estimate can be slightly improved by exploiting the recursive relations for iterates in (1.136) and (1.137). Subtracting this equalities and applying the triangle inequality yield
\[
\|x_i - \tilde{x}_i\|_2 \leq \alpha_0 \sum_{i=1}^{i} \|\tilde{d}_m\|_2 \leq \alpha_0 \epsilon \ \text{cond}(A)i.
\]

Therefore, we get
\[
\|x - \tilde{x}_i\|_2 \leq \alpha_0 \epsilon \ \text{cond}(A)i + \|x_i - \tilde{x}_i\|_2 \leq \alpha_0 \epsilon \ \text{cond}(A)i + q\|x\|_2
\leq 2\epsilon i \|A^{-1}\|_2 + q\|x\|_2
\leq (2\epsilon + q\|\tilde{b}\|_2)\|A^{-1}\|_2.
\] (1.145)
Error bound (1.145) shows that the iterates of inexact method tend to an \( \epsilon \)-neighborhood of the solution \( x \). The factor in (1.145) multiplying \( \epsilon \) grows linearly with respect to the iteration number. Numerical experiments [206] show that in practice the expected dependence on \( i \) is more favorable.

Consider the bound (1.142), which controls the accuracy of matrix-vector multiplication: due to (1.143) the factor \( \|\bar{r}_i\|_2^{-1} \) increases to the level of \( \epsilon^{-1} \) for sufficiently large \( i \). Therefore, when the iterations progress, the norm of the perturbation matrix \( E \) may increase and become of the same order as \( \|A\|_2 \) (the matrix-vector multiplication may be rather inaccurate).

### 1.4.9 An inexact conjugate gradient method

Now we discuss which of the conclusions about the convergence of inexact methods remain valid for the inexact CG method. In the CG method, iterates and residuals non-linearly depend on perturbations introduced in each iteration if matrix-vector products are evaluated approximately. This makes the analysis of the inexact CG method rather involved, and only partial convergence results are available at this moment. Consider the recurrent form of the CG method given in Algorithm 1.2. Accordingly, the inexact version reads

\[
\begin{align*}
q &= (A + E_i)\bar{p}_i, \\
\alpha_i &= (\bar{r}_{i-1}, \bar{r}_{i-1})/(q_i, p_i), \\
\bar{x}_i &= \bar{x}_{i-1} + \alpha_i p_i, \\
\bar{r}_i &= \bar{r}_{i-1} - \alpha_i q_i, \\
\beta_i &= (\bar{r}_i, \bar{r}_i)/(\bar{r}_{i-1}, \bar{r}_{i-1}), \\
\bar{p}_{i+1} &= \bar{r}_i + \beta_i p_i,
\end{align*}
\]  

(1.146)

where \( \bar{x}_0 = 0, \bar{p}_0 = \bar{r}_0 = b \).

Similarly to the analysis of the Richardson method, let us estimate the defect \( d_i = \bar{r}_i - (b - A\bar{x}_i) \) between computed and “exact” residuals. The relation for \( \bar{x}_i \) in (1.146) yields

\[
b - A\bar{x}_i = (b - A\bar{x}_{i-1}) - \alpha_i A\bar{p}_i.
\]

We subtract this from

\[
\bar{r}_i = \bar{r}_{i-1} - \alpha_i (A + E_i)\bar{p}_i
\]

to get

\[
d_i = d_{i-1} - \alpha_i E_i\bar{p}_i.
\]

This identity immediately implies the estimate

\[
\|d_m\|_2 \leq \sum_{i=0}^{m-1} |\alpha_i| \|E_i\|_2 \|\bar{p}_i\|_2 \leq m \max_i |\alpha_i| \|E_i\|_2 \|\bar{p}_i\|_2. 
\]

(1.147)

If the matrix-vector product accuracy is driven by \( |\alpha_i| \|\bar{p}_i\|_2 \), that is,

\[
\|E_i\|_2 \leq \varepsilon \|A\|_2 (|\alpha_i| \|\bar{p}_i\|_2)^{-1},
\]

then a bound similar to (1.144) immediately follows. However, \( \alpha_i \) is not available at the stage of the method when the matrix-vector product is computed, and it is not a priori clear if \( |\alpha_i| \|\bar{p}_i\|_2 \) decreases in the course of iterations, leading to a relaxation strategy. Comparing (1.147) to (1.140), one may also look for an estimate on coefficients \( \alpha_i \), but a
reasonable estimate of $|x_j|$ for the inexact method is not easy to access. Instead, in [206] the following estimate is proved on the product $|x_j||\tilde{p}_i|_2$:

\[ |x_j||\tilde{p}_i|_2 \leq \sqrt{7}||H^H_i||_2(||\tilde{r}_i||_2^2 + ||\tilde{r}_{i+1}||_2^2)^{\frac{1}{2}}, \]  

(1.148)

where $H^H_i$ is the pseudo-inverse of the matrix $H_i$ from the inexact Lanczos decomposition associated with (1.146) and such that $||H^H_i||_2 \leq ||H^H_j||_2$ for $i \leq j$. It may be argued that it is reasonable to assume $||H^H_i||_2 \leq ||A^{-1}||_2$. Estimates (1.147) and (1.148) yield

\[ ||\tilde{r}_m - (b - A\tilde{x}_m)||_2 \leq m^\frac{1}{2}||H^H_m||_2 \max_i ||E_i||_2 ||\tilde{r}_i||_2 + ||\tilde{r}_{i+1}||_2^{\frac{1}{2}}. \]

If one assumes that the norms of perturbation matrices are bounded as

\[ ||E_i||_2 \leq \varepsilon ||A||_2(1 + ||\tilde{r}_i||_2 + ||\tilde{r}_{i+1}||_2^{\frac{1}{2}})^{-\frac{1}{2}} \quad \text{and} \quad ||H^H_i||_2 \leq \varepsilon ||A^{-1}||_2, \]  

(1.149)

then

\[ ||\tilde{r}_m - (b - A\tilde{x}_m)||_2 \leq \varepsilon \varepsilon \text{cond}(A)m^\frac{1}{2}. \]  

(1.150)

Thus, with the assumptions (1.149), the convergence to zero of recursive residuals implies that the exact residuals tend to an $\varepsilon$-neighborhood of 0. Note that the bound on $||E_i||_2$ allows the accuracy of the matrix-vector multiplication to decrease if the iterates tend to the solution. The vector $\tilde{r}_{i+1}$ is not available on the iteration $i$ of the method. To make the accuracy bound computable, one should avoid using it.

It remains an open question whether the inexact CG method with the inner accuracy tolerance as in (1.149) (or with similar strategies relaxing the tolerance in the progress of iterations) can be proved to converge to some neighborhood of the solution $x$. In other words, we are missing a proof that $||\tilde{r}_i||$ or $||b - A\tilde{x}_i||$ would stagnate at some level of order $O(\varepsilon)$. Numerical experience suggests that in many cases inexact Krylov subspace methods with relaxed inner accuracy are successful and the $\varepsilon$-accuracy is attained by the iterates, although the convergence may be nonmonotone.

### 1.4.10 Preconditioning

If an iteration method for the system $Ax = b$ is not converging sufficiently fast, then one may try running the method for an equivalent system of the form $ABy = b$. Such a substitution is called preconditioning and the matrix $B$ is called a preconditioner. The “best” choice is $B^{-1} = A$ (in particular, this indicates that convergence can always be improved with preconditioning). However, such an “obvious” choice is not practical.

The matrix $AB$ is not necessarily computed explicitly. Instead, one multiplies a vector with $AB$ in two subsequent steps: First, multiply the solution with the matrix $B$, and second, multiply the solution with $A$. Such preconditioning is sometimes called explicit. Instead of a matrix $B$ approximating $A^{-1}$, one is often given a matrix $C = B^{-1}$ that approximates $A$. In this case, to compute a matrix-vector product with $B$ one needs to solve a system of linear algebraic equations with matrix $C$. Such preconditioning is sometimes called implicit.

As an alternative to $ABy = b$, one may consider the equivalent system $BAY = Bb$, with all of the above remaining valid. To distinguish between this two ways of preconditioning, $ABy = b$ is called the right preconditioning and $BAY = Bb$ the left preconditioning.
1.4. Analysis of the conjugate gradient method

1.4.11 Preconditioning, the Hermitian case

If the matrices $A$ and $B$ are Hermitian, the product matrix $AB$ is not necessarily Hermitian. However, it is still Hermitian in a generalized sense.

Matrices can be considered as operators in a space with a special scalar product. As an example, we can set

$$\langle x, y \rangle_D = \langle Dx, y \rangle$$

for a matrix $D = D^* > 0$.

A matrix $M$ is called $D$-Hermitian if

$$\langle Mx, y \rangle_D = \langle x, My \rangle_D \quad \forall x, y \in \mathbb{C}^n$$

and positive $D$-definite if $\langle Mx, x \rangle_D > 0$ for all $x \neq 0$.

Assume $A$ and $B$ are Hermitian positive definite matrices. It is easy to check that the matrix $AB$ is $B$-Hermitian and positive $B$-definite. Hence, the system $ABx = b$ can be solved with a standard CG method substituting a usual Euclidean inner product $\langle \cdot, \cdot \rangle$ with $\langle \cdot, \cdot \rangle_B$:

$$\hat{x}_i = \text{argmin} \{ f(\hat{x}) : \hat{x} = x_0 + y, y \in \mathcal{K}(AB, \hat{r}_0) \},$$

with $f(\hat{x}) = \frac{1}{2} \langle AB \hat{x}, \hat{x} \rangle_B - \text{Re}(b, \hat{x})_B,$

(1.151)

This gives the recursive relations for the CG method applied to solve $AB\hat{x} = b$:

$$\begin{align*}
\hat{r}_0 &= b - AB\hat{x}_0, \quad \hat{p}_1 = \hat{r}_0, \\
\alpha_i &= \langle Br_{i-1}, \hat{r}_{i-1} \rangle / \langle BAB\hat{p}_i, \hat{p}_i \rangle, \\
\hat{x}_i &= \hat{x}_{i-1} + \alpha_i \hat{p}_i, \\
\hat{r}_i &= \hat{r}_{i-1} - \alpha_i AB\hat{p}_i, \\
\beta_i &= \langle \hat{r}_i, \hat{r}_i \rangle / \langle \hat{r}_{i-1}, \hat{r}_{i-1} \rangle, \\
p_{i+1} &= \hat{r}_i + \beta_i \hat{p}_i.
\end{align*}$$

It is convenient to make the following substitution:

$$\begin{align*}
\bar{x}_i &= B^{-1}x_i, \quad \bar{r}_i = r_i, \quad \bar{p}_i = Bp_i.
\end{align*}$$

As a result, we arrive at the preconditioned conjugate gradient (PCG) method:

**Algorithm 1.4. PCG method.**

Input: Matrix $A \in \mathbb{C}^{n \times n}$, $A = A^* > 0$, and preconditioner matrix $B \in \mathbb{C}^{n \times n}$, $B = B^* > 0$, right-hand side vector $b \in \mathbb{C}^n$, initial approximation $x_0 \in \mathbb{C}^n$, maximum number of iterations $N_{\text{max}}$, stopping criterion.

Set $r_0 := b - Ax_0$, $p_1 = Br_0$ (Initialization);

For $i = 1, \ldots, N_{\text{max}}$ or until stopping criterion is met, do

1. $\alpha_i := \langle r_{i-1}, Br_{i-1} \rangle / \langle Ap_i, p_i \rangle,$
2. $x_i := x_{i-1} + \alpha_i p_i,$
3. $r_i := r_{i-1} - \alpha_i Ap_i,$
4. $\beta_i := \langle r_i, Br_i \rangle / \langle r_{i-1}, Br_{i-1} \rangle,$
5. $p_{i+1} := Br_i + \beta_i p_i.$

End

Output: New approximation $x_i$ to the solution of $Ax = b$ after iterations are stopped.
We note that $\mathbf{x}_i$ is an approximate solution and the vector $r_i = b - A\mathbf{x}_i$ is its residual for the original system.

If $B$ and $A$ are Hermitian positive definite matrices, then all eigenvalues of $AB$ are real and positive (explain why!). One typically looks for a preconditioner $B$ such that $\text{sp}(AB) \in \{m, M\}$ and $\text{cond}(AB) = M/m$ is possibly small. Based on the minimization property (1.151) one shows that the estimate (1.121) remains valid with the spectra bounds for the preconditioned matrix, $\text{sp}(AB) \in \{m, M\}$ (we leave checking this as an exercise). Even if $\text{cond}(AB)$ is fairly large, the results about the superconvergence of the CG method suggest that the convergence should be fast if the majority of eigenvalues (not all, but almost all eigenvalues) are well clustered.

### 1.4.12 Number of iterations and alternative convergence estimates

According to the estimate (1.121), the inequality $\|e_i\|_A \leq \varepsilon \|e_0\|_A$ is valid if

$$2 \left( \frac{1 - x^{-1/2}}{1 + x^{-1/2}} \right)^i \leq \varepsilon \iff i \geq \ln(2\varepsilon^{-1})/\ln \frac{1 + x^{-1/2}}{1 - x^{-1/2}}.$$  

where $x = M/m$ is the spectral condition number for the matrix $A = A^* > 0$. Note that

$$\ln \frac{1 + x^{-1/2}}{1 - x^{-1/2}} = \ln \left( 1 + \frac{2x^{-1/2}}{1 - x^{-1/2}} \right) \leq \frac{2x^{-1/2}}{1 - x^{-1/2}}.$$

Therefore,

$$i \geq \frac{1}{2} \ln(2\varepsilon^{-1})x^{1/2} \Rightarrow \|e_i\|_A \leq \varepsilon \|e_0\|_A.$$

This bound on the number of iterations is exact on the class of all possible Hermitian positive definite matrices with a given condition number, but for many particular matrices from this class the estimate can be too pessimistic. Therefore, when building a preconditioner, it is not always necessary to minimize the condition number of $AB$. It can be also useful to look at other quantities. One example is the following $K$-condition number:

$$K(A) = \frac{(\text{tr}A/n)^n}{\text{det} A}. \quad (1.153)$$

**Theorem 1.31** (see [110]). The residuals of the PCG method (1.4.11) satisfy the estimates

$$\|r_i\|_B \leq \left( K(AB)^{1/i} - 1 \right)^{1/2} \|r_0\|_B.$$

**Corollary 1.32.** If the iteration number satisfies $i \geq \log_2 K(AB) + \log_2 \varepsilon^{-1}$, then $\|r_i\|_B \leq \varepsilon \|r_0\|_B$.

**Proof.** The simple inequality $t - 1 \leq (t/2)^2$ for all $t \in \mathbb{R}$ yields

$$(K(AB)^{1/i} - 1)^{1/2} \leq K(AB)/2^i.$$  

Hence, $\|r_i\|_B \leq \varepsilon \|r_0\|_B$ is satisfied if $K(AB)/2^i \leq \varepsilon$. Applying $\log_2$ to both sides of the estimate and performing elementary computations yield the result. \qed
1.4. Analysis of the conjugate gradient method

Proof of the theorem. For the sake of simplicity, consider the case $B = I$. Assume that the residuals $r_0, \ldots, r_i$ are nonzero. From step 2 of Algorithm 1.2 for the conjugate gradient method we get

$$Ap_i = \frac{1}{\alpha_i}(r_{i-1} - r_i), \quad Ap_{i+1} = \frac{1}{\alpha_{i+1}}(r_i - r_{i+1}), \quad Ap_{i+1} = Ar_i + \beta_i Ap_i.$$  

This implies

$$AR_{i+1} = R_{i+1}A^T - \frac{1}{\alpha_{i+1}}r_{i+1}(0, 0, 1)^T, \quad R_{i+1} = [r_0, r_1, \ldots, r_i],$$

(1.154)

where

$$T_{i+1} = \begin{bmatrix}
\frac{1}{s_1} & -\frac{\beta_1}{s_1} & \ldots & -\frac{\beta_i}{s_i} \\
-\frac{s_1}{s_2} & \frac{1}{s_2} + \frac{\beta_1}{s_1} & \ldots & -\frac{\beta_i}{s_i} \\
\vdots & \ddots & \ddots & \vdots \\
-\frac{s_1}{s_i} & \ldots & -\frac{s_{i-1}}{s_i} & \frac{1}{s_i} + \frac{\beta_i}{s_i}
\end{bmatrix}, \quad i \geq 1.$$

The leading submatrices $\Delta_j$ of $T_m$ satisfy

$$\det \Delta_j = \left(\frac{1}{s_j} + \frac{\beta_{j-1}}{s_{j-1}}\right) \det \Delta_{j-1} = \frac{\beta_{j-1}}{s_{j-1}} \det \Delta_{j-2}, \quad \det \Delta_1 = \frac{1}{s_1}, \quad \det \Delta_2 = \frac{1}{s_1s_2}.$$

Using this recursion, one finds

$$\det T_{i+1} = \prod_{j=1}^{i+1} \frac{1}{s_j}.$$

The coefficients $\beta_j$ are computed from $\beta_j = \langle r_j, r_j \rangle / \langle r_{j-1}, r_{j-1} \rangle$; cf. step 4 of Algorithm 1.2. This yields

$$\prod_{j=1}^{i} \beta_j = \frac{\|r_i\|^2}{\|r_0\|^2},$$

Denote by $f(a_1, \ldots, a_{i+1})$ the ratio between the arithmetic and geometric means of $a_1, \ldots, a_{i+1}$ ($f(\ldots) \geq 1$). Fix an arbitrary $0 < \theta < 1$ and denote $B = (K(T_{i+1}))^{1/(i+1)}$. Elementary computations show

$$B = \left(\sum_{j=1}^{i+1} \frac{1}{\alpha_j} + \sum_{j=1}^{i+1} \frac{\beta_j}{\alpha_j} \right) \left(\prod_{j=1}^{i+1} \frac{1}{\alpha_j}\right)^{-1/(i+1)}$$

$$= (1-\theta)^{1/(i+1)} \sum_{j=1}^{i+1} \frac{1}{s_j} - \frac{\theta}{s_{i+1}} + \theta^{1/(i+1)} \sum_{j=1}^{i+1} \frac{\beta_j}{s_j} + \frac{\theta^{1/(i+1)}}{(1-\theta)^{1/(i+1)}} \left(\prod_{j=1}^{i+1} \beta_j\right)^{1/(i+1)}$$

$$= (1-\theta)^{1/(i+1)} f\left(\frac{1}{\alpha_1}, \ldots, \frac{1}{\alpha_i}, \frac{1}{\alpha_{i+1}}\right)$$

$$+ \theta^{1/(i+1)} f\left(\frac{\beta_1}{\alpha_1}, \ldots, \frac{\beta_i}{\alpha_i}, \frac{\theta}{\alpha_{i+1}} \right) \left(\|r_i\|^2\right)^{1/(i+1)}.$$
This immediately implies
\[ B \geq (1 - \theta)^{1/(i+1)} + \theta^{1/(i+1)} \left( \frac{||r_i||_2^2}{||r_0||_2^2} \right)^{1/(i+1)}, \]
and thus
\[ \frac{||r_i||_2^2}{||r_0||_2^2} \leq \left( \frac{B - (1 - \theta)^{1/(i+1)}}{\theta^{1/(i+1)}} \right)^{i+1}. \]

Now we look for a minimum of the right-hand side with respect to \( \theta \) and find
\[ ||r_m||_2 \leq \left( K(T_{i+1})^{1/i} - 1 \right)^{1/2} ||r_0||_2. \]

It remains to show that \( K(T_{i+1}) \leq K(A) \). We note that \( T_{i+1} \) is the leading submatrix of the order \( i + 1 \) in the matrix
\[ \tilde{A} = \begin{bmatrix} T_{i+1} & S \\ S^* & H \end{bmatrix}, \]
which is unitary similarly to the original matrix \( A \). This equality follows, for example, from (1.154), with \( i + 1 = n \), if one recalls that \( r_i \) form an orthogonal system and \( \beta_i = ||r_i||^2/||r_{i+1}||^2 \). The quantity \( K(A) \) is invariant under a unitary transformation (explain why!). For the block-diagonal matrix
\[ D = \begin{bmatrix} T_{i+1} & 0 \\ 0 & H \end{bmatrix} \]
we get \( \text{tr}A = \text{tr}D \) and \( \det D^{-1}A \leq 1 \) (the proof of the last estimate we leave as Exercise 1.4.4). Hence, \( K(A) \geq K(D) \). Denote by \( \mu_1, \ldots, \mu_{i+1} \) and \( \mu_{i+2}, \ldots, \mu_n \) eigenvalues of the matrix blocks \( T \) and \( H \). Let \( s = (\mu_1 + \cdots + \mu_{i+1})/(i+1) \). We find
\[ K(D) = (f(\mu_1, \ldots, \mu_n))^n = (f(s, \ldots, s, \mu_{i+2}, \ldots, \mu_n))^n (f(\mu_1, \ldots, \mu_{i+1}))^{i+1} \geq (f(\mu_1, \ldots, \mu_{i+1}))^{i+1} = K(T_{i+1}). \]

Theorem 1.31 is proved.

1.4.13 A linear iteration as a preconditioner

To solve the system \( Az = b \), assume a linear iteration of the form
\[ z^{i+1} = Mz^i + Nb, \quad (1.155) \]
where \( M \) is the iteration matrix. A few examples of simple linear iterations, for example Jacobi and SOR methods, were introduced in Section 1.1. More sophisticated linear iterations, such as multigrid cycles, will be studied in Chapter 3. Here we discuss how and why any linear convergent iteration of the form (1.155) can be used to define a preconditioner for the matrix \( A \) and thus be embedded in a Krylov subspace method. In particular, this will lead us to the concept of inner-outer iteration.
1.4. Analysis of the conjugate gradient method

It is reasonable to assume that the solution \( z = A^{-1}b \) is the fixed point for (1.155):

\[
z = Mz + Nb.
\]

This can be rewritten as

\[
A^{-1}b = MA^{-1}b + Nb.
\]

Since \( b \) can be arbitrary, the later implies

\[
A^{-1} = MA^{-1} + N \Rightarrow N = (I - M)A^{-1}.
\]

Assume that iteration (1.155) geometrically converges with a factor \( \xi < 1 \), that is, \( \|M\| = \xi < 1 \), where \( \|\cdot\| \) is an operator norm. Consider a linear operator \( B : \mathbb{C}^n \to \mathbb{C}^n \): For any vector \( x \in \mathbb{C}^n \), we define \( y = Bx \) as follows. The vector \( y \) is an approximate solution to the system \( Az = x \) resulting after \( k \) iterations of the method (1.155) applied to solve \( Az = x \), with the zero initial guess \( y^0 = 0 \):

\[
\begin{align*}
y^0 &= 0, \\
y^{i+1} &= My^i + Nx, \quad i = 0, \ldots, k - 1, \\
y &= y^k.
\end{align*}
\]

(1.157)

The algorithm (1.157) defines a linear operator \( B : \mathbb{C}^n \to \mathbb{C}^n \). Without causing confusion, the matrix of this operator is also denoted by \( B \). We will see in a moment that this \( B \) can be interpreted as an approximate inverse of \( A \) and thus considered as the preconditioner for \( A \). Instead of applying a Krylov subspace method to \( Az = b \), one applies it to \( BAz = Bb \) or \( ABx = b \).

Since the iteration method is linear, the algorithm (1.157) gives

\[
y = M^k y^0 + \tilde{N}x = \tilde{N}x,
\]

with \( \tilde{N} = (\sum_{j=0}^k M^j)N \). We use relation (1.156) to obtain formally

\[
B = \tilde{N} = (I - M^k)A^{-1}.
\]

(1.158)

The equality (1.158) serves only for analysis and never for computing \( B \). In practice, the matrix \( B \) is never computed explicitly. Instead, the algorithm (1.157) defines the rule for calculating the product of the matrix \( B \) and an arbitrary vector \( x \). Computing \( y = Bx \) essentially consists of executing \( k \) linear iterations (1.155).

**Theorem 1.33.** Let \( B \) be a preconditioner for a matrix \( A \in \mathbb{C}^{n \times n} \) defined by \( k \) iterations of the linear iteration process (1.157), with an iteration matrix \( M \) such that \( \|M\| = \xi < 1 \). Then we have

\[
\|BA\| \leq 1 + \xi^k, \quad \text{cond}(BA) \leq \frac{1 + \xi^k}{1 - \xi^k}, \quad (1 - \xi^k)\|x\|^2 \leq \text{Re}\langle BAx, x \rangle \quad \forall x \in \mathbb{C}^n.
\]

**Proof.** One gets from (1.158) the equality \( BA = I - M^k \). The estimate for \( \|BA\| \) follows from

\[
\|I - M^k\| \leq \|I\| + \|M^k\| \leq \|I\| + \|M\|^k \leq (1 + \xi^k).
\]
We also have
\[
\| (I - M^k)^{-1} \| = \sup_{x \neq 0} \frac{\| (I - M^k)^{-1} x \|}{\| x \|} = \sup_{y \neq 0} \frac{\| y \|}{\| (I - M^k) y \|} \leq \sup_{y \neq 0} \frac{\| y \|}{\| y \| - \| M^k y \|} \leq \sup_{y \neq 0} (1 - \| M^k \|) \| y \| \leq (1 - \xi^k)^{-1}.
\]
Therefore, it holds that
\[
\hat{\eta} := \text{cond}(BA) = \| I - M^k \| \| (I - M^k)^{-1} \| \leq \frac{1 + \xi^k}{1 - \xi^k}.
\]
Similarly, we estimate for arbitrary \( x \in \mathbb{C}^m \):
\[
\Re \langle BA x, \hat{x} \rangle = \| x \|^2 - \Re \langle M^k x, \hat{x} \rangle \geq \| x \|^2 - \langle M^k x, \hat{x} \rangle \geq (1 - \| M^k \|) \| x \|^2 \geq (1 - \xi^k) \| x \|^2.
\]

We illustrate the advantage of using the linear method (1.155) to define a preconditioner for a Krylov subspace method rather than a standalone solver. Assume for a moment that \( A \) is Hermitian positive definite and \( M \) is Hermitian, yielding that \( BA \) is Hermitian positive definite (explain why!). Thanks to (1.121), we can estimate the convergence factor of the preconditioned CG applied to solve \( Az = b \), with \( B \) defined through (1.157):
\[
\sqrt{\frac{\text{cond}(AB) - 1}{\text{cond}(AB) + 1}} = \sqrt{\frac{1 + \xi^k}{1 - \xi^k} - 1} = \frac{1 - \sqrt{1 - \xi^k}}{\xi^k} < \xi^k.
\]
Therefore, the convergence factor of the PCG method is less than the one of simple iteration (1.155). Besides this advantage, the CG method, like the Krylov subspace solver, may benefit from eigenvalue clustering (demonstrating superlinear convergence) as well as other matrix properties, as discussed in this section.

If a linear method is a convergent iteration, then the GMRES method with corresponding preconditioner enjoys the guaranteed convergence rate. Indeed, the above theorem and estimate (1.95) yield, for the residual of the GMRES method,
\[
\| r_i \| \leq \frac{2 \xi^k}{1 + \xi^k} \| r_{i-1} \|.
\] (1.159)

We recall that the above estimate is likely pessimistic, since it is based on one search direction on every iteration instead of the full Krylov subspace. Therefore, the actual convergence can be better (and in practice is usually much better). It is also worth noting that the estimate (1.95) was proved only for positive definite matrices, while (1.159) is valid for arbitrary nonsingular \( A \): A convergent linear iteration ensures the preconditioned matrix is positive definite!

If several iterations of (1.155) are performed to define the preconditioner, i.e., \( k > 1 \) in (1.157), then the Krylov subspace method is called outer iteration, and the linear method is called inner iteration. In the following chapters, we shall often study different ways of and approaches to defining preconditioners to accelerate the convergence of Krylov
subspace methods. *Quite often these preconditioners will be defined and studied in the form of a convergent linear iteration.*

## Exercises

1.4.1. Let $0 < m \leq a \leq b \leq M$ and let all eigenvalues of a Hermitian positive definite matrix $A$ belong to an interval $[a, b]$, with the exception of $k$ eigenvalues lying in $[m, a]$ and $l$ eigenvalues from $[b, M]$. Prove the following estimate for the $A$-norms of the errors in the CG method:

$$
\|e_i\|_A \leq 2 \left( \frac{M}{m} \right)^k \left( \frac{1 - \sqrt{b}}{1 + \sqrt{b}} \right)^{i-k-l} \|e_0\|_A.
$$

1.4.2. Assume $A$ is a Hermitian positive definite matrix of size $n$ with 1 standing on the main diagonal. Prove that for any positive definite diagonal matrix $D$ it holds that

$$
\text{cond}_2(A) \leq n \text{cond}_2(DAD).
$$

1.4.3. Prove that for the CG method the inequality $\|r_i\|_2 \leq \varepsilon \|r_0\|_2$ holds for all $i \geq k(\varepsilon)$, with $k(\varepsilon) \leq c \ln \varepsilon^{-1}/\ln \ln \varepsilon^{-1}$ and $c$ independent of $\varepsilon$. (Use Theorem 1.31 and inequality $t - 1 \leq (\sigma - 1)^{i-1} (t/\sigma)^i$, $\sigma, t > 1$.)

1.4.4. Consider a positive definite Hermitian $2 \times 2$ block matrix $A$ and $2 \times 2$ block matrix $D = \text{block diag}(A)$:

$$
A = \begin{pmatrix} H_1 & H \\ H^* & H_2 \end{pmatrix}, \quad D = \begin{pmatrix} H_1 & 0 \\ 0 & H_2 \end{pmatrix}.
$$

It is clear that $D$ is also positive definite Hermitian. Show inequality $\det(D^{-1}A) \leq 1$.

*Hint.* Argue that

$$
\det(D^{-1}A) = \det(D^{-1/2}AD^{-1/2}) = \det \left( \begin{pmatrix} I & \tilde{H} \\ \tilde{H}^* & I \end{pmatrix} \right) = \det(I - \tilde{H}^* \tilde{H}).
$$

1.4.5. Deduce from the minimization property (1.151) of the preconditioned CG method that the $A$-norm of its error satisfies the estimate (1.121) if $\text{sp}(AB) \in [m, M]$. 
Chapter 2
Toeplitz Matrices and Preconditioners

2.1 Introduction to Toeplitz Matrices

2.1.1 Toeplitz matrices and polynomials

A matrix $A$ with entries $a_{ij}$ is called a Toeplitz matrix if

$$a_{i_1 j_1} = a_{i_2 j_2} \quad \text{whenever} \quad i_1 - j_1 = i_2 - j_2.$$  

This means that $a_{ij}$ is a function of $i - j$. One may say that Toeplitz matrices reproduce a well-known shift-invariance property for a function $f(x,y)$ of two variables when $f(x,y)$ depends only on the difference $x - y$.

Here is a portrait of a square Toeplitz matrix of order $n = 4$:

$$A = \begin{bmatrix}
a_0 & a_{-1} & a_{-2} & a_{-3} \\
a_1 & a_0 & a_{-1} & a_{-2} \\
a_2 & a_1 & a_0 & a_{-1} \\
a_3 & a_2 & a_1 & a_0
\end{bmatrix}.$$  

Each matrix diagonal consists of entries all equal to each other. This property gives rise to another name for Toeplitz matrices: Sometimes they are called isodiagonal matrices. The numbers $a_{-3}, a_{-2}, a_{-1}, a_0, a_1, a_2, a_3$ are called generators of $A$. The generators are independent parameters that completely define $A$. In a general case, a Toeplitz matrix of order $n$ is determined by $2n - 1$ generators.

In algebra, Toeplitz matrices are commonly associated with polynomials. Given any two polynomials

$$a(x) = a_0 + a_1 x + \cdots + a_m x^m$$  

and

$$b(x) = b_0 + b_1 x + \cdots + b_n x^n,$$

we construct the corresponding lower triangular Toeplitz matrices $A$ and $B$ in the following way. Matrices $A$ and $B$ will be of the same size, which can be any number $N \geq m + n + 1$. Next, we define $A$ by its first column: The first column contains $a_0, \ldots, a_m$ and is extended by zeros to be of height equal to $N$, and similarly, the first column of $B$ begins with $b_0, \ldots, b_n$, followed again by zeros. A simple and important observation is that the product

$$c(x) = a(x)b(x)$$
is a polynomial whose coefficients form the first column of the matrix product

\[ C = AB. \]

These coefficients occupy the first \( m + n + 1 \) entries of the first column of \( AB \); other entries in this column are zeros. Here is an illustration for \( m = n = 2 \) and \( N = 7 \):

\[
\begin{bmatrix}
  c_0 & c_1 & c_2 & c_3 & c_4 & 0 & 0 \\
  c_1 & c_0 & c_2 & c_3 & c_4 & 0 & 0 \\
  c_2 & c_1 & c_0 & c_2 & c_3 & c_4 & 0 \\
  c_3 & c_2 & c_1 & c_0 & c_2 & c_3 & c_4 \\
  c_4 & c_3 & c_2 & c_1 & c_0 & c_2 & c_3 \\
  0 & 0 & c_4 & c_3 & c_2 & c_1 & c_0 \\
  0 & 0 & 0 & 0 & c_4 & c_3 & c_2 \\
\end{bmatrix}
= 
\begin{bmatrix}
  a_0 & a_1 & a_2 & 0 & 0 & 0 & 0 \\
  a_1 & a_0 & a_1 & a_2 & 0 & 0 & 0 \\
  a_2 & a_1 & a_0 & a_1 & a_2 & 0 & 0 \\
  0 & a_2 & a_1 & a_0 & a_1 & a_2 & 0 \\
  0 & 0 & a_2 & a_1 & a_0 & a_1 & a_2 \\
  0 & 0 & 0 & a_2 & a_1 & a_0 & a_1 \\
\end{bmatrix}
\begin{bmatrix}
  b_0 & 0 & 0 & 0 & 0 & 0 & 0 \\
  b_1 & b_0 & 0 & 0 & 0 & 0 & 0 \\
  b_2 & b_1 & b_0 & 0 & 0 & 0 & 0 \\
  0 & b_2 & b_1 & b_0 & 0 & 0 & 0 \\
  0 & 0 & b_2 & b_1 & b_0 & 0 & 0 \\
  0 & 0 & 0 & b_2 & b_1 & b_0 & 0 \\
\end{bmatrix}.
\]

Besides a nice bond with polynomials, we observe that the set of lower triangular Toeplitz matrices possesses an eminent algebraic structure: This set is an algebra.

**Lemma 2.1.** The sum and product of any lower (upper) triangular Toeplitz matrices are lower (upper) triangular Toeplitz matrices.

The proof is straightforward and left to the reader.

**Corollary 2.2.** If a lower (upper) triangular Toeplitz matrix is nonsingular, then its inverse is a lower (upper) triangular Toeplitz matrix.

**Proof.** By the Caley–Hamilton theorem, a matrix is a root of its characteristic polynomial. Thus, if

\[ \det(\lambda I - A) = \lambda^n + \alpha_{n-1} \lambda^{n-1} + \cdots + \alpha_1 \lambda + \alpha_0, \]

then

\[ A^n + \alpha_{n-1} A^{n-1} + \cdots + \alpha_1 A + \alpha_0 I = 0. \]

If \( A \) is nonsingular, then \( \alpha_0 \neq 0 \), and we easily derive that

\[ A^{-1} = -\frac{1}{\alpha_0} (A^{n-1} + \alpha_{n-1} A^{n-2} + \cdots + \alpha_1 I). \]

If \( A \) is a lower (upper) triangular Toeplitz matrix, then, by Lemma 2.1, the right-hand side of the last identity is also a lower (upper) triangular Toeplitz matrix.

Denote by \( b \) and \( c \) the first columns of lower triangular \( B \) and \( C = AB \). It obviously holds that

\[ c = Ab. \]

The link to polynomials leads to the useful observation about computing a matrix-vector product with a Toeplitz matrix: *If a product of two polynomials of order \( n \) can be computed
2.1. Introduction to Toeplitz Matrices

In $W(n)$ arithmetic operations, then the product of a Toeplitz matrix and an arbitrary vector can be computed in $O(W(n))$ arithmetic operations. To verify this claim, it suffices to note that a Toeplitz matrix is a sum of lower and upper triangular Toeplitz matrices.

It is well known that the product of two polynomials of complex coefficients can be computed using

$$W(n) = O(n \log_2 n)$$

elementary arithmetical and logical operations [3, 208]. Therefore, a matrix-vector multiplication with a Toeplitz matrix of order $n$ is of the same complexity.

In this section, we study iterative solvers for linear systems $Ax = b$ with a nonsingular Toeplitz matrix $A$. From the above complexity estimate we conclude that each matrix-vector product evaluation is inexpensive. Hence, we need only to take a closer look at the number of iterations needed for an iterative method to converge with a prescribed tolerance. As we know, this number can be minimized by choosing a good preconditioner.

We note that cost-efficient direct, i.e., noniterative, solvers for Toeplitz matrices are also available. Direct solvers with complexity $O(n^2)$ have been known since the 1940s [130]. Since then they have been modified and improved many times; see, e.g., [102]. Under certain restrictions on Toeplitz matrices, direct solvers with complexity as low as $O(n \log^2 n)$ have been proposed [4, 24, 46, 64, 147].

One way or another, direct solvers for Toeplitz systems are in close connection with a wonderful algebraic structure of the corresponding inverse matrices: The inverse to a Toeplitz matrix is a sum of two matrices, where each is a product of two Toeplitz matrices [84, 195]. One may say that the inverses to Toeplitz matrices are still Toeplitz-like and that, moreover, it is possible to construct algorithms for matrices of various Toeplitz-like structures [102, 109]. No similar claim is known for multilevel Toeplitz matrices. Later in this chapter, we shall consider multilevel Toeplitz matrices for which iterative solvers have no alternative. For general concepts of structures in multilevel matrices see [199].

2.1.2 Circulant matrices and Fourier matrices

A matrix $A$ of order $n$ with entries $a_{ij}$ is called a circulant matrix if

$$a_{i_1j_1} = a_{i_2j_2} \quad \text{whenever} \quad i_1 - j_1 = i_2 - j_2 \pmod{n}.$$  

A circulant matrix is thus a particular case of a Toeplitz matrix. For $n = 4$ a circulant matrix is as follows:

$$A = \begin{bmatrix}
a_0 & a_3 & a_2 & a_1 \\
a_1 & a_0 & a_3 & a_2 \\
a_2 & a_1 & a_0 & a_3 \\
a_3 & a_2 & a_1 & a_0
\end{bmatrix}.$$  

Any circulant matrix is completely defined by any of its columns or any of its rows. In particular, $A$ is determined by its first column $a$, and the quantities $a_0, \ldots, a_{n-1}$ are the generators of $A$. More precisely,

$$A = [a, Pa, \ldots, P^{n-1}a],$$

where $P$ is the cyclic downward shift matrix of the form

$$P = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\
1 & 0 & 1 & \cdots & 0 \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
0 & \cdots & 1 & 0 & 1 \end{bmatrix}.$$
The matrices \( I, P, \ldots, P^{n-1} \) yield a basis in the linear space of circulant matrices of order \( n \). Obviously, it holds
\[
A = a_0 I + a_1 P + \cdots + a_{n-1} P^{n-1}.
\]
Now, take any complex number \( \epsilon \) such that \( \epsilon^n = 1 \). Then
\[
[1, \epsilon, \ldots, \epsilon^{n-1}]P = \epsilon [1, \epsilon, \ldots, \epsilon^{n-1}]
\]
and, hence,
\[
[1, \epsilon, \ldots, \epsilon^{n-1}]P^k = \epsilon^k [1, \epsilon, \ldots, \epsilon^{n-1}].
\]
By multiplying both sides by \( a_k \) and summing up for \( k = 0, 1, \ldots, n-1 \) we obtain
\[
[1, \epsilon, \ldots, \epsilon^{n-1}]A = \lambda(\epsilon) [1, \epsilon, \ldots, \epsilon^{n-1}],
\]
where
\[
\lambda(\epsilon) = a_0 + a_1 \epsilon + \cdots + a_{n-1} \epsilon^{n-1}.
\]
In the matrix notation, we have
\[
FA = DF,
\]
where \( D \) is a diagonal matrix of the form
\[
D = \text{diag}(\lambda(1), \lambda(\epsilon), \ldots, \lambda(\epsilon^{n-1}))
\]
or, equivalently,
\[
D = \text{diag}(Fa), \tag{2.1}
\]
and \( F \) is the so-called Fourier matrix:
\[
F = \begin{bmatrix}
1 & 1 & 1 & \cdots & 1 \\
1 & \epsilon & \epsilon^2 & \cdots & \epsilon^{n-1} \\
1 & \epsilon^2 & \epsilon^{2^2} & \cdots & \epsilon^{2^{(n-1)}} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & \epsilon^{n-1} & \epsilon^{(n-1)^2} & \cdots & \epsilon^{(n-1)(n-1)} \\
\end{bmatrix} \tag{2.2}
\]
with
\[
\epsilon = \exp\{i2\pi/n\}, \quad i^2 = 1.
\]
It is easy to check that \( F^*F = n \cdot I \). The proof of this reduces to the summation of geometric progressions. Hence, the matrix \( n^{-1/2}F \) is unitary and
\[
F^{-1} = n^{-1} F^*.
\]
We actually have proved the following theorem.

**Theorem 2.3.** For any circulant matrix \( A \) of order \( n \) with the first column \( a \), it holds that
\[
A = F^{-1} DF = n^{-1} F^* DF,
\]
where \( D \) is a diagonal matrix containing the eigenvalues of \( A \), and \( F \) is the Fourier matrix given by (2.2). The eigenvalues of \( A \) are the entries of the vector \( Fa \), where \( a \) is the first column of \( A \).
2.1. Introduction to Toeplitz Matrices

The fast Fourier transform

The fast Fourier transform shows that the columns of $F^{-1}$, and hence of $F^*$, are the eigenvectors of $A$. Since $F^*$ coincides with $F$, a matrix where the entries of $F$ are replaced with their complex conjugates, and $e^{u-1} = e^{-1}$, it follows that $F^*$ can be obtained from $F$ by a permutation of columns. Thus, we conclude that the columns of $F$ are the permuted eigenvectors of $A$.

We see from Theorem 2.3 and the formula (2.1) that the eigenvalues of any circulant matrix are easily found via one matrix-by-vector multiplication with the Fourier matrix. Moreover, all circulant matrices of the same order $n$ have common eigenvectors which are the columns of the Fourier matrix.

Further, a matrix-vector product with a circulant matrix reduces to three matrix-vector products with the Fourier matrix. Since circulants form an algebra, the inverse to any nonsingular circulant matrix is also a circulant matrix and can be found using the formula

$$A^{-1} = n^{-1} F^* D^{-1} F.$$ 

In order to find $A^{-1}$, it is sufficient to compute its first column, as it is a generator of $A^{-1}$. In practice, this reduces to two matrix-by-vector multiplications with $F$.

If we need to solve a linear system $Ax = b$, this reduces to three matrix-by-vector multiplications with $F$ (Exercise 2.1.1).

More algebraic properties of circulant matrices and possible generalizations can be found in [63].

2.1.3 The fast Fourier transform

We already apprehend the importance of computation of $y = Fx$. This operation with $x$ as an input and $y$ as an output is called the Fourier transform.

By a fast Fourier transform we mean any algorithm for finding the Fourier transform with complexity of $O(n \log_2 n)$. The fast Fourier transform is ubiquitous in numerical analysis and signal processing and is included in the list of the 10 most important algorithms of the 20th century according to [69]. Here we review an algebraic approach to derive one of such algorithms.

Suppose for a moment that $n = 2^d$. We show how a Fourier transform of order $n$ can be reduced to a couple of Fourier transforms of order $n/2$.

When doing this, it is convenient to numerate rows and columns by indices ranging from 0 to $n - 1$. Thus,

$$F = F_n = \begin{bmatrix} e^{i \frac{2\pi st}{n}} \end{bmatrix}, \quad 0 \leq s, \ t \leq n - 1.$$ 

Permute the rows of $F_n$ to make a subsequence of all even rows of $F_n$ that is followed by a subsequence of all odd rows of $F_n$. This permutation of rows can be done via premultiplication of $F_n$ by a permutation matrix $\Pi_n$. We have

$$\Pi_n F_n = \begin{bmatrix} [e_n^{2kl}] & [e_n^{2k(n/2+l)}] \\ [e_n^{2(k+1)l}] & [e_n^{2(k+1)(n/2+l)}] \end{bmatrix}, \quad 0 \leq k, \ l \leq n/2 - 1.$$ 

Let us recognize that $\Pi_n F_n$ has the following block form:

$$PF = \begin{bmatrix} F_{n/2} & F_{n/2} \\ W_n/2^* F_{n/2} & -W_n/2^* F_{n/2} \end{bmatrix} = \begin{bmatrix} I_{n/2} & 0 \\ 0 & W_n/2 \end{bmatrix} \begin{bmatrix} F_{n/2} & 0 \\ 0 & F_{n/2} \end{bmatrix} \begin{bmatrix} I_{n/2} & I_{n/2} \\ I_{n/2} & -I_{n/2} \end{bmatrix},$$

where

$$P = \begin{bmatrix} I_{n/2} & 0 \\ 0 & W_n/2 \end{bmatrix} \begin{bmatrix} I_{n/2} & I_{n/2} \\ I_{n/2} & -I_{n/2} \end{bmatrix} = \begin{bmatrix} W_n/2 & 0 \\ 0 & W_n/2 \end{bmatrix}.$$
Chapter 2. Toeplitz Matrices and Preconditioners

where

\[ W_{n/2} = \text{diag}(1, \epsilon_n, \ldots, \epsilon_n^{n/2-1}), \quad \epsilon_n = \exp\{i2\pi/n\}. \]

It becomes evident that the initial transform of order \(n\) reduces to the two similar transforms of order \(n/2\). It is easy to calculate that the cost of this reduction is \(n\) additions/subtractions and \(n/2\) multiplications by the entries of the diagonal matrix \(W_{n/2}\). Since the number of recursive steps is equal to \(d = \log_2 n\), in total we need only

\[
\sum_{d}^{n + 2 \cdot (n/2) + \cdots + 2^{d-1} \cdot (n/2^{d-1})} = nd = n \log_2 n
\]

additions/subtractions and \(\frac{1}{2} n \log_2 n\) multiplications.

Consequently, the complexity of multiplication of a circulant matrix by a vector is \(O(n \log_2 n)\) provided that \(n\) is a power of 2.

Now, consider a Toeplitz matrix of order \(n\) where \(n\) is an arbitrary positive integer. We can always embed it as a block into a larger circulant matrix, whose order is a power of 2 in between \(2n\) and \(4n\). This proves that the complexity of multiplication of an arbitrary Toeplitz matrix of order \(n\) by a vector is \(O(n \log_2 n)\) for any \(n\).

If \(n\) is not a power of 2, then the complexity of the Fourier transform of order \(n\) is still \(O(n \log_2 n)\). It suffices to note that \(F_n\) reduces to a Toeplitz matrix via multiplications from both sides by a diagonal matrix:

\[ F_n = D_n T_n D_n, \]

where

\[ T_n = [\epsilon_n^{(s-t)^2/2}], \quad D_n = \text{diag}(\epsilon_n^{s^2/2}), \quad 0 \leq s, t \leq n-1. \]

Exercises

2.1.1. Prove that solving a system of linear algebraic equations with a circulant matrix of order \(n\) can be done by invoking three Fourier transforms of order \(n\) plus \(O(n)\) arithmetic operations.

2.1.2. Prove that the arithmetic complexity of solving a system of linear algebraic equations with a circulant matrix of order \(n\), where \(n\) is not necessarily a power of 2, is \(O(n \log_2 n)\).

2.1.3. Prove that the complexity of computing \(A^{-1}\) for a nonsingular circulant matrix \(A\) of order \(n\) is \(O(n \log_2 n)\).

2.1.4. Consider a circulant \(A\) of order \(n\) and multiply all its elements below the main diagonal by some \(\epsilon \neq 0\). Denote the new matrix by \(A_\epsilon\). Prove that the complexity of computing \(A_\epsilon^{-1}\) is \(O(n \log_2 n)\).

2.1.5. Let \(A\) be a nonsingular lower triangular Toeplitz matrix. Prove that the complexity of computing \(A^{-1}\) is \(O(n \log_2 n)\).

2.1.6. Prove that the coefficients of the polynomial \(p(x) = \prod_{i=1}^{n}(x - x_i)\), with given roots \(x_1, \ldots, x_n\) can be computed in \(O(n \log_2^2 n)\) operations.

2.1.7. Prove that the product of two Toeplitz matrices of order \(n\) can be computed in \(O(n^2)\) operations and cannot be found in \(o(n^2)\) operations.
2.2. Preconditioners and applications

2.2.1. Circulant preconditioners

Circulant matrices are naturally considered as preconditioners for Toeplitz matrices. Let $A$ be a Toeplitz matrix with generators $a_k$, where $-n+1 \leq k \leq n-1$, and let $C$ be a circulant matrix with generators $c_k$, where $0 \leq k \leq n-1$.

A circulant matrix $C$ is called a simple circulant preconditioner for $A$ if

$$c_k = a_k \quad \forall -n/2 < k < n/2. \quad (2.3)$$

For $n = 4$ we obtain

$$C = \begin{bmatrix} a_0 & a_{-1} & c_2 & a_1 \\ a_1 & a_0 & a_{-1} & c_2 \\ c_2 & a_1 & a_0 & a_{-1} \\ a_{-1} & c_2 & a_1 & a_0 \end{bmatrix}.$$

As we observe in this example, the condition (2.3) does not determine the value of $c_2$. In general, (2.3) does not define $c_{n/2}$ if $n$ is even.

In applications of circulant preconditioners, it is usually assumed that $a_k \to 0$ as $k \to \infty$. Thus, we can impose, for simplicity, that $c_{n/2} = 0$ if $n$ is even.

If $n$ is odd, then $C$ is completely defined by (2.3). For example, for $n = 5$ we obtain

$$C = \begin{bmatrix} a_0 & a_{-1} & a_{-2} & a_2 & a_1 \\ a_1 & a_0 & a_{-1} & a_{-2} & a_2 \\ a_2 & a_1 & a_0 & a_{-1} & a_{-2} \\ a_{-1} & a_{-2} & a_1 & a_0 & a_{-1} \\ a_{-2} & a_2 & a_1 & a_0 & a_{-1} \end{bmatrix}.$$

A circulant matrix $C$ is called an optimal circulant preconditioner for $A$ if

$$\|A - C\|_F = \min_{B \text{ is a circulant}} \|A - B\|_F. \quad (2.4)$$

We recall that $\|\cdots\|_F$ denotes the Frobenius matrix norm; see section 1.1. Using the Toeplitz structure of $A$ we easily find from (2.4) that

$$c_k = \frac{(n-k)a_k + ka_{k-n}}{n}, \quad 0 \leq k \leq n-1.$$

We should note at this time that the definition (2.4) makes sense as well if $A$ is an arbitrary matrix. If $a_{ij}$ are the entries of $A$, then

$$c_k = n^{-1} \sum_{i-j=k \pmod{n}} a_{ij}. \quad (2.5)$$

Lemma 2.4. Suppose that $A$ is an arbitrary matrix of order $n$ and $C$ is its optimal circulant preconditioner. Then it holds that

$$\text{diag}(FCF^{-1}) = \text{diag}(FAF^{-1}),$$

where $F$ is the Fourier matrix of order $n$. 
Proof. It is easy to check that
\[ FAF^{-1} = QAQ^{-1}, \quad FCF^{-1} = QCQ^{-1}, \]
where \( Q = n^{-1/2}F \) and \( Q^{-1} = n^{-1/2}F^* \) are both unitary matrices. Since the Frobenius norm of a matrix does not change after multiplication from any side by any unitary matrix (explain why!), we obtain
\[ ||A - C||_F = ||\hat{A} - \hat{C}||_F, \]
where \( \hat{A} = FAF^{-1} \) is the Fourier image of \( A \) and \( \hat{C} = FCF^{-1} \) is the Fourier image of \( C \). Since \( C \) is a circulant, by Theorem 2.3 its Fourier image is a diagonal matrix consisting of the eigenvalues of \( C \). Thus, the minimization over circulants in (2.4) reduces to the minimization over all diagonal matrices. Therefore, we get
\[ ||\hat{A} - \hat{C}||_F = \min_{D \text{ is diagonal}} ||\hat{A} - D||_F. \]
It trivially follows that \( \hat{C} = \text{diag}(\hat{A}) \) and this completes the proof. \( \square \)

Corollary 2.5. If \( A \) is a Hermitian positive definite matrix, then its optimal circulant preconditioner \( C \) is Hermitian positive definite.

Proof. Note that \( F^{-1}AF \) is also a Hermitian positive definite matrix, and hence its main diagonal consists of real positive numbers. Thus, all the eigenvalues of \( C \) are positive.

Recall the definition of the \( K \)-condition number for \( M \), a matrix of size \( n \times n \) (see section 1.4):
\[ K(M) = \frac{(\text{tr}(M)/n)^n}{\det M}. \]
The next corollary of Lemma 2.4 shows another minimization property of the optimal circulant preconditioner.

Corollary 2.6. Assume that \( A \) is a Hermitian positive definite matrix and \( C \) is its optimal circulant preconditioner. Then
\[ K(C^{-1}A) = \min_{B \text{ is a nonsingular circulant}} K(B^{-1}A), \]
where \( K(C^{-1}A) \) is the \( K \)-condition number of \( C^{-1}A \).

Proof. Let \( \hat{a}_i = \hat{A}_{ii} \) and \( \hat{c}_i = \hat{C}_{ii} \). Then, one easily verifies that
\[ K(C^{-1}A) = K(\hat{C}^{-1}\hat{A}) = \left( \frac{\sum_{i=1}^n \hat{a}_i / \hat{c}_i}{n} \right)^n \frac{\prod_{i=1}^n \hat{c}_i}{\det A}. \]
Thanks to the inequality between the arithmetic and geometric means, we get
\[ \left( \frac{\sum_{i=1}^n \hat{a}_i / \hat{c}_i}{n} \right)^n \geq \prod_{i=1}^n \hat{a}_i / \hat{c}_i. \]
Hence,
\[ \left( \frac{\sum_{i=1}^n \hat{a}_i / \hat{c}_i}{n} \right)^n \prod_{i=1}^n \hat{c}_i \geq \prod_{i=1}^n \hat{a}_i \iff K(C^{-1}A) \geq \frac{\prod_{i=1}^n \hat{a}_i}{\det A}. \]
2.2. Preconditioners and applications

We know that the equality holds iff

\[ \hat{a}_1/\hat{e}_1 = \cdots = \hat{a}_n/\hat{e}_n =: h \]

for some number \( h \). The optimal circulant preconditioner satisfies these equations with \( h = 1 \), and hence it minimizes \( K(C^{-1}A) \) over all nonsingular circulants.

The idea of using circulants as preconditioners was probably first considered in [151]. The simple circulant preconditioner was introduced by Strang [189], and the optimal circulant preconditioner was proposed by Chan [58]. Then, theoretical analysis appeared in a series of papers; see, in particular, [53], and see [54], one of the first survey papers on circulant preconditioners. A good introduction to the topic and some developments can be found in the books [56, 149].

We see that the positive definiteness of a matrix is inherited by its optimal circulant preconditioner. Moreover, this preconditioner minimizes the \( K \)-condition number of the preconditioned matrix over the space of all circulant matrices.

We remark that a simple circulant preconditioner, in general, does not possess any of these properties. However, in many particular applications the positive definiteness property still holds for a simple circulant preconditioner, and it may outperform the optimal preconditioner, sometimes significantly.

2.2.2 Analysis of optimal preconditioners

Let us consider a real Lebesgue-integrable \( 2\pi \)-periodic function \( f(t) \) and its formal Fourier series

\[ f(t) = \sum_{k=-\infty}^{\infty} a_k \exp\{ikt\}. \]

Later on we assume that the values \( f(t) \) are defined for all \( t \), although it may happen that the series converge not for all values of \( t \).

Given the Fourier series, we may naturally associate a function \( f(t) \) with a variety of Toeplitz matrices. First of all, these are Toeplitz matrices with generators \( a_k \):

\[ A_n = [a_{k-l}], \quad 0 \leq k, l \leq n. \]

Thus, \( f(t) \) generates a sequence of Toeplitz matrices \( A_n \), with \( n = 1, 2, \ldots \). For any \( A_n \) there exists the optimal circulant preconditioner \( C_n \).

As we know from the general theory, the convergence of iterative methods based on the Krylov subspaces is largely governed by the distribution of the eigenvalues of \( C_n^{-1}A_n \). We shall study the spectrum of \( C_n^{-1}A_n \) using the smoothness properties of \( f(t) \). This function is called the generating function or the symbol for \( A_n \) and \( C_n \).

**Lemma 2.7.** For any column vector \( x = [x_0, x_1, \ldots, x_{k-1}]^T \), the inner product \( \langle A_n x, x \rangle \) can be expressed by the formula

\[ \langle A_n x, x \rangle = \frac{1}{2\pi} \int_{-\pi}^\pi f(t) \left| \sum_{k=0}^{n-1} x_k \exp\{ikt\} \right|^2 dt. \]

**Proof.** Recall that the Fourier series coefficients can be expressed as

\[ a_k = \frac{1}{2\pi} \int_{-\pi}^\pi f(t) \exp\{-ikt\} dt. \]
Therefore,
\[
\frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \left| \sum_{k=0}^{n-1} x_k \exp\{ikt\} \right|^2 dt
\]
\[
= \sum_{k=0}^{n-1} \bar{x}_k \sum_{l=0}^{n-1} x_l \left( \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \exp\{-i(k-l)t\} dt \right) = \sum_{k=0}^{n-1} \bar{x}_k \sum_{l=0}^{n-1} a_{k-l} x_l = \langle A_n x, x \rangle. \]

Corollary 2.8. Suppose that \( f(t) \) is not a constant on a set of nonzero Lebesgue measure and \( \epsilon_{\text{min}} \leq f(t) \leq \epsilon_{\text{max}} \) for all \( t \).

Then all eigenvalues \( \lambda \) of the Toeplitz matrix \( A_n \) and the optimal circulant preconditioner \( C_n \) satisfy the inequalities
\[
\epsilon_{\text{min}} < \lambda < \epsilon_{\text{max}}
\]
for any \( n = 1, 2, \ldots \).

**Proof.** It is clear that the claim is true for Toeplitz matrices \( A_n \), since any eigenvalue of \( A_n \) is a value of \( \langle A_n x, x \rangle \), where \( x \) is the corresponding eigenvector of unity norm. For the optimal circulants \( C_n \), we deduce from Lemma 2.4 that
\[
\langle C_n x, x \rangle = \langle A_n x, x \rangle,
\]
where \( x \) is any column of the Fourier matrix \( F_n \). Hence, (2.6) holds true for any eigenvector of \( C_n \). This completes the proof.

We say that \( f(t) \) belongs to the Wiener class of functions with the bound \( \epsilon > 0 \) if
\[
\sum_{k=-\infty}^{\infty} |a_k| \leq \epsilon < +\infty. \tag{2.7}
\]
A consequence of (2.7) is that the Fourier expansion of \( f(t) \) is an absolutely convergent series that tends to \( f(t) \) at every point \( t \).

**Lemma 2.9.** Assume that \( f(t) \) is in the Wiener class with a bound \( \epsilon > 0 \). Then for any \( \epsilon > 0 \) there exist integers \( r = r(\epsilon, \epsilon) \) and \( N = N(\epsilon) \) such that for any \( n \geq N \) the difference \( A_n - C_n \) is split as follows:
\[
A_n - C_n = R_n + E_n,
\]
where
\[
\text{rank} R_n \leq r \quad \text{and} \quad \|E_n\|_\infty \leq \epsilon.
\]

**Proof.** Fix an arbitrary \( \epsilon > 0 \). Due to (2.7), we can find an integer \( s = s(\epsilon) \) such that
\[
\sum_{|k| \geq s} |a_k| \leq \epsilon/4.
\]
Let \( C_n = [c_{k-l}] \). Then \( c_k = c_l \) whenever \( k = l \) mod \( n \). Thus,
\[
c_{-k} = c_{n-k} = \frac{(n-k)a_{n-k} + ka_{n-k}}{n},
\]
and hence,
\[ c_{-k} - a_{-k} = \frac{k}{n} (a_{-k} - a_{-k}). \]

Therefore, for all sufficiently large \( n \) we obtain
\[ \sum_{k=0}^{n-1} |c_{-k} - a_{-k}| = \sum_{k=0}^{n-1} \frac{k}{n} |a_{-k} - a_{-k}| \leq \sum_{k \geq s} (|a_k| + |a_{-k}|) \leq \frac{\varepsilon}{4} + \frac{cs}{n} \leq \frac{\varepsilon}{2}. \]

By the same arguments, for sufficiently large \( n \) we have
\[ \sum_{k=0}^{n-1} |c_k - a_k| \leq \frac{\varepsilon}{2}. \]

Let us define a matrix \( R_n \) as follows:
\[ (R_n)_{ij} = \begin{cases} (A_n - C_n)_{ij}, & |i - j| \geq n - s, \\ 0, & \text{otherwise}. \end{cases} \]

Since \( R_n \) has a zero band and only nonzero entries in the \( s \times s \) blocks at the corners, we conclude that
\[ \text{rank } R_n \leq 2s, \]
so we can set \( r = 2s. \)

We will need the following well-known result on eigenvalues distribution of a Hermitian matrix perturbed by a low rank matrix. The next theorem is known as an interlacing theorem for eigenvalues; see, e.g., [86].

**Theorem 2.10.** Assume \( A = A^* \) and \( B = A + E \) for some \( E = E^* \), \( \text{rank}(E) = 1 \). Then the eigenvalues of matrices \( A \) and \( B \) interlace: It holds that either
\[ \lambda_i(A) \in [\lambda_i(B), \lambda_{i+1}(B)] \quad \text{for } i = 1, \ldots, n-1 \]
or
\[ \lambda_i(B) \in [\lambda_i(A), \lambda_{i+1}(A)] \quad \text{for } i = 1, \ldots, n-1. \]

For each \( n = 1, 2, \ldots \) consider \( n \)-element sets \( \Lambda_n = \{ \lambda_i^{(n)} \}_{i=1}^n \). Assume that \( \lambda_i^{(n)} \) are complex numbers. We say that a complex number \( \alpha \) is a proper cluster for the sets \( \Lambda_n \), with \( n = 1, 2, \ldots \), if for any \( \varepsilon > 0 \) there is a positive number \( c(\varepsilon) \) such that for all sufficiently large \( n \) the number of elements \( \lambda \in \Lambda_n \) such that \( |\lambda - \alpha| > \varepsilon \) does not exceed \( c(\varepsilon) \). This means that for all sufficiently large \( n \), almost all elements of \( \Lambda_n \) lie in the \( \varepsilon \)-vicinity of \( \alpha \).

**Theorem 2.11.** Assume that \( f(t) \) belongs to the Wiener class with the bound \( c > 0 \) and \( f(t) \geq c_{\text{min}} > 0 \) for all \( t \). Then all the eigenvalues \( \lambda \) of the preconditioned matrices \( C^{-1}_n A_n \) are located in the closed interval \([c_{\text{min}}/c, c/c_{\text{min}}]\), and \( \alpha = 1 \) is a proper cluster for the eigenvalue sets.
Proof. Since \( f(t) \) is real, the matrices \( A_n \), and hence \( C_n \), are Hermitian. The eigenvalues of \( C_n^{-1} A_n \) coincide with those of Hermitian matrices \( C_n^{-1/2} A_n C_n^{-1/2} \), and hence they are real positive numbers. By Corollary 2.8, any eigenvalue of \( A_n \) and any one of \( C_n \) belongs to the closed interval \( [c_{\min}, c] \). Consequently,

\[
\lambda(C_n^{-1} A_n) \leq \|C_n^{-1} A_n\|_2 \leq \frac{c}{c_{\min}}.
\]

Similarly, it holds that

\[
\frac{1}{\lambda(C_n^{-1} A_n)} = \lambda(A_n^{-1} C_n) \leq \|A_n^{-1} C_n\|_2 \leq \frac{c}{c_{\min}}.
\]

Further, by Lemma 2.9,

\[
C_n^{-1/2} A_n C_n^{-1/2} = P_n + Q_n,
\]

where

\[
P_n = I + C_n^{-1/2} E_n C_n^{-1/2} \quad \text{and} \quad Q_n = C_n^{-1/2} R_n C_n^{-1/2}.
\]

Since the matrices \( A_n \) and \( C_n \) are Hermitian, the matrices \( E_n \) and \( R_n \) are Hermitian as well. The eigenvalues of \( P_n \) are located in the interval \( [1 - c \varepsilon, 1 + c \varepsilon] \). When we add to \( P_n \) a Hermitian matrix of rank bounded by \( r \), then only at most \( r \) eigenvalues may leave the interval \( [1 - c \varepsilon, 1 + c \varepsilon] \) (check that this follows from Theorem 2.10). It remains to observe that rank \( Q_n \leq r \).

Suppose that we solve linear systems \( A_n x_n = b_n \) by the conjugate gradient (CG) method with the preconditioner \( C_n \). Theorem 2.11 and convergence results for the CG method (see subsection 1.4.2), imply that the number of iterations to ensure the error norm to be less than \( \varepsilon \) times the initial error norm can be bounded independent of \( n \). Moreover, from the clustering property of the eigenvalues we infer that for any \( 0 < q < 1 \) and sufficiently large \( n \) the ratio between the error at step \( k \) and the initial one is bounded by \( cq^k \), where the constant \( c \) depends on \( q \). In such cases, one says that the convergence is superlinear. This type of convergence behavior is typically observed when the condition number is rather high, but most eigenvalues are clustered near a single value. In effect, after several initial steps, the method starts to converge with a rate determined only by eigenvalues from the interval \( [1 - c \varepsilon, 1 + c \varepsilon] \).

As we have seen in this section, the analysis of the preconditioners’ quality hinges on the concept of a spectral cluster. This concept does not make much sense if we are given one particular matrix. It acquires a rigorous definition only if we consider a sequence of matrices. Earlier in this section we introduced one-point clusters. Nonetheless, more general clusters are of outstanding importance for mathematics, in particular, for the analysis of iterative solvers. We recommend the paper [201] as a systematic introduction to the general concept of cluster and its relations to the concept of distribution. In that paper, one can also find some simple tools of matrix theory providing an easy way to prove and generalize several Szego-like theorems on distributions and clustering of eigenvalues and singular values of multilevel Toeplitz matrices. It should be emphasized that the analysis for non-Hermitian matrices is by far more complicated and not in perfect shape as in the case of Hermitian matrices. For non-Hermitian matrices, however, we can more or less easily analyze the distribution of singular values. The relations between the eigenvalue and singular value clusters were first studied in [204, 205]. It was shown in these
papers that, under amazingly mild assumptions an existence of one-point eigenvalue cluster follows from the existence of a one-point cluster for singular values. General clusters were also discussed in the monograph [202]. Classical approaches to asymptotic spectral distributions for Toeplitz matrices can be found in [90, 26].

2.2.3 Analysis of simple preconditioners

Denote by $S_n$ the simple circulant preconditioner for $A_n$. To study the spectrum of $S_n^{-1}A_n$, we largely follow the same line as we did with the optimal circulants, and obtain similar results.

There is, however, one difference in the analysis of optimal and simple circulants: the inequalities $c_{\text{min}} \leq f(t) \leq c_{\text{max}}$ do not imply that the eigenvalues of $S_n$ belong to the interval $[c_{\text{min}}, c_{\text{max}}]$ for all $n$. Nevertheless, this claim is valid in an asymptotic form. We need to extend the interval to $[c_{\text{min}} - \delta, c_{\text{max}} + \delta]$ with any sufficiently small $\delta > 0$, then all the eigenvalues of $S_n$ fall inside the extended interval for all sufficiently large $n$. To prove this, it is enough to assume that $f(t)$ is in the Wiener class. In the study of optimal circulants, this condition was not imposed from the very beginning, but appeared one step later in Lemma 2.9 and Theorem 2.11.

Denote by $f_s(t)$ the partial sum of the Fourier series:

$$f_s(t) = \sum_{k=-s}^{s} a_k \exp(ikt).$$

Lemma 2.12. The eigenvalues of the simple circulant preconditioner $S_n$ are given by the formula

$$\lambda_k = f_s\left(\frac{(n-1)/2}{n} \cdot \frac{2\pi}{n} \right), \quad 1 \leq k \leq n. \quad (2.8)$$

Proof. If $n = 2m + 1$, then the first column of $S_n$ is

$$(a_0, a_1, \ldots, a_m, a_{-m}, \ldots, a_{-1})^T.$$

According to Theorem 2.3, the eigenvalues of $S_n$ are

$$\lambda_k = \sum_{l=0}^{m-1} a_k \exp\left(\frac{2\pi}{n} kl\right)$$

$$= a_0 + \sum_{l=1}^{m} a_l \exp\left(\frac{2\pi}{n} kl\right) + \sum_{l=1}^{m} a_{-l} \exp(ik(n-l)) = f_m\left(\frac{2\pi}{n} k\right).$$

We have $m = [(n-1)/2]$.

If $n = 2m$ then the first column of $S_n$ is

$$(a_0, a_1, \ldots, a_{m-1}, 0, a_{-m+1}, \ldots, a_{-1})^T,$$

and the eigenvalues of $S_n$ read

$$\lambda_k = a_0 + \sum_{l=1}^{m-1} a_l \exp\left(\frac{2\pi}{n} kl\right) + \sum_{l=1}^{m-1} a_{-l} \exp(ik(n-l)) = f_{m-1}\left(\frac{2\pi}{n} k\right).$$

We finish the proof, noticing that $m - 1 = [(n-1)/2]$.
Corollary 2.13. Let \( c_{\min} \leq f(t) \leq c_{\max} \) for all \( t \) and assume (2.7). For any \( \delta > 0 \) there is \( N = N(\delta) \) such that for any \( n \geq N \) all the eigenvalues of \( S_n \) belong to \( [c_{\min} - \delta, c_{\max} + \delta] \).

**Proof.** It suffices to note that (2.7) implies the uniform convergence of the partial Fourier sums to the generating function \( f(t) \).

Lemma 2.9 holds as it was stated if \( C_n \) is replaced by \( S_n \). Only this change is needed to adjust the proof. Below we reconsider Theorem 2.11 and adapt it to the case of simple circulants.

**Theorem 2.14.** Assume that \( f(t) \) belongs to the Wiener class with a bound \( c > 0 \), and let \( f(t) \geq c_{\min} > c_0 > 0 \) for all \( t \). Then for all sufficiently large \( n \) all the eigenvalues of \( S_n^{-1} A_n \) are located in the interval \( c_0 / \epsilon \leq \lambda \leq c / \epsilon_0 \) and \( \alpha = 1 \) is a proper cluster of the eigenvalue sets.

We skip the proof, since it largely mimics the proof of Theorem 2.11.

### 2.2.4 Which circulant preconditioner is better?

Under the assumptions of Theorems 2.11 and 2.14 we observe only a minor difference between optimal and simple circulant preconditioners.

Moreover, even without any preconditioner, the number of iterations required by the CG method to meet a fixed tolerance \( \epsilon \) for the relative error norm depends only on \( \epsilon \) and not on \( n \). However, this is no longer true if \( f(t) = 0 \) at some \( t \). In this case, simple and optimal preconditioners demonstrate essentially different convergence properties.

Let us consider a more general definition of a cluster for a sequence of \( n \)-element sets \( \Lambda_n = \{ \lambda^{(n)}_j \}, n = 1, 2, \ldots \). An intuitive definition is as follows: This is a complex number \( \alpha \) where all but a relatively small number of the points \( \lambda^{(n)}_i \) are accumulated. A rigorous definition is slightly more involved.

Consider any \( \epsilon > 0 \) and all outliers for the \( \epsilon \)-neighborhood of \( \alpha \):

\[
\Gamma_\epsilon(n) = \{ \lambda^{(n)}_i : |\lambda^{(n)}_i - \alpha| > \epsilon, 1 \leq i \leq n \}.
\]

We say that \( \alpha \) is a *general cluster* for the sets \( \Lambda_n \) if

\[
|\Gamma_\epsilon(n)| = o(n) \quad \forall \epsilon > 0,
\]

that is,

\[
\lim_{n \to 0} \frac{|\Gamma_\epsilon(n)|}{n} = 0 \quad \forall \epsilon > 0.
\]

We call \( \gamma_\epsilon(n) = |\Gamma_\epsilon(n)| \) (the size of the set \( \Gamma_\epsilon(n) \)) the *outlier function*.

Previously we introduced a more restrictive notion of cluster, the so-called *proper cluster*. In that case, the outlier function is bounded uniformly in \( n \). For a general cluster, this function may grow in \( n \) to infinity, but not too fast compared to \( n \), e.g., like \( \sqrt{n} \) or \( \log n \).

Assume that a \( 2\pi \)-periodic function \( f(t) \) is nonnegative but has a finite number of zeros on each period:

\[
f(t_1) = \cdots = f(t_k) = 0, \\
-\pi < t_1 < \cdots < t_k \leq \pi.
\]

The function \( f(t) \) generates a family of Toeplitz matrices \( A_n \), optimal circulants \( C_n \), and simple circulants \( S_n \). We shall prove the existence of a cluster for sets \( \Lambda_n \) containing the...
2.2. Preconditioners and applications

It turns out that the behavior of $\gamma_\epsilon(n)$ is notably different for optimal and simple circulants.

Denote by $f^{(p)}(t + 0)$ and $f^{(p)}(t - 0)$ the right- and left-hand $p$th order derivatives of $f(t)$ at $t$, respectively. Assume that for every $t_j$ such that $f(t_j) = 0$, there exist positive integers $p_j^+$ and $p_j^-$ such that

$$f^{(1)}(x_j + 0) = \cdots = f^{(p_j^--1)}(x_j + 0) = 0, \quad f^{(p_j^+-1)}(x_j + 0) \neq 0,$$
$$f^{(1)}(x_j - 0) = \cdots = f^{(p_j^--1)}(x_j - 0) = 0, \quad f^{(p_j^+-1)}(x_j - 0) \neq 0, \tag{2.12}$$

and let

$$p := \max\{p_j^\pm : 1 \leq j \leq v\}. \tag{2.13}$$

Thus, all zeros are of finite order, and the maximal order is $p$.

Suppose that a $2\pi$-periodic nonnegative function $f(t)$ is such that its $m$th derivative is piecewise continuous and has a bounded derivative on each continuity interval. Denote by $K_m^{(p)}$ the set of all such functions with a finite number of zeros, and the maximal order according to the conditions (2.11)–(2.13) is $p$.

Even though one cannot guarantee that the simple circulants $S_n$ for functions from $K_m^{(p)}$ are positive definite, it is easy to make them positive definite. As any circulant, $S_n$ is represented by its spectral decomposition $S_n = F_n^{-1}D_nF_n$, with a diagonal eigenvalue matrix $D_n$. Let us transform $D_n$ into $\tilde{D}_n$ by replacing any strictly negative eigenvalue with some positive number. These numbers can be different and their exact values are not important. We are interested only in the number of modified eigenvalues, which determine the number of outliers in the cluster. Construct a new circulant

$$\tilde{S}_n = F_n^{-1}\tilde{D}_nF_n.$$  

We call $\tilde{S}_n$ a modified simple circulant.

Note that optimal circulants for functions from $K_m^{(p)}$ are automatically positive definite according to Corollary 2.8.

**Theorem 2.15.** Suppose a function $f \in K_m^{(p)}$ generates Toeplitz matrices $A_n$, optimal circulants $C_n$, and modified simple circulants $\tilde{S}_n$. Then the eigenvalues of $C_n^{-1}A_n$ possess a general cluster with the outlier function

$$\gamma_\epsilon(n) = \begin{cases} O(n^{p/(p+1)}), & m > 1, \\ O(n^{p/(p+1)} \ln n), & m = 1, \end{cases} \tag{2.14}$$

while the eigenvalues of $\tilde{S}_n^{-1}A_n$ have a general cluster at unity with the outlier function

$$\gamma_\epsilon(n) = O(n^{p/(p+m)}). \tag{2.15}$$
Theorem 2.15 was originally proved in [200]. We shall give the proof in the next subsection. Here we discuss the meaning of this theorem and its link to some classical results in mathematics.

Theorem 2.15 reveals how the smoothness of \( f(t) \) affects the growth of the outlier function. We see a fundamental difference between simple and optimal circulants: increasing smoothness essentially affects the behavior of \( \gamma_\varepsilon(n) \) in the case of simple circulant preconditioners, and affect it very slightly in the case of optimal circulants. This is in good agreement with the well-known convergence results for different methods of summation of the Fourier series: Increasing smoothness makes the partial sums converge faster, but does not influence the rate of convergence for the so-called Cesaro sums. The latter sums, however, may converge under milder smoothness assumptions, even when the partial sums diverge. Nevertheless, the convergence of partial sums is accelerated by the smoothness, while for the Cesaro sums this is not the case. Thus, optimal circulants are like Cesaro sums, while simple ones are like partial sums. The latter connection was already manifested in (2.8). The former connection is supported by Lemma 2.16.

We introduce notation for the Cesaro and partial sums:

\[
\sigma_n(t) = \frac{1}{n} \sum_{s=0}^{n-1} f_s(t), \quad f_s(t) = \sum_{k=-k}^{k} a_k \exp\{ikt\}.
\]

**Lemma 2.16.** The eigenvalues of optimal circulants \( C_n \) are given by

\[
\lambda_k(C_n) = \sigma_n\left(\frac{2\pi n k}{n}\right), \quad 1 \leq k \leq n,
\]

where \( \sigma_n(t) \) is the \( n \)th Cesaro sum for the Fourier expansion of \( f(t) \).

**Proof.** The claim directly follows from Lemma 2.7. We note only that the eigenvalues of \( C_n \) are equal to the values of \( \langle A_n x, x \rangle \) if \( x \) is any column of the Fourier matrix of order \( n \). □

The extreme case of smoothness is when \( f(t) \) is a trigonometric polynomial

\[
f(t) = \sum_{k=-\mu_1}^{\mu_2} a_k \exp\{ikt\}. \tag{2.16}
\]

**Theorem 2.17.** Given a nonnegative trigonometric polynomial \( f(t) \) of the form (2.16), assume that it is not identically zero. Let it generate Toeplitz matrices \( A_n \) and modified simple circulants \( S_n \). Then the eigenvalues of \( S_n^{-1} A_n \) have a proper cluster at unity, that is,

\[
\gamma_\varepsilon(n) = O(1) \quad \text{for any} \quad \varepsilon > 0.
\]

**Proof.** Assume first that \( S_n \) are positive definite. Consider \( N = \mu_1 + \mu_2 \), with \( \mu_1, \mu_2 \) from (2.16), and for all \( n \geq N + 1 \) define the zero-band matrices

\[
\Delta^{(N)}_n = \begin{bmatrix}
\ddots & \ddots & \ddots \\
\ddots & a_N & \ddots \\
\ddots & \ddots & \ddots \\
a_{-N} & 0 & \ddots \\
\ddots & \ddots & \ddots \\
a_{-1} & \ddots & \ddots \\
a_0 & \ddots & \ddots \\
a_1 & \ddots & \ddots \\
\end{bmatrix}. \tag{2.17}
\]
2.2. Preconditioners and applications

Then for all sufficiently large \( n \) we have

\[
A_n - S_n + \Delta^{(N)}_n = 0.
\]

Consequently, the Hermitian matrix \( S_n^{-1/2}(A_n - S_n)S_n^{-1/2} \) differs from the zero matrix by a matrix whose rank is upper-bounded by \( 2N \). It follows that \( S_n^{-1}A_n - I_n \) may have at most \( 2N \) nonzero eigenvalues.

Now, let us replace \( S_n \) with \( \tilde{S}_n \). Any \( f(t) \) of the form (2.16) coincides with its partial Fourier sums \( f_n(t) \) for all \( n \geq \max(\mu_1, \mu_2) \). By Lemma 2.12, the eigenvalues of \( S_n \) are the values of partial sums. Consequently, the eigenvalues of \( S_n \) are all nonnegative for sufficiently large \( n \), and the number of zero eigenvalues is not greater than the number of zeros of \( f(t) \). This implies that the rank of \( S_n - \tilde{S}_n \) does not exceed the number of zeros of \( f(t) \). Consequently, \( A_n \) differs from \( \tilde{S}_n \) by a matrix, whose rank does not exceed \( 2N \) plus the number of zeros of \( f(t) \).

A particularly renowned example of a trigonometric polynomial is

\[
f(t) = 2 - 2\cos t = 2 - \exp\{-it\} - \exp\{it\}. \tag{2.18}
\]

The corresponding Toeplitz matrices

\[
A_n = \begin{bmatrix}
2 & -1 & & & \\
-1 & 2 & -1 & & \\
& \ddots & \ddots & \ddots & \\
0 & -1 & 2 & -1 & \\
& & 0 & -1 & 2
\end{bmatrix}
\]

arise in the discretization of the one-dimensional Laplace operator. The optimal and simple circulants read

\[
C_n = \begin{bmatrix}
2 & \frac{1-n}{n} & \frac{1-n}{n} & & \\
\frac{1-n}{n} & 2 & \frac{1-n}{n} & & \\
& \ddots & \ddots & \ddots & \\
0 & \frac{1-n}{n} & 2 & \frac{1-n}{n} & \\
& & 0 & \frac{1-n}{n} & 2
\end{bmatrix},
\]

\[
S_n = \begin{bmatrix}
2 & -1 & -1 \\
-1 & 2 & -1 & 0 \\
& \ddots & \ddots & \ddots \\
0 & -1 & 2 & -1 \\
-1 & -1 & 2
\end{bmatrix}.
\]

Obviously, \( S_n \) are singular and of rank \( n - 1 \). We obtain modified simple circulants \( \tilde{S}_n \) by changing only one zero eigenvalue. Theorems 2.15 and 2.17 are clearly illustrated by the behavior of \( \gamma_\epsilon(n) \) shown in Table 2.1.

Another popular example of generating function is

\[
f(t) = 2\pi \left| \sin \frac{t}{2} \right|. \tag{2.19}
\]
Table 2.1. The outlier function $\gamma(n)$ with $\varepsilon = 0.1$ for optimal and simple circulants.

<table>
<thead>
<tr>
<th>$n$</th>
<th>64</th>
<th>128</th>
<th>256</th>
<th>512</th>
<th>1024</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma(n)$ for $C_n$</td>
<td>12</td>
<td>16</td>
<td>26</td>
<td>40</td>
<td>77</td>
</tr>
<tr>
<td>$\gamma(n)$ for $S_n$</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 2.2. The outlier function for the generating function (2.19), $\varepsilon = 0.1$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>64</th>
<th>128</th>
<th>256</th>
<th>512</th>
<th>1024</th>
<th>2048</th>
<th>4096</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma^C(n)$</td>
<td>6</td>
<td>6</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>14</td>
</tr>
<tr>
<td>$\gamma^S(n)$</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

The corresponding Toeplitz matrices

$$A_n = \begin{bmatrix} 1 & \frac{1}{1/(k-l)^2} \\ \frac{1}{1/(k-l)^2} & 1 \end{bmatrix}, \quad 1 \leq k, l \leq n,$$

appear in the works by Lifanov on the discrete vortex method for subsonic flow problems. In this case, $p = 1$ and $m = 1$. According to Theorem 2.15, for the spectra of a matrix preconditioned with optimal and modified simple circulants, we have

$$\gamma^C(n) = O(n^{1/2}), \quad \gamma^S(n) = O(n^{1/2}\ln n).$$

As we can see from Table 2.2, these estimates are rather adequate.

### 2.2.5 Proof of estimates for the outlier function

Given an arbitrary function $f(t) \in K^m$, we consider its Fourier expansion

$$f(t) = \sum_{k=-\infty}^{\infty} a_k \exp(ikt)$$

and the corresponding $n$th partial sums

$$f_n(t) = \sum_{k=-n}^{n} a_k \exp(ikt)$$

and $n$th Cesaro sums

$$\sigma_n(t) = \frac{1}{n} \sum_{k=-n}^{n-1} f_k(t) = \sum_{k=-n}^{n} \left(1 - \frac{|k|}{n}\right) a_k \exp(ikt).$$

**Lemma 2.18.** If $f \in K^m$, then

$$|f(t) - f_n(t)| \leq \varepsilon \frac{1}{n^m}, \quad (2.20)$$

$$|f(t) - \sigma_n(t)| \leq \varepsilon \left\{ \begin{aligned} \frac{1}{n^n}, & \quad m > 1, \\
\frac{1}{n^m}, & \quad m = 1, \end{aligned} \right. \quad (2.21)$$

where $\varepsilon$ does not depend on $n$, and the estimates hold uniformly in $t$. 
2.2. Preconditioners and applications

Proof. From the definition of \( f_n \) and \( \sigma_n \) it follows that

\[
|f(t) - f_n(t)| = \left| \sum_{|k| > n} a_k \exp(ikt) \right| \leq \sum_{|k| > n} |a_k|,
\]

\[
|f(t) - \sigma_n(t)| \leq \frac{1}{n} \sum_{k=\pm n} |k||a_k| + \sum_{|k| > n} |a_k|.
\]

It is an easy matter to check that

\[
a_k = O\left( \frac{1}{k^{m+1}} \right).
\]

Now the estimates (2.20) and (2.21) follow by elementary computations. \( \Box \)

Lemma 2.19. Let \( t_k = t_0 + kh, \ k \in \mathbb{Z} \), be a uniform grid on the real axis with the step-size \( h \), and let \( M \) be a finite union of nonintersecting intervals, with total length equal to \( d \). Then the number of nodes \( t_k \in M \) does not exceed \( c + d/h \), where \( c \) depends on the number of intervals but not on \( h \).

We skip the proof, as the assertion of the lemma is obvious.

Proof of Theorem 2.15. Take an arbitrary, but sufficiently small, \( \delta > 0 \), and define

\[
M(\delta) := \{ t \in [-\pi, \pi]: f(t) \leq \delta \}.
\]

Since \( f(t) \in K_m^p \), the set \( M(\delta) \) can be embedded into a union of nonintersecting intervals with total length

\[
d(\delta) = O(\delta^{1/p}), \quad (2.22)
\]

provided that \( \delta \) is sufficiently small. (This can be readily checked by using the Taylor expansion in zeros \( t_i \).)

The eigenvalues of \( S_n \) are expressed by (2.8). Denote by \( \beta_\delta(n) \) the number of those eigenvalues which are less than or equal to \( \delta \). Let \( \phi_\delta(t) \) be the characteristic function of the closed interval \([-\delta, \delta]\):

\[
\phi_\delta(t) = \begin{cases} 
1, & |t| \leq \delta, \\
0, & |t| > \delta.
\end{cases}
\]

Using Lemma 2.19, we obtain

\[
\frac{\beta_\delta(n)}{n} \leq \frac{1}{n} \sum_{k=0}^{n-1} \phi_\delta \left( f(\frac{2\pi k}{n}) \right)
\]

\[
\leq \frac{1}{n} \sum_{k=0}^{n-1} \phi_{\delta + O(n^{-m})} \left( f(\frac{2\pi k}{n}) \right).
\]

Note that \( \phi_{\delta + O(n^{-m})}(f(\frac{2\pi k}{n})) = 1 \) iff \( \frac{2\pi k}{n} \in M(\delta + O(n^{-m})) \). Therefore, we apply Lemma 2.18 and (2.22) to get

\[
\frac{\beta_\delta(n)}{n} = O \left( (\delta + O(n^{-m}))^{1/p} + \left( \frac{1}{n} \right) \right).
\]
Therefore,

$$\beta_{\delta}(n) \leq c_1 \delta^{1/p} n + c_2 n^{1-m/p} + c_3,$$

where $c_1, c_2, c_3 > 0$ are independent of $n$ and $\delta$.

Assume from now on that $S_n$ is positive definite. For the sake of the proof, we approximate $S_n$ by other circulants using the spectral decomposition of $S_n$. For each eigenvalue, we check whether it is less than $\delta$ and, if so, replace it with $\delta$. In this way we obtain a new circulant $S_n(\delta)$. By construction, it holds that

$$\|S_n^{-1/2}(\delta)\|_\infty \leq \delta^{-1/2}$$

and

$$\text{rank}(S_n - S_n(\delta)) \leq \beta_{\delta}(n).$$

From the spectral decomposition it follows that the estimate for the ranks remains valid if we substitute the circulants with a matrix function of them. In particular,

$$\text{rank}(S_n^{-1/2} - S_n^{-1/2}(\delta)) \leq \beta_{\delta}(n).$$

Consider a matrix $\Delta_n^{(N)}$ of the form (2.17). Due to Lemma 2.18, we get

$$\|A_n - S_n + \Delta_n^{(N)}\|_\infty \leq c \sum_{k \geq N} \frac{1}{k^{m+1}} \leq \frac{c_2}{N^m},$$

where $c_2 > 0$ is independent of $n$ and $N$. Then

$$S_n^{-1/2}(\delta)(A_n - S_n + \Delta_n^{(N)})S_n^{-1/2}(\delta) = S_n^{-1/2}(A_n - S_n)S_n^{-1/2} + \Delta_n^{(N)}S_n^{-1/2}(\delta),$$

where

$$\Delta_n^{(N)}(\delta) = (S_n^{-1/2}(\delta) - S_n^{-1/2})(A_n - S_n)S_n^{-1/2}$$

$$+ S_n^{-1/2}(\delta)(A_n - S_n)(S_n^{-1/2}(\delta) - S_n^{-1/2})$$

$$+ S_n^{-1/2}(\delta)\Delta_n^{(N)}S_n^{-1/2}(\delta).$$

From this identity, estimate (2.24), and the structure of $\Delta_n^{(N)}$, we deduce

$$\text{rank} \Delta_n^{(N)}(\delta) \leq 2(c_1 \delta^{1/p} n + c_2 n^{1-m/p} + c_3 + N).$$

At the same time, estimates (2.23) and (2.25) imply

$$\|S_n^{-1/2}(\delta)(A_n - S_n + \Delta_n^{(N)})S_n^{-1/2}(\delta)\|_\infty \leq \frac{c_2}{\delta N^m}.$$  \hfill (2.27)

Take an arbitrary $\varepsilon > 0$ and make the right-hand side of (2.27) equal to $\varepsilon$. It suffices to choose

$$N := \left\lceil \left( \frac{c_2}{\delta \varepsilon} \right)^{1/m} \right\rceil.$$  \hfill (2.26)

From this choice, we infer from (2.26) that the number of eigenvalues of

$$S_n^{-1/2}A_n S_n^{-1/2} - I$$
outside the \( \epsilon \)-vicinity of zero does not exceed

\[
\left[ 2 \left( c_1 \delta^{1/p} n + c_2 n^{1-m/p} + c_3 \left( \frac{c_0}{\delta} \right)^{1/m} \right) \right].
\]

Let us set

\[
\delta := n^{pm/(p+m)}.
\]

Then, since \( n^{1-m/p} \leq n^{p/(p+m)} \), we obtain

\[
\text{rank } \Delta_n(\delta) = O(n^{p/(p+m)}).
\]

This completes the proof in the case of simple circulants under the assumption that they are positive definite.

Tackling the modified simple circulants needs only a minor change: they differ from the original circulants by rank, and the same estimates as for the rank of \( S_n - S_n(\delta) \) apply. The proof for optimal circulants is accomplished by the same scheme—the difference is only in the norm estimates. So we skip this part and suggest it to the reader as an exercise.

For numerical tests and further discussion of modified preconditioners, see [190].

### 2.2.6 Are there better circulants?

In general, a preconditioner may be viewed as an approximation to a given matrix. However, the quality of this approximation should be understood in a much broader sense rather than as a distance between two matrices in some metric induced by a matrix norm. Suppose we can decompose

\[
A_n = P_n + R_n + E_n,
\]

where \( E_n \) is sufficiently small in some norm and \( R_n \) is of sufficiently small rank; then this indicates that \( P_n \) is a good preconditioner for \( A_n \). In order to be more precise and define a quantitative measure of the quality of \( P_n \), we examine how small the norm and rank can be for circulant preconditioners. In the Hermitian case, if \( \| E_n \|_2 = \epsilon \), then the rank of \( R_n \) is directly linked with the number of the eigenvalues of \( P_n^{-1} A_n \) falling outside the \( \epsilon \)-vicinity of unity. It can be characterized by how fast the outlier function \( \gamma_\epsilon(n) \) defined above grows with respect to \( n \).

Consider several cases where the simple and optimal circulants lead to poor behavior of \( \gamma_\epsilon(n) \), whereas some other circulants lead to better estimates for the outlier function.

**Lemma 2.20.** Assume that \( A_n \) is a lower triangular Toeplitz matrix of order \( n \) with the first column’s elements

\[
a_k = \rho^k, \quad 0 \leq k \leq n, \quad \rho \neq 1.
\]

Then there exists a circulant matrix \( P_n \) such that

\[
A_n = P_n + R_n
\]

with

\[
\text{rank } R_n = 1.
\]

**Proof.** We look for \( R_n \) with the entries

\[
(R_n)_{kl} = c \rho^{k-l}, \quad 0 \leq k, l \leq n - 1.
\]
The rank of such a matrix is 1 (explain why). Then \( P_n := A_n - R_n \) is a circulant iff
\[
\rho^k - c \rho^k = -c \rho^{k-n}, \quad 0 \leq k \leq n - 1,
\]
and this is fulfilled iff
\[
c = \frac{1}{1 - \rho^{-n}}. \quad \square
\]

**Lemma 2.21.** Assume that \( A_n \) is a lower triangular Toeplitz matrix of order \( n \) with the first column’s elements
\[
a_k = k^m, \quad 0 \leq k \leq n - 1,
\]
where \( m \) is a nonnegative integer. Then there exists a circulant matrix \( P_n \) such that
\[
A_n = P_n + R_n,
\]
with
\[
\text{rank } R_n \leq m + 2.
\]

**Proof.** Begin with \( m = 0 \) and introduce a first-degree polynomial
\[
\psi(x) = \frac{1}{n} x.
\]
It is easy to check that
\[
\psi(x) - \psi(x - n) = 1 \quad \forall \ x.
\]
We define \( R_n \) by
\[
(R_n)_{kl} = \psi(k - l), \quad 0 \leq k, l \leq n - 1.
\]
For the columns of \( R_n \) it holds that \( R_n(:, l) = R_n(:, 1) + \frac{l - 1}{n} (1, \ldots, 1)^T \). From this one easily finds \( \text{rank } R_n \leq 2 \) (show this). Since
\[
\psi(k) - \psi(k - n) = 1, \quad k = 0, 1, \ldots, n - 1,
\]
the matrix \( T_n - R_n \) is a circulant.

Now, assume by induction that we already constructed a polynomial \( p_m(x) \) with the properties
\[
p_m(x) - p_m(x - n) = x^m \quad \forall \ x, \quad \deg p_m \leq m + 1.
\]
Then \( R_n \), with the entries \( (R_n)_{kl} := p_m(k - l) \), is of rank not greater than \( m + 2 \), and the matrix \( P_n := A_n - R_n \) is a circulant.

Let \( \phi(x) = \int_0^x p_m(t) \, dt \). Then
\[
\phi(x) - \phi(x - n) + \phi(-n) = \frac{x^{m+1}}{m+1},
\]
and the polynomials
\[
p_{m+1}(x) := (m + 1)\phi(x) - (m + 1)\phi(-n)\phi(x)
\]
enjoy the properties
\[
p_{m+1}(x) - p_{m+1}(x - n) = x^{m+1} \quad \forall \ x, \quad \deg p_{m+1} \leq m + 2.
\]
This completes the proof of the induction step. \( \square \)
2.2. Preconditioners and applications

As one further illustration, consider a Toeplitz matrix of the form

\[ A_n = \begin{bmatrix} 1 & \frac{1}{n} & \cdots & \frac{n-1}{n} \\ \vdots & \frac{1}{n} & \ddots & \vdots \\ \frac{1-n}{n} & \frac{1}{n} & \ddots & \frac{1}{n} \\ \frac{1-n}{n} & \frac{1}{n} & \cdots & 1 \end{bmatrix}. \]

Neither simple nor optimal circulants are close to \( A_n \) in the sense of the splitting (2.28). However, this does not mean that there is no good circulant preconditioner for \( A_n \). According to the above proof, set

\[ P_n := \begin{bmatrix} 1 & \frac{1}{n} & \cdots & \frac{n-1}{n} \\ \vdots & \frac{1}{n} & \ddots & \vdots \\ \frac{1-n}{n} & \frac{1}{n} & \ddots & \frac{1}{n} \\ \frac{1-n}{n} & \frac{1}{n} & \cdots & 1 \end{bmatrix}. \]

Then \( P_n \) is a circulant such that

\[ A_n - P_n = [(k - l)/n], \quad 0 \leq k, l \leq n - 1. \]

Hence,

\[ A_n - P_n = \begin{bmatrix} 0 \\ 1/n \\ \vdots \\ (n-1)/n \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 1/n \\ \vdots \\ (n-1)/n \end{bmatrix}, \]

which evidently implies that \( A_n \) can be decomposed as \( A_n = P_n + R_n + E_n \), with

\[ \text{rank } R_n = 2, \quad E_n = 0. \]

**Lemma 2.22.** Assume that \( A_n \) is a lower triangular Toeplitz matrix, with the first column’s elements

\[ a_k = \begin{cases} 0, & k = 0, \\ \rho^k k^{-\alpha}, & 1 \leq k \leq n-1, \end{cases} \]

where \( \alpha > 0 \). Then for any \( \varepsilon > 0 \), there exists a circulant matrix \( P_n \) such that

\[ A_n = P_n + R_n + E_n, \]

where

\[ |(E_{n})_{kl}| \leq |(A_{n})_{kl}| \varepsilon, \quad 0 \leq k, l \leq n - 1, \]

\[ \text{rank } R_n \leq \ln \varepsilon^{-1} (c_0 + c_1 \ln \varepsilon^{-1} + c_2 \ln n). \]

**Proof.** The claim emanates from a well-known fact (cf. [223]) about the approximation of the function \( f(x) = x^{-\alpha} \) by a finite sum of exponents: There exist frequencies \( f_i \) and weights \( w_i \) such that

\[ k^{-\alpha} - \sum_{l=1}^{r} w_i \exp(-f_i k) \leq k^{-\alpha} \varepsilon \]
with

\[ r \leq \ln \varepsilon^{-1} \left( c_0 + c_1 \ln \varepsilon^{-1} + c_2 \ln n \right). \]

We use this fact and apply Lemmas 2.20 and 2.21 to complete the proof. \(\square\)

**Theorem 2.23.** Assume that a Toeplitz matrix is generated by the formal Fourier expansion of a piecewise analytic function of the form

\[ f(z) = g(z) + \sum_{i=0}^{\infty} \sum_{t=0}^{\infty} A_{st}(z - \zeta_s)^t \log(z - \zeta_s), \]

where

\[ z = \exp\{it\}, \quad |\zeta_s| = 1, \]

and the function \(g(z)\) is analytic in a ring around \(|z| = 1\). Then for any \(\varepsilon > 0\) there exists a circulant matrix \(P_n\) and a splitting

\[ A_n = P_n + R_n + E_n, \]

with

\[ |(E_n)_{kl}| \leq |(A_n)_{kl}| \varepsilon, \quad 0 \leq k, l \leq n - 1, \]

\[ \text{rank } R_n \leq \ln \varepsilon^{-1} \left( c_0 + c_1 \ln \varepsilon^{-1} + c_2 \ln n \right) + c_3, \]

and the constants \(c_0, c_1, c_2, c_3\) independent of \(n\) and \(\varepsilon\).

The proof of the theorem follows from Lemmas 2.20, 2.21, and 2.22; cf. [227].

Simple and optimal circulants are still attractive because they are constructed only from the elements of a given matrix, while the existence of better circulants was proved under some assumptions on the generating function. One may ask: Can we still construct those better circulants using only the entries of a given matrix?

Although this is not as trivial as constructing simple and optimal circulants, fortunately we can do this. To present the idea, we consider the splitting (2.28), with \(E_n = 0\), i.e., \(A_n = P_n + R_n\), where \(P_n\) is a circulant and \(\text{rank } R_n \leq r\). We assume that such a splitting exists and, based on this knowledge, we need to find \(P_n\). To this end, introduce

\[ \hat{A}_n := F_n A_n F_n^{-1}, \quad D_n := F_n P_n F_n^{-1}, \]

where \(F_n\) is the Fourier matrix of order \(n\). Then the problem of finding \(P_n\) reduces to the following: Given \(\hat{A}_n\), find a diagonal matrix \(D_n\) such that

\[ \text{rank} (\hat{A}_n - D_n) \leq r. \]

Assume that \(r = 1\) and \(\hat{R}_n := \hat{A}_n - D_n \neq 0\). Using the MATLAB notation, let \(\hat{R}_n(i,:)\) and \(\hat{R}_n(:,j)\) denote the \(i\)th row and \(j\)th column of \(\hat{R}_n\), respectively. If this row and column are nonzero, then it also holds that

\[ \hat{R}_n(i,j) = (\hat{R}_n)_{ij} \neq 0 \quad (\text{since } \text{rank}(\hat{R}_n) = 1) \]

and \(R_n\) is given by a rank-1 skeleton (dyadic) decomposition

\[ \hat{R}_n = \frac{1}{R_n(i,j)} \hat{R}_n(i,:) \hat{R}_n(:,j). \tag{2.29} \]

Given \(\hat{A}_n\), we know all the entries in \(\hat{R}_n(i,:)\) and \(\hat{R}_n(:,j)\) except for \(\hat{R}_n(i,i)\) and \(\hat{R}_n(j,j)\).
2.2. Preconditioners and applications

Suppose that \( i \neq j \). Then (2.29) reveals all the entries of \( \hat{R}_n \) except for \( \hat{R}_n(i,i) \) and \( \hat{R}_n(j,j) \). Thus, \( D_n \) is almost recovered: Only two diagonal entries are not found. If \( \hat{R}_n(k,k) \neq 0 \) for some \( k \neq i,j \), then

\[
\hat{R}_n = \frac{1}{\hat{R}_n(k,k)} \hat{R}_n(k,:) \hat{R}_n(:,k),
\]

This formula recovers all the entries of \( \hat{R}_n \). Hence, the diagonal matrix \( D_n \) is found and allows us to get the circulant \( P_n \). The case \( r > 1 \) is treated similarly. The only difference is that we leave unrecovered more than two diagonal entries.

The above computations can be done quickly due to a special representation of the off-diagonal entries of the Fourier image \( \hat{A}_n \).

**Lemma 2.24.** Assume \( A_n \) is designed by generators \( \{a_{1-n}, \ldots, a_{n-1}\} \), and let

\[
\begin{bmatrix}
v_0 \\
v_1 \\
\vdots \\
v_{n-1}
\end{bmatrix} = F_n \begin{bmatrix}
an_{1-n} - a_1 \\
\vdots \\
an_1 - a_{n-1}
\end{bmatrix}.
\]

Then the off-diagonal entries of \( \hat{A}_n \) are given by the formula

\[
(\hat{A}_n)_{kl} = \frac{v_k - v_l}{n(e^{k-l} - 1)}, \quad 0 \leq k, l \leq n - 1, \quad k \neq l,
\]

where

\[
\varepsilon = \exp\left\{ \frac{2\pi}{n} \right\}.
\]

**Proof.** By definition, \( \hat{A}_n = \frac{1}{n} F_n A_n F_n^* \). In the elementwise form it reads

\[
(\hat{A}_n)_{kl} = \frac{1}{n} \sum_{s=0}^{n-1} \sum_{\beta=0}^{n-1} \varepsilon^{s \alpha_s - \beta} \varepsilon^{-s \beta} = \frac{1}{n} \sum_{s=0}^{n-1} \sum_{\beta=0}^{n-1} \varepsilon^{(s-\beta) \alpha_s} \varepsilon^{(s-\beta)(-\beta)}.
\]

Set \( s = \alpha - \beta \) and rearrange the above summation as follows:

\[
(\hat{A}_n)_{kl} = \frac{1}{n} \sum_{s=1}^{n-1} \left( \varepsilon^{-s \alpha_s} \sum_{\beta=0}^{n-1} \varepsilon^{(k-l)\beta} + \varepsilon^{s \alpha_s} \sum_{\beta=0}^{n-1} \varepsilon^{-(k-l)\beta} \right)
\]

\[
= \frac{1}{n} \sum_{s=1}^{n-1} \left( \varepsilon^{-s \alpha_s} \frac{1 - \varepsilon^{(k-l)s}}{\varepsilon^{k-l} - 1} + \varepsilon^{s \alpha_s} \frac{\varepsilon^{(k-l)(n-s)} - 1}{\varepsilon^{k-l} - 1} \right)
\]

\[
= \frac{1}{n(e^{k-l} - 1)} \sum_{s=1}^{n-1} \varepsilon^{s \alpha_s} (a_{n-s} - a_s)(1 - \varepsilon^{-(k-l)s})
\]

\[
= \frac{1}{n(e^{k-l} - 1)} \left( \sum_{s=1}^{n-1} \varepsilon^{s \alpha_s} (a_{n-s} - a_s) - \sum_{s=1}^{n-1} \varepsilon^{s \alpha_s} (a_{n-s} - a_s) \right).
\]
Corollary 2.25. If the quantities $v_k$ are computed, then any individual entry of the Fourier image $\hat{A}_n$ can be computed in a finite number of operations that does not depend on $n$.

For more details and numerical results for typical test matrices, we refer the reader to [161]. Recently, a similar technique was used for the construction of new matrix algebra preconditioners [197]. Other useful types of circulant preconditioners are considered in [57, 115, 164].

2.2.7 Multilevel Toeplitz matrices and multilevel circulants

Matrices arising in multidimensional problems may naturally possess a multilevel structure. Suppose $A = [a(i, j)]$ is a matrix of order

$$n = n_1 \ldots n_p.$$ 

Then the row and column indices can be put in one-to-one correspondence with multi-indices

$$i \leftrightarrow \beta := (i_1, \ldots, i_p), \quad j \leftrightarrow \alpha := (j_1, \ldots, j_p)$$

by the rule

$$i = (i_1 - 1)n_2 \ldots n_p + \cdots + (i_{p-1} - 1)n_p + i_p,$$

$$j = (j_1 - 1)n_2 \ldots n_p + \cdots + (j_{p-1} - 1)n_p + j_p,$$

where

$$1 \leq i_1, j_1 \leq n_1, \ldots, 1 \leq i_p, j_p \leq n_p.$$ 

We find it convenient to omit commas and brackets in multi-indices and write

$$\beta = i_1 \ldots i_p, \quad \alpha = j_1 \ldots j_p.$$ 

The usage of this notation should be clear from the context and should not be confused with a product of numbers. With this notation, we can write

$$a(i, j) = a(i_1 \ldots i_p, j_1 \ldots j_p).$$

We can naturally consider $A$ as a block matrix with a hierarchy of blocks. Let us set

$$A(i_1, j_1) := [a(i_1 \ldots i_p, j_1 \ldots j_p)].$$

Then $A$ can be viewed as a block matrix of block size $n_1 \times n_1$ with the blocks $A(i_1, j_1)$, each one being of size $(n_2 \ldots n_p) \times (n_2 \ldots n_p)$. We say that this block partitioning of $A$ is the 1st level partitioning and $A(i_1, j_1)$ are the 1st level blocks. Similarly, $A$ consists of the blocks

$$A(i_1 i_2, j_1 j_2) := [a(i_1 \ldots i_p, j_1 \ldots j_p)].$$

We say that these blocks belong to the 2nd level. In the same manner, the blocks

$$A(i_1 \ldots i_k, j_1 \ldots j_k) := [a(i_1 \ldots i_p, j_1 \ldots j_p)]$$

constitute the $k$th level of $A$. Each block of the level $k - 1$ is a block matrix composed of smaller blocks of the level $k$. When a matrix $A$ is endowed with such a structure, we
call $A$ a **multilevel matrix** or, more precisely, a **$p$-level matrix**, provided that the $p$-level blocks are considered as matrices with scalar entries.

Block matrices of each level may possess some structure. If on every level each block is a block Toeplitz matrix, then $A$ is called a **multilevel Toeplitz matrix**. In this case $a(i_1 \ldots i_p, j_1 \ldots j_p)$ depends only on the differences $i_k - j_k$, that is,

$$a(i_1 \ldots i_p, j_1 \ldots j_p) = F(i_1 - j_1, \ldots, i_p - j_p).$$

If on every level each block is a block circulant matrix, then $A$ is called a **multilevel circulant matrix**. In this case, $a(i_1 \ldots i_p, j_1 \ldots j_p)$ depends only on the differences $i_k - j_k$ modulo $n_k$, that is,

$$a(i_1 \ldots i_p, j_1 \ldots j_p) = F((i_1 - j_1)(\mod n_1), \ldots, (i_p - j_p)(\mod n_p)).$$

For the sake of brevity, one can introduce componentwise operations on multi-indices and write

$$a(\beta, \alpha) = F(\beta - \alpha)$$

for a multilevel Toeplitz matrix and

$$a(\beta, \alpha) = F((\beta - \alpha)(\mod n)), \quad n := (n_1, \ldots, n_p)$$

for a multilevel circulant matrix.

Multilevel circulants are used as preconditioners for multilevel Toeplitz matrices. We define the simple and optimal multilevel circulant preconditioners similarly to the one-level case. Using the link with multivariate Fourier series, we can prove certain results about clusters in the spectra of preconditioned matrices. These results are similar to those of the one-level case. However, in the study of asymptotic behavior of the outlier function we should assume that $n = n_1 \ldots n_p \to \infty$ is such that

$$n_1 \to \infty, \ldots, n_p \to \infty$$

happens simultaneously. This is a special flavor of the multilevel case and somewhat limits what we can achieve by the presented methods of analysis.

Consider a formal Fourier series in $p$ variables,

$$f(t_1, \ldots, t_p) = \sum_{k_1 = -\infty}^{\infty} \cdots \sum_{k_p = -\infty}^{\infty} a(k_1, \ldots, k_p) \exp\{i(k_1 t_1 + \cdots + k_p t_p)\},$$

where

$$i^2 = 1.$$

We say that $f$ belongs to the Wiener class with bound $c$ if

$$\sum_{k_1 = -\infty}^{\infty} \cdots \sum_{k_p = -\infty}^{\infty} |a(k_1, \ldots, k_p)| \leq c < +\infty.$$

We associate with $f$ a $p$-level Toeplitz matrix

$$A_n = [a(i_1 - j_1, \ldots, i_p - j_p)]$$
and $p$-level optimal circulants $C_n$ and $p$-level simple circulants $S_n$. The optimal ones minimize the Frobenius norm $\|A_n-P_n\|_F$ over all $p$-level circulants $P_n$. The simple circulants are constructed blockwise: on the 1st level we find the simple block circulant using a direct extension of the one-level definition; on the 2nd level we find a simple block circulant for each block of the 1st level, and so on. Here is an illustration for a two-level Toeplitz symmetric matrix:

$$
A_{(3,3)} = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\
2 & 1 & 2 & 0 & 4 & 5 & 0 & 7 & 8 \\
3 & 1 & 1 & 0 & 0 & 4 & 0 & 0 & 7 \\
4 & 0 & 0 & 1 & 2 & 3 & 4 & 5 & 6 \\
5 & 4 & 0 & 2 & 1 & 2 & 0 & 4 & 5 \\
6 & 5 & 4 & 3 & 2 & 1 & 0 & 0 & 4 \\
7 & 0 & 0 & 4 & 0 & 0 & 1 & 2 & 3 \\
8 & 7 & 0 & 5 & 4 & 0 & 2 & 1 & 2 \\
9 & 8 & 7 & 6 & 5 & 4 & 3 & 2 & 1
\end{bmatrix},
$$

$$
S_{(3,3)} = \begin{bmatrix}
1 & 2 & 2 & 4 & 5 & 0 & 4 & 0 & 5 \\
2 & 1 & 2 & 0 & 4 & 5 & 0 & 4 & 0 \\
2 & 1 & 1 & 5 & 0 & 4 & 0 & 5 & 4 \\
4 & 0 & 5 & 1 & 2 & 2 & 4 & 5 & 0 \\
5 & 4 & 0 & 2 & 1 & 2 & 0 & 4 & 5 \\
0 & 5 & 4 & 2 & 2 & 1 & 5 & 0 & 4 \\
4 & 5 & 0 & 4 & 0 & 5 & 1 & 2 & 2 \\
0 & 4 & 5 & 5 & 4 & 0 & 2 & 1 & 2 \\
5 & 0 & 4 & 0 & 5 & 4 & 2 & 2 & 1
\end{bmatrix},
$$

**Theorem 2.26.** Assume that $f(t_1, \ldots, t_p)$ is given by its formal Fourier series, belongs to the Wiener class, and is bounded from below by a positive constant uniformly in $t_1, \ldots, t_p$. Then $A_n$ and the optimal $p$-level circulants $C_n$ are Hermitian positive definite matrices, for all sufficiently large $n_1, \ldots, n_p$, the simple circulant $S_n$ are Hermitian positive definite, and the eigenvalues of $S_n^{-1}A_n$ and $C_n^{-1}A_n$ have a general cluster at unity.

We skip the proof, as it repeats the steps of the one-level proof, although it is somewhat more involved due to the multilevel notation.

It should be emphasized that in the multilevel case we prove existence of a general cluster and we are not able to show that the cluster is proper. Moreover, as a rule, it is not a proper cluster. Even if the outlier function grows quite moderately, in the multilevel case it must grow in $n$. This is the kind of negative result proved in [179, 180]. Preconditioning techniques for some particular cases of two-level matrices are considered in [107].

### 2.2.8 Applications and remarks

As a first example, consider the equation of Love from electrostatics (see Dellwo [65]):

$$
u(x) = \frac{1}{\pi} \int_{-1}^{1} \frac{d}{d^2 + (x-t)^2} u(t) \, dt = 1.$$


To solve it, we can introduce a uniform grid \( x_i := b(i - 1/2) \) with the step-size \( b = 2/n \) and reduce the problem to a linear algebraic system of the form

\[
\sum_{j=1}^{n} a_{ij} u_j = 1, \quad 1 \leq i \leq n,
\]

where

\[
a_{ij} = \delta_{i-j} - \frac{db}{\pi(d^2 + (x_i - x_j)^2)} = \delta_{i-j} - \frac{db}{\pi(d^2 + (i - j)^2 b^2)}, \quad 1 \leq i, j \leq n,
\]

\[
\delta_k = \begin{cases} 1 & \text{if } k = 0, \\ 0 & \text{otherwise}. \end{cases}
\]

Obviously, the coefficient matrix for this system is a Toeplitz matrix.

This example may appear to be too simple to demonstrate the full power of iterative solvers for Toeplitz matrices. However, it clearly shows how this kind of structured matrix can arise in the numerical solution of integral equations.

Let us describe a more general situation. Consider a \( d \)-dimensional domain \( \Omega \) and an integral equation

\[
u(x) + \int_{\Omega} G(x, y)u(y)dy = f(x), \quad x \in \Omega,
\]

where the kernel \( G(x, y) \) is invariant under some transformation \( R \):

\[
G(Rx, Ry) = G(x, y), \quad x, y \in \mathbb{R}^d.
\]

Assume also that \( \Omega \) is a union of disjoint domains obtained from one of them via the transformation \( R \):

\[
\Omega = \bigcup_{1 \leq j \leq n} \Omega_j, \quad \Omega_{j+1} = R\Omega_j, \quad 1 \leq j \leq n - 1.
\]

Suppose that an approximate solution is sought in the form

\[
u(x) = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} \phi_{ij}(x),
\]

where \( \phi_{ij}(x) \) is supported on \( \Omega_j \) and partially invariant under the same transformation \( R \):

\[
\phi_{i,j+1}(Rx) = \phi_{ij}(x), \quad x \in \Omega_j, \quad 1 \leq i \leq m, \quad 1 \leq j \leq n - 1.
\]

Upon application of the Galerkin method with the FORTRAN-like numbering of the unknowns \( a_{ij} \), we arrive at a linear algebraic system with a block Toeplitz matrix of coefficients. All blocks are of size \( m \times m \) and may possess no structure, but any block diagonal parallel to the main one consists of the same blocks.

In many situations of practical interest there could be more than one transformation of invariance. In such cases the blocks as well possess some structure, e.g., could be also block Toeplitz matrices and, if there are still more transformations, we can eventually obtain multilevel Toeplitz matrices.
Integral equations of the classical potential theory have very special kernels. As a typical example, consider Symm’s equation

\[-\frac{1}{2\pi} \int_{\partial \Omega} \ln|x-y|u(y)ds = f(x), \quad x \in \partial \Omega.\]

This is a single-layer potential equation related to the Dirichlet boundary value problems for the Laplace equation in two dimensions. Formally Symm’s equation is an integral equation of the first kind, and the well-posedness crucially depends on the choice of functional classes for the solution and the right-hand side. If the contour \(\Omega\) is a circle, then the transformation given by the left-hand side integral is a rotation over the center of this circle. This case is especially simple, as we directly obtain a linear algebraic system with a circulant matrix of coefficients. In the general case, the matrices are neither circulant nor Toeplitz. However, circulants generally serve as good preconditioners. Without giving a rigorous proof, we can explain it using the fact of spectral equivalence of matrices for different smooth contours [203].

Important practical examples are the following equations arising in the modeling of subsonic flow around a thin airfoil given as an open contour:

\[
\int_{-1}^{1} u(y) \frac{1}{x-y} dy + \int_{-1}^{1} G_1(x,y)u(y)dy = f_1(x),
\]

(2.30)

\[
\int_{-1}^{1} u(y) \frac{1}{(x-y)^2} dy + \int_{-1}^{1} G_2(x,y)u(y)dy = f_2(x).
\]

(2.31)

Equation (2.30) has the singularity of Cauchy type, and the first integral term in (2.31) is understood in the sense of Hadamard. Simple discretizations for these equations are provided by the discrete vortex method [132], where one uniform grid is used for the values of the kernel function and another grid is shifted by the half step-size and used for the collocation points. In this method, and for some more advanced projection methods as well, the first term in both equations yields a Toeplitz matrix. Circulant preconditioners for both cases are the perfect choice.

An airfoil in three dimensions is frequently modeled as an open smooth surface \(\Omega\), and the equation for the vortex density function reads

\[
\frac{1}{4\pi} \int_{\Omega} \frac{\partial}{\partial n_x} \frac{\partial}{\partial n_y} \left(\frac{1}{|x-y|}\right) u(y)ds(y) = f(y), \quad x \in \Omega,
\]

where \(n_x\) and \(n_y\) are the normals to \(\Omega\) at the corresponding points. If \(\Omega\) is flat, e.g., a square on the plane, then this equation acquires the form

\[
\frac{1}{4\pi} \int_{\Omega} \frac{u(y_1,y_2)}{(x_1-y_1)^2 + (x_2-y_2)^2}dy_1dy_2 = f(x_1,x_2).
\]

Again, discretization on uniform grids results in a Toeplitz matrix. In the general case, Toeplitz matrices can be used as efficient preconditioners.

Toeplitz structure is observed and widely used also in more complicated cases, e.g., when solving the volume integral equation in electrodynamics. We skip the exact formulation, as it is beyond the scope of this book. In rather abstract terms the equation can be

\footnote{For Symm’s equation, the standard choice consists of the Sobolev spaces \(H^{-1/2}\) and \(H^{1/2}\) for the solution and right-hand side, respectively.}
written as follows:

\[
\gamma(x)u(x) + \int_\Omega G(|x-y|)u(y) d\omega(y) = f(x), \quad \Omega = [0,1]^3.
\]

A discretization may lead to a matrix which is the sum of a diagonal matrix and a three-level Toeplitz matrix. Note also that the claim depends on an actual discretization method applied. In any case, we obtain a sum of three-level matrices where the first is highly sparse and the second is Toeplitz at every level. The structure of the matrix provides us with a fast matrix-vector multiplication algorithm, which is important for iterative solvers. However, a choice of good preconditioner for such problems is still a challenge.

Multilevel Toeplitz matrices are prolific in the works of imaging sciences; cf. [99, 59]. Toeplitz structure emerges as well in the solution of differential equations discretized on uniform grids; cf. [52, 55, 108].

Closing this chapter, we emphasize that there are still a lot of open questions in the design of efficient iterative solvers for multilevel Toeplitz matrices. Most likely, one has to take into account other kinds of structures, which should be combined with the Toeplitz structure. Tensor-related and block low-rank structures are considered in [159, 160, 111].

## Exercises

2.2.1. Prove that the complexity of computing $A^{-1}$ for a nonsingular $p$-level circulant $A$ of order $n$ is $O(n \log_2 n)$.

2.2.2. Let all the eigenvalues of a Hermitian matrix $A$ belong to the interval $[a, b]$. Consider a matrix $B = A + R$, where $R = R^*$ and $\text{rank} R \leq r$. Prove that $B$ has at most $r$ eigenvalues outside the interval $[a, b]$.

2.2.3. Consider a sequence of Toeplitz matrices $A_n = [a_{i-j}]_{n \times n}$ and a corresponding sequence of optimal circulants $C_n = [c_{i-j}]_{n \times n}$. Assume $\sum_{k=-\infty}^{\infty} |a_k|^2 < +\infty$. Prove that $||A_n - C_n||_2^2 = o(n)$ holds. Furthermore, assuming $||C_n^{-1}||_2 ||A_n - C_n||_F = o(\sqrt{n})$, prove that the singular values of matrices $C_n^{-1}A_n$ have a cluster in $1$.

2.2.4. Consider a sequence of Toeplitz matrices $A_n = [a_{i-j}]_{n \times n}$ and a corresponding sequence of optimal circulants $C_n = [c_{i-j}]_{n \times n}$. Prove that the eigenvalues of matrices $C_n$ are the partial Cesaro sums:

$$
\lambda_k(C_n) = \sigma_n \left( \frac{2\pi}{n} \right), \quad 0 \leq k \leq n; \quad \sigma_n(t) = \sum_{k=0}^{n} \left( 1 - \frac{k}{n} \right) a_k e^{ikt}.
$$
Chapter 3  
Multigrid Preconditioners

3.1 The introductory section

In this chapter, we introduce an optimal, efficient, and widely used approach to accelerate convergence of Krylov subspace methods by preconditioning: multigrid iterations or multigrid methods. Although this approach applies to matrices originating from a variety of mathematical models and applications, the multigrid method was originally introduced, extensively developed, and shown to be especially efficient for solving algebraic problems resulting from the discretization of partial differential equations. The superior performance of multigrid (MG) preconditioners owes much to properties of differential problems and discretization methods such as stability and accuracy. We will need these properties to understand and analyze multigrid methods. Therefore, we start the chapter with an elementary introduction to finite element methods, which is the discretization method of choice for most of this book. Several, more advanced finite element methods and their properties will be given in those places of the text where they are used.

In this introductory section, we recall the elementary properties of the finite element method. As already mentioned, the finite element method will be our main tool for approximating a differential problem by a system of algebraic equations. We start with formulation of a differential problem and some estimates for its solution. We proceed with the reduction of the differential problem to a computable discrete problem, and further we recall estimates on the difference between solutions of these two problems. In this section, we introduce necessary definitions and results, using a simple one-dimensional model problem as an example. The foundations of the finite element method in full generality and further details can be found in many monographs; see, e.g., Ciarlet [61] or Braess [31].

A multigrid method first appeared and was studied in the pioneering papers of Fedorenko [76, 77] and Bakhvalov [12] in the 1960s. However, it did not draw enough attention by that time. Only starting with the paper of Brandt [42] the method was widely acknowledged and further developed by many mathematicians and engineers. The modern theory of the method was largely founded in the 1980s and 1990s. The book of Hackbusch [96] was one of the first on the subject. In that influential work, an abstract mathematical framework was developed and many applications of multigrids were described. The theory and practice of multigrid methods continues to be extended and developed into the present day.
Chapter 3. Multigrid Preconditioners

3.1.1 Preliminaries

Assume $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$, is a bounded domain with piecewise smooth boundary, satisfying the Lipschitz condition. We shall use the following standard functional spaces and notation.

The space $L^2 = L^2(\Omega)$ is the Lebesgue space of all integrable function $v(x)$ on $(\Omega)$, with the finite integral $\int_{\Omega} v^2(x) \, dx$; the integral $(u, v) := \int_{\Omega} u(x) v(x) \, dx$ defines the inner product for $L^2(\Omega)$; $C^k = C^k(\Omega)$ is the space of $k$ time continuously differentiable functions on $\Omega$; $H^k = H^k(\Omega)$ is a subspace of $L^2(\Omega)$ of functions such that all mixed derivatives\(^7\) of orders $0, \ldots, k$ are from $L^2(\Omega)$; and $H^0_0 = H^0_0(\Omega)$ is the subspace of functions from $H^1(\Omega)$ vanishing on boundary of $\Omega$. The inner product in $H^1(\Omega)$ (and in $H^1_0(\Omega)$) is defined by

$$(u, v)_{H^1} := \int_{\Omega} u v + \sum_{i=1}^d \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i} \, dx.$$ 

The Lebesgue space $L^2$ and the Sobolev spaces $H^k$ and $H^1_0$ are Hilbert spaces.

For functions from $H^1_0(\Omega)$, the Friedrichs inequality holds:

$$\left( \int_{\Omega} u^2 \, dx \right)^{\frac{1}{2}} \leq C_{\text{F}}(\Omega) \left( \int_{\Omega} \sum_{i=1}^d \left( \frac{\partial u}{\partial x_i} \right)^2 \, dx \right)^{\frac{1}{2}},$$

(3.1)

where a constant $C_{\text{F}}(\Omega)$ is independent of $u$. With the help of the Friedrichs inequality, one concludes that

$$(u, v)_t := \int_{\Omega} \sum_{i=1}^d \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i} \, dx$$

(3.2)

defines an inner product in $H^1_0(\Omega)$, and this inner product defines a norm $\| \cdot \|_1$ on $H^1_0(\Omega)$ equivalent to the $H^1$-norm.

After possible redefinition on a set of measure zero, any function from $H^1(\Omega)$ is continuous on $\Omega$, and it holds that $\sup_{x \in \Omega} |u| \leq C \| u \|_{H^1(\Omega)}$, with a constant $C$ depending only on $\Omega$. More on the properties of Sobolev spaces can be found in [1, 184].

3.1.2 A two-point boundary value problem

We start with the simple example of the one-dimension diffusion equation:

$$\begin{align*}
- (k(x) u')' &= f & \text{on } (0,1), \\
u(0) &= u(1) = 0.
\end{align*}$$

(3.3)

The diffusion coefficient $k(x)$ is assumed to be piecewise smooth, and $1 \leq k(x) \leq K$. We multiply the first equation in (3.3) by an arbitrary function $v(x) \in H^1_0(0,1)$ and integrate the resulting relation by parts on the interval $[0, 1]$. Thanks to the homogeneous boundary conditions, we obtain that a solution $u(x)$ to (3.3), if it exists, satisfies

$$\int_0^1 k(x) u' v' \, dx = \int_0^1 f v \, dx \quad \forall \, v(x) \in H^1_0(0,1).$$

(3.4)

\(^7\text{Strictly speaking, a special extension of the classical notion of derivative is used. One can find the definition of generalized derivative and a discussion in [122].}\)
A function $u$ from $H^1_0(0,1)$, satisfying (3.4), is called the weak solution to the two-point boundary value problem (3.3). If the integral on the right-hand side of (3.4) defines a linear bounded functional on $H^1_0(0,1)$ as happens for $f \in L^2$, then the solution to (3.4) exists and is unique.\footnote{The result follows from the Lax–Milgram lemma: Assume $a(u, v)$ is a bilinear continuous form defined on a Hilbert space $H$ such that $a(u, v) \geq \|v\|^2$ for any element from $H$, and $f(v)$ is a linear bounded functional. Then there exists the unique $u \in H$ satisfying equality $a(u, v) = f(v)$ for all $v \in H$.} Moreover, if the solution $u(x)$ to (3.4) and the coefficient $k(x)$ are smooth functions (e.g., $u \in C^2(0,1)$, $k \in C^1(0,1)$), it can be shown that the weak solution $u(x)$ also satisfies (3.3) in the usual pointwise sense.

Let $f(x) \in L^2(0,1)$. Setting $v = u$ in (3.4) and applying the Cauchy and the Friedrichs inequalities leads to the stability estimate (check!)

$$\|u\|_{H^1} = \left(\int_0^1 u'^2 + (u')^2 \, dx \right)^\frac{1}{2} \leq c \left(\int_0^1 f^2 \, dx \right)^\frac{1}{2} = c \|f\|_{L^2},$$

for $u$ solving (3.4). Additionally assume that $k(x) \in C^1(0,1)$; then $u \in H^2(0,1)$ and we can estimate the $L^2$-norm of the solution second derivative:

$$\|u''\|_{L^2} = \left(\int_0^1 (u'')^2 \, dx \right)^\frac{1}{2} \leq c \left(\int_0^1 f^2 \, dx \right)^\frac{1}{2} = c \|f\|_{L^2}.$$  \hspace{1cm} (3.6)

The estimates (3.5) and (3.6) are the basic \textit{a priori}, i.e., known to hold without any further information about $u(x)$, estimates for the solution of (3.3). Note that to show (3.6) we need the following additional assumption on the regularity (smoothness) of the given data: $k(x) \in C^1(0,1)$.

### 3.1.3 Finite element method: A two-point boundary value problem

To define an approximate solution to (3.3), consider a finite element method. The weak formulation (3.4) of the problem is the departure point for this discretization technique.

First, we choose a subspace $\mathbb{U}_h$ of a \textit{finite dimension} approximating the space $H^1_0(0,1)$ and such that $\mathbb{U}_h \subset H^1_0(0,1)$. What is meant by “$\mathbb{U}_h$ approximates $H^1_0(0,1)$” will become clearer later. As a simple example, assume that $\mathbb{U}_h$ consists of all piecewise linear continuous functions on $[0,1]$ with respect to the partition of the interval $[0,1]$ into $N$ equal intervals of the length $h = 1/N$. Assume also $v_h(0) = v_h(1) = 0$ for all functions from $\mathbb{U}_h$.

The finite element solution to (3.3) is a function $u_h$ from $\mathbb{U}_h$, satisfying the equality

$$\int_0^1 k(x)u_h'v_h' \, dx = \int_0^1 f v_h \, dx \quad \forall v_h(x) \in \mathbb{U}_h.$$  \hspace{1cm} (3.7)

### 3.1.4 Finite element method: Systems of algebraic equations

Since $\mathbb{U}_h$ is a subspace of a finite dimension, then the search for the solution to (3.7) can be reduced to solving a system of algebraic equations. In practice, this can be done by choosing a basis in $\mathbb{U}_h$. Let $\{\phi_k^i\}$, $i = 1, \ldots, N - 1$, be a basis in $\mathbb{U}_h$ consisting of the functions (see Figure 3.1)

$$\phi_k^i(x) = \begin{cases} \frac{x}{h} - (i - 1) & \text{if } x \in [(i - 1)h, ih], \\ \frac{x}{h} + i + 1 & \text{if } x \in [ih, (i + 1)h], \\ 0 & \text{otherwise.} \end{cases}$$
Chapter 3. Multigrid Preconditioners

Since both sides in (3.7) are linear with respect to \( \psi_h \), (3.7) holds for all \( \psi_h \) if

\[
\int_0^1 k(x) \psi_i'(x)^2 \psi_j'(x) \, dx = \int_0^1 f(x) \psi_i(x) \, dx \quad \forall \; i = 1, \ldots, N - 1.
\]  

(3.8)

Figure 3.1. Basis functions for piecewise linear elements.

We look for the solution \( u_h \) in the form of its decomposition with respect to the basis. Let \( z \) be the vector of the coefficients \( \{z_i\} \in \mathbb{R}^{N-1} \) of this decomposition:

\[
u_h = \sum_{j=1}^{N-1} z_j \psi_j h.
\]  

(3.9)

Insert the decomposition (3.9) into (3.8) to get

\[
\sum_{j=1}^{N-1} z_j \int_0^1 k(x)(\psi_j'(x))^2 \psi_i'(x) \, dx = \int_0^1 f(x) \psi_i(x) \, dx \quad \forall \; i = 1, \ldots, N - 1.
\]  

(3.10)

Relation (3.10) gives rise to the system of the linear algebraic equations, where the unknown vector is the vector of the coefficients \( z \in \mathbb{R}^{N-1} \):

\[
A_h z = b.
\]  

(3.11)

The matrix elements and the right-hand side are calculated from

\[
a_{ij} = \int_0^1 k(x)(\psi_j'(x))^2 \psi_i'(x) \, dx, \quad b_i = \int_0^1 f(x) \psi_i(x) \, dx.
\]  

(3.12)

As before, \( \langle y, z \rangle \) denotes the Euclidean inner product, \( ||y||_2 := \langle y, y \rangle^{1/2} \). To emphasize the mesh dependence, we shall use notation \( \mathbb{R}_h \) for the Euclidean space of coefficients of finite element functions, e.g., for the present example, \( \mathbb{R}_h = \mathbb{R}^{N-1} \).

Relation (3.9) defines the isomorphism \( P_h : \mathbb{R}_h \to U_h \) between the finite element function space \( U_h \) and the space of coefficients \( \mathbb{R}_h \). The important property of \( P_h \) is that up to a scaling factor, the isomorphism preserves an \( h \)-independent equivalence of the norms. More precisely, it holds that

\[
c_{\text{min}} \sqrt{b} ||z||_2 \leq ||P_h z||_2 \leq C_{\text{max}} \sqrt{b} ||z||_2,
\]  

(3.13)

with some constants \( c_{\text{min}} \) and \( C_{\text{max}} \) independent of \( h \) (check this!).
For arbitrary \( z, y \in \mathbb{R}_b \) and \( u_h = P_h z, v_h = P_h y \), the following equality holds:

\[
\langle A_h z, y \rangle = \int_0^1 k(x) u'_h v'_h \, dx.
\] (3.14)

The proof of (3.14) is left to the reader as an exercise. The identity (3.14) may serve as an alternative to (3.12) for the definition of the matrix \( A_h \) and will be frequently use in the analysis of the multigrid methods.

**Example 3.1.** Let \( k(x) = 1 \); then \( A_h \) from (3.11) is the tridiagonal matrix:

\[
A_h = \frac{1}{b} \begin{bmatrix}
2 & -1 & 0 \\
-1 & 2 & -1 \\
& \ddots & \ddots & \ddots \\
0 & -1 & 2 & -1 \\
\end{bmatrix} \in \mathbb{R}^{(N-1) \times (N-1)}. \tag{3.15}
\]

**Example 3.2.** Consider (3.3) with \( k(x) = 1 \) once again, but with different boundary conditions

\[
u(0) = 0, \quad \frac{\partial u}{\partial x}(1) = a.
\] (3.16)

To apply a finite element method for such a problem, one should first consider an appropriate weak formulation. Instead of \( H^1_0(0,1) \) as the basic functional space, one considers the subspace \( \bar{H}^1 \) of \( H^1(0,1) \) such that all functions from \( \bar{H}^1 \) vanish only on the left endpoint of the interval: \( u(0) = 0 \). Performing the integration by parts as in the previous example yields the relation

\[
\int_0^1 ku'v' \, dx = \int_0^1 f v \, dx + a(v(1)) \quad \forall v \in \bar{H}^1. \tag{3.17}
\]

The function \( u \) from \( \bar{H}^1 \) that satisfies (3.17) is the weak solution to (3.3) with the boundary conditions (3.16). The Neumann condition in the right endpoint \( x = 1 \) is accounted by the integral formulation (3.17).

The finite element space \( U_h \) is defined as all piecewise linear continuous functions on \([0,1]\) vanishing in \( x = 0 \). There are no extra conditions in \( x = 1 \). This corresponds to adding one extra basis function (a “half” of the one shown in Figure 3.1). The matrix \( A_h \) from (3.11) takes the following form (nonzero elements are only on three diagonals):

\[
A_h = \frac{1}{b} \begin{bmatrix}
2 & -1 & 0 \\
-1 & 2 & -1 \\
& \ddots & \ddots & \ddots \\
0 & -1 & 2 & -1 \\
\end{bmatrix} \in \mathbb{R}^{N \times N}. \tag{3.15}
\]

### 3.1.5 The Poisson equation

As another example, consider the Poisson problem in two space dimensions: Find the function \( u(x, y) \) satisfying

\[
-\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f \quad \text{in} \quad \Omega \in \mathbb{R}^2, \quad u|_{\partial \Omega} = 0. \tag{3.18}
\]
Multiplying (3.18) by an arbitrary function \( v(x,y) \) from \( H^1_0(\Omega) \) and integrating by parts, we get the weak formulation of the problem: Find \( u \in H^1_0(\Omega) \) satisfying the equality

\[
\int_{\Omega} \frac{\partial u}{\partial x} \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \frac{\partial v}{\partial y} \, dx \, dy = \int_{\Omega} f \, v \, dx \, dy \quad \forall \, v \in H^1_0(\Omega).
\] (3.19)

Similarly to the diffusion problem above, the finite element method replaces in (3.19) the Hilbert space \( H^1_0(\Omega) \) by a finite dimension space \( U_h \). To define \( U_h \), we assume that \( \Omega \) is divided into a finite number of triangles such that any two triangles have either an empty intersection, or a whole edge, or a vertex. The set of all such triangles we denote by \( T \) and call triangulation. Hence, we have \( \Omega = \bigcup_{T \in T} T \). Two important characteristics of a triangulation \( T \) are the regularity constant \( \beta \) and the maximal triangle diameter \( h \) defined from

\[
\beta := \sup_{T \in T} \frac{\rho(T)}{r(T)}, \quad h := \sup_{T \in T} r(T),
\] (3.20)

where \( r(T) \) and \( \rho(T) \) are the diameters of the superscribed and subscripted circles for \( T \in T \), respectively. If a family of triangulations \( \{T_h\}_{h>0} \) parameterized by some positive real parameter \( h \) is given, then it is said to satisfy a minimal angle condition if

\[
0 < \beta_h = \inf_{k>0} \beta_h, \quad \text{where } \beta_h := \sup_{T \in T_h} \frac{\rho(T)}{r(T)).
\] (3.21)

It is easy to see that the uniform lower bound on \( \beta_h \) yields that angles of triangles in \( \{T_h\}_{h>0} \) are uniformly bounded from below by a positive constant. Both parameters \( \beta \) and \( h \) are important for characterizing approximation properties of a finite element method for a given triangulation. In particular, \( \beta \) should not be too close to 0. For the case when \( \Omega = (0,1) \times (0,1) \), an example of regular uniform triangulation is shown in Figure 3.2.

We denote by \( P_k(T) \) the space of polynomials of degree less than or equal to \( k \) on \( T \) and let \( U_h \) be the space of all piecewise affine continuous functions (on every triangle, functions from \( U_h \) are polynomials of degree 1):

\[
U_h := \{ v \in C(\Omega) \mid v \in P_k(T) \text{ for any } T \in T \}.
\] (3.22)

It is clear that any function from \( U_h \) is fully defined by its values in the vertices of triangles. Let us enumerate all internal vertices, i.e., vertices lying in the domain interior, as shown in Figure 3.2. The \( i \)th basis function from \( U_h \) is defined as the piecewise linear continuous function taking the value 1 in the \( i \)th vertex and 0 in all other vertices. The stiffness matrix \( A_h \) is defined through the identity that is similar to (3.14):

\[
\langle A_h u, v \rangle = \int_{\Omega} \frac{\partial u_h}{\partial x} \frac{\partial v_h}{\partial x} + \frac{\partial u_h}{\partial y} \frac{\partial v_h}{\partial y} \, dx \, dy \quad \forall \, u, v \in \mathbb{R}, \quad u_h = P_h u, \quad v_h = P_h v.
\] (3.23)

The stiffness matrix is now the \( N \times N \)-block tridiagonal matrix

\[
A_h = \begin{bmatrix}
D & E & 0 & & \\
E & D & E & & \\
& \ddots & \ddots & \ddots & \\
0 & E & D & & \\
& & & & \end{bmatrix}.
\] (3.24)

3.1. The introductory section

The blocks $D$ and $E$ are the $N \times N$ matrices

$$D = \begin{bmatrix}
4 & -1 & 0 \\
-1 & 4 & -1 \\
\vdots & \ddots & \ddots \\
-1 & \ddots & -1 \\
0 & 4 & -1 
\end{bmatrix}, \quad E = -I, \quad b = \frac{1}{N+1}.$$

Another simple and natural way to build $\mathbb{U}_b$ is to divide $\Omega$ into elementary rectangles and consider all continuous functions on $\Omega$ to be bilinear for every rectangle. Any such function is uniquely defined by its values in the vertices of the rectangles. The matrix $A_h$ takes the same form as in (3.24), with the blocks

$$D = \frac{1}{3} \begin{bmatrix}
8 & -1 & 0 \\
-1 & 8 & -1 \\
\vdots & \ddots & \ddots \\
-1 & \ddots & -1 \\
0 & -1 & 8 
\end{bmatrix}, \quad E = \frac{1}{3} \begin{bmatrix}
-1 & -1 & 0 \\
-1 & -1 & -1 \\
\vdots & \ddots & \ddots \\
-1 & -1 & -1 \\
0 & -1 & -1 
\end{bmatrix}.$$

3.1.6 Approximation properties of finite elements

Now we try to explain what we mean by “the space $\mathbb{U}_b$ approximates the space $H^1_0(\Omega)$.”

Consider arbitrary $v(x)$ from $H^2(\Omega) \cap H^1_0(\Omega)$. Recall that we may assume $v(x)$ is a continuous function. Therefore, the Lagrangian interpolant $v_h \in \mathbb{U}_b$ is well-defined: $v_h(x) = v(x)$ if $x$ is a vertex of a triangle from $\mathcal{T}$. With the help of a simple scaling argument, one then shows the following interpolation estimates (cf. [61, 31]):

$$\inf_{v_h \in \mathbb{U}_b} \|v - v_h\|_{H^1} \leq C_1 b \|v\|_{H^1}, \quad \inf_{v_h \in \mathbb{U}_b} \|v - v_h\|_{L^2} \leq C_2 b^2 \|v\|_{H^1}, \quad (3.25)$$
where the constants \( C_1 \) and \( C_2 \) are independent of \( \nu \), but may depend on \( \beta \) from (3.20). Actually, the norm \( \|v\|_{H^1} \) on the right-hand side of (3.25) can be replaced by the semi-norm: \( \|v\|_{H^1}^2 := \int_T \sum_{i,j=1}^2 \frac{\partial v}{\partial x_i} \frac{\partial v}{\partial x_j} \, dx \).

The estimate (3.25) tells us that a sufficiently smooth function from \( H^1_0(\Omega) \) can be approximated by an element from \( \mathbb{U}_h \) with any prescribed accuracy in the \( H^1 \)-norm for sufficiently small \( h \).

The estimates given below show that the finite element solution is a reasonable approximation to the solution of the differential problem. The proofs of these estimates can be found in many textbooks on finite element method, in particular, [61, 31, 45]. Let \( u(x) \) be the solution to (3.18) and let \( u(x) \in H^2(\Omega) \). Let \( u_h \) be the finite element solution. It holds that

\[
\|u - u_h\|_{H^1} \leq C \, h \|u\|_{H^1}, \tag{3.26}
\]

where a constant \( C \) is independent of \( u \), but may depend on \( \beta \) and \( \Omega \).

The next result shows an estimate on the difference between \( u \) and \( u_h \) in the \( L^2 \)-norm. This estimate holds under an additional assumption on \( \Omega \): we assume that \( \Omega \) is a convex polygon. The estimate is stronger than the one in (3.26):

\[
\|u - u_h\|_{L^2} \leq C \, h^2 \|u\|_{H^1}, \tag{3.27}
\]

where a constant \( C \) is independent of \( u \), but may depend on \( \beta \) and \( \Omega \).

### 3.1.7 Stiffness matrices

In a more general setting, the Galerkin method for an abstract problem reads as follows: Find \( u_h \in \mathbb{U}_h \) satisfying

\[
a(u_h, v_h) = f(v_h) \quad \forall \ v_h \in \mathbb{U}_h,
\]

where \( \mathbb{U}_h \) is a finite-dimensional subspace of a Hilbert space, \( \dim(\mathbb{U}_h) = N \), and \( a(\cdot, \cdot) \) is a bilinear continuous form. A corresponding linear system of algebraic equations arises once a basis in \( \mathbb{U}_h \) is fixed. As we already know, the choice of basis defines the canonical isomorphism \( P_h \) between \( \mathbb{R}^N \) and \( \mathbb{U}_h \). The matrix of the system then satisfies

\[
(Au, v) = a(u_h, v_h) \quad \forall \ u, v \in \mathbb{R}^N, \quad \text{and} \quad u_h = P_h u, \ v_h = P_h v.
\]

Thus, properties of the stiffness matrix \( A \) are ruled by the choice of the basis, the properties of \( \mathbb{U}_h \), and the bilinear form. Obviously, if \( a(\cdot, \cdot) \) is symmetric or elliptic, then the matrix \( A \) is symmetric or positive definite, respectively. In addition, assume

\[
\text{supp}(u_h) \cap \text{supp}(v_h) = \emptyset \quad \Rightarrow \quad a(u_h, v_h) = 0; \tag{3.28}
\]

here \( \text{supp} \) denotes the support of a function. This implication holds for bilinear forms arising in finite element methods.

Another important property of \( A \) follows from the choice of basis in a finite element method, the minimal angle condition (3.21), and (3.28). To formulate it, assume we are given a family of triangulations, satisfying (3.21) and corresponding families of finite element spaces \( \mathbb{U}_h \) and stiffness matrices \( A_h \) parameterized by \( h \in (0, 1] \). It is natural to assume \( \dim(\mathbb{U}_h) \to \infty \) if \( h \to 0 \). In \( \mathbb{U}_h \) we choose a nodal basis \( \{\phi_i^h\}_{i=1}^{N_h} \) such that the support of \( \phi_i^h \) belongs to triangles sharing the node \( i \). According to the condition (3.21), a number of triangles sharing a single node is uniformly bounded over the whole family of triangulations (explain why!). This implies that support of \( \phi_i^h \) may have nonempty
overlap only with supports of basis functions whose total number is uniformly bounded. Since \((A_h)_{i,j} = a(\phi_h^i, \phi_h^j)\) and (3.28) holds, we have
\[
\max_{i=1,\ldots,N} \sum_{j=1}^{N} |\text{sign}(A_h)_{i,j}| \leq C,
\]
where \(C\) is some constant independent of \(h\). We obtained that the number of nonzero entries in any row of an entire matrix \(A_h\) is uniformly bounded independent of the matrix size. Therefore, we conclude that the number of nonzero entries in a matrix \(A_h \in \mathbb{R}^{N \times N}\) is \(O(N)\), i.e., it is proportional to the matrix size. Matrices with such a property are called sparse matrices.

The sparsity property is important for us. In particular, it follows that a matrix-vector product with such a matrix has the complexity \(O(N)\).

---

Exercises

3.1.1. Let \(f(x) \in L^2(0,1)\). Show the estimate (3.5) for the solution of (3.4).

3.1.2. Let \(f(x) \in L^2(0,1)\) and \(k(x) \in C^1(0,1)\); then the solution of (3.4) is from \(H^2(0,1)\). Show the estimate (3.6).

3.1.3. Check the equivalence (3.13) for the case of a one-dimensional finite element method and varying mesh size \(h\). How do the constants \(c_{\text{min}}\) and \(C_{\text{max}}\) depend on mesh size variation?

3.1.4. Verify that for arbitrary \(z,y \in \mathbb{R}\), \(u_h = P_h z\), \(v_h = P_h y\) the following equality holds:
\[
\langle A_h z, y \rangle = \int_0^1 k(x) u_h' v_h' \, dx.
\]

3.1.5. For arbitrary \(v(x) \in H^2(0,1) \cap H_0^1(0,1)\) prove the following interpolation estimate:
\[
\inf_{v_h \in U_h} \int_0^1 (v' - v_h')^2 \, dx \leq b^2 \int_0^1 (v'')^2 \, dx,
\]
and for arbitrary \(v(x) \in H_0^1(0,1)\) check
\[
\inf_{v_h \in U_h} \int_0^1 (v - v_h)^2 \, dx \leq b^2 \int_0^1 (v')^2 \, dx.
\]

3.1.6. Let \(u(x)\) be the solution to (3.4) and \(u(x) \in H^2(0,1)\). Let \(u_h\) be the finite element solution. Prove the estimate
\[
\int_0^1 (u' - u_h')^2 \, dx \leq K^2 b^2 \int_0^1 (u'')^2 \, dx.
\]

3.1.7. Let \(k(x) \in C^1(0,1)\) and the assumptions from Exercise 3.1.6 hold. Prove the estimate
\[
\int_0^1 (u - u_h)^2 \, dx \leq C b^2 \int_0^1 (u' - u_h')^2 \, dx.
\]

3.1.8. As defined in Section 3.1.7, is it true that an inverse of a sparse stiffness matrix is also a sparse matrix?
3.2 Two-grid iteration

Consider the two-point boundary value problem (3.3) with \( k(x) = 1 \) and apply the finite element method, as in Example 3.1. Doing so, we get the system of algebraic equations

\[
Az = b
\]  

(3.32)

with the tridiagonal matrix \( A \) as in (3.15).

The Gauss elimination can be efficiently applied to solve this particular system. Nevertheless, it is instructive to consider this system for the purpose of introducing the multigrid method. It is straightforward to check that the computational complexity of the Gauss elimination method for the discrete Poisson system from Section 3.1.5 is already \( O(N^{3.2}) \) for both linear and bilinear finite elements, where \( N \) is the number of unknowns in (3.51). The optimal complexity would be \( O(N) \). In the case of partial differential equations posed in three spatial dimensions, the gap between the optimal complexity and the complexity of a direct solver may become even larger. If the number of unknowns is large, one should look for an iterative method to solve the system.

3.2.1 The smoothing property of basic iterative methods

Let us consider one particular iterative method from Section 1.1 (Example 1.5): the Jacobi method. Additionally, we introduce a relaxation parameter \( w \): For a given \( z^0 \) compute for \( i = 0, 1, 2, \ldots \)

\[
z^{i+1} = z^i - wD^{-1}(Az^i - b).
\]  

(3.33)

We recall that \( D = \text{diag}(A) \) is the diagonal matrix built from the main diagonal elements of \( A \). Since for this particular example all diagonal elements are the same, the iteration (3.33) is equivalent to the Richardson method:

\[
z^{i+1} = z^i - \frac{w}{2}(Az^i - b).
\]  

(3.34)

The convergence rate of (3.34) can be estimated once eigenvalues of \( A \) are known. One verifies by direct computations that

\[A\psi^k = \lambda_k \psi^k,
\]

where \( \psi^k = \{\sin(\pi k j h)\}_{j=1}^{N-1}, \quad \lambda_k = 4h^{-1}\sin(\pi k h/2), \quad k = 1, \ldots, N-1. \)  

(3.35)

Note that \( \lambda_{\text{min}} \leq \pi^2 b, \lambda_{\text{max}} \sim 2b^{-1}, \) and so \( \text{cond}(A) = \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}} \geq \frac{\pi}{2}b^{-2} \). Therefore, with the optimal choice of \( w \), the convergence factor of (3.34) can be estimated by some \( q = 1 - O(h^2) \), and the number of iterations required for solving the system with relative accuracy \( \varepsilon \) can be estimated by \( O(h^{-2} \ln \varepsilon^{-1}) \). Clearly, we are far from the optimal computational complexity \( O(h^{-1}) \) for this example. The conclusion is that the Jacobi method cannot be accepted as an efficient solver for this system. The same conclusion is valid for the Gauss–Seidel method and other algorithms from Section 1.1. Note that the eigenvalues of \( A \) do not show any clustering, which may otherwise benefit a Krylov subspace method. Thus, preconditioning is essential for building an efficient iterative method.

Now, consider only “high-frequency” eigenvectors of matrix \( A \): these are \( \psi^k \) for \( k \geq N/2 = \frac{1}{2\pi} \). Note that for such \( k \), it holds that

\[2 = 4\sin^2\frac{\pi}{4} \leq h\lambda_k \leq 4.\]
3.2. Two-grid iteration

Let \( w = 1/2 \) in (3.34); then for the eigenvalues of the iteration matrix \( S = I - \frac{w}{2}A \), we have

\[
|\lambda_k(S)| = \left| 1 - \frac{b}{4} \lambda_k(A) \right| \leq \frac{1}{2} \text{ for } k \geq \frac{N}{2}.
\]  

(3.36)

The iteration matrix is responsible for the evolution of the error vector; cf. (1.12). Thus, the estimate (3.36) shows that on high frequencies the relaxed Jacobi method has the rate of convergence greater than or equal to \( q = 0.5 \). More insight can be gained if one considers the initial error decomposition using the orthogonal system of the eigenvectors:

\[
e^0 = \sum_{k=1}^{N} c_k \psi^k.
\]  

(3.37)

For the error vector on iteration \( m \) we get

\[
e^m = S^m e^0 = \sum_{k=1}^{N} c_k S^m \psi^k = \sum_{k=1}^{N} c_k [\lambda_k(S)]^m \psi^k.
\]  

(3.38)

Due to (3.36), the coefficients near eigenfunctions with index \( k \geq \frac{N}{2} \) in the error decomposition (3.38) are reduced on each iteration by at least a factor of 2. Therefore, after only very few iterations the impact of high frequencies in the error vector \( e = z^m - z \) is not significant. Since for large \( k \) the functions \( \psi^k \) are changing rapidly, it is common to say that the error become “smooth.” Iterations leading to such behavior of the error are commonly called smoothing iterations, and the iterative method is called smoother.

Summarizing, the following strategy of finding an approximate solution to (3.11) looks reasonable:

**Step 1.** A number of smoothing iterations of Jacobi type.

**Step 2.** A number of some iterations for damping low-frequency components in the error, to reduce coefficients multiplying \( \psi^k, k < \frac{N}{2} \), in the error decomposition (3.38).

**Step 3.** If the required accuracy was not reached, go to Step 1.

In the classical (geometric) multigrid method, Step 2 is called the coarse grid correction.

3.2.2 ■ Coarse grid correction

We proceed with two simple, but important, observations:

(a) If the error \( e \) is a mildly varying function, then we expect that it can be well approximated on a grid with a smaller number of nodes \( 9 \) (although we usually do not know \( e \) explicitly, it is important to be aware of the existence of such a coarse grid function approximating \( e \)).

(b) The error vector \( e \) solves the equation

\[
A e = d, \text{ where } d = A z' - b.
\]  

(3.39)

\( ^9 \)Strictly speaking, \( e \) is not a function, but a vector of coefficients from \( \mathbb{R}^h \), corresponding to a finite element function from \( U_h \). For the finite element function, it makes sense to consider an approximation on a coarser grid.
If we find $e$, the solution is obtained from $z = z^i - e$. Unfortunately, the system (3.39) is similar to the original system (3.11), and it can be as hard to solve exactly as the original. However, in view of observation (a), we may try to find an approximation $e_H$ to $e_b$ on a coarser grid with mesh size $H$ (for example, we consider a grid with the step $H = 2h$). This can be done by solving an approximation to the problem (3.39) on the coarser grid:

$$A_H e_H = d_H.$$  

If $p : \mathbb{R}_H \to \mathbb{R}_h$ is a prolongation operator from the coarse grid to the fine grid, then the new approximation $z^{i+1}$ to solution $z$ would be

$$z^{i+1} = z^i - p e_H.$$  

Relation (3.41) is called the coarse grid correction.

The following questions remain open:

a. How do we define the prolongation operator $p$?

b. How do we choose a restriction operator $r : \mathbb{R}_h \to \mathbb{R}_H$? The restriction operator is needed to define the residual $d_H = r d$ on the coarse grid in (3.40).

c. How do we define the coarse grid operator $A_H$ in (3.40)?

With these new notations, the coarse grid correction (3.41) can be rewritten as

$$z^{i+1} = z^i - p A_H^{-1} r(Az^i - b).$$  

The next two subsections address questions a–c.

### 3.2.3 Prolongation and restriction

If the system of linear algebraic equations comes from a finite element method, it is natural to define the prolongation and restriction operators in the following canonical way.

For the finite element method it is common to have $U_H \subset U_h$, where $U_H$ is a finite element space corresponding to the coarse grid. In this case, the operator $\bar{p}$ from $U_H$ to $U_h$ is trivial: $\bar{p}(u_H) = u_H$, and the prolongation in the spaces of coefficients $p : \mathbb{R}_H \to \mathbb{R}_h$ can be naturally defined thanks to the isomorphisms (see subsection 3.1.3) $P_h : \mathbb{R}_h \to U_h$ and $P_H : \mathbb{R}_H \to U_H$:

$$p : \mathbb{R}_H \xrightarrow{P_H} U_H \subset \mathbb{R}_h \xleftarrow{P^{-1}_h} \mathbb{R}_h.$$  

Thus, we get

$$p := P^{-1}_h P_H.$$  

For piecewise linear elements, the prolongation $p$ is illustrated in Figure 3.3, where the $i$th vector coefficient corresponds to the $i$th node. In a general case, $p$ is a (sparse) matrix that can be computed from given $P_h$ and $P_H$.

Consider the $L^2$-projection into the finite element space $\tilde{r} : U_h \to U_H$ defined by the relation

$$\langle \tilde{r} v_h, u_H \rangle = \langle v_h, u_H \rangle \quad \forall v_h \in U_h, u_H \in U_H.$$  

The canonical restriction $r : \mathbb{R}_h \to \mathbb{R}_H$ is built from the analogue to (3.44) in the space of coefficients

$$\langle r P^{-1}_h v_h, P^{-1}_H u_H \rangle = \langle P^{-1}_h v_h, P^{-1}_H u_H \rangle \quad \forall v_h \in U_h, u_H \in U_H.$$
3.2. Two-grid iteration

This relation gives \( r^* = P_H^{-1} P_H \), where \( r^* \) is an operator adjoint to \( r \), i.e., \( \langle r^*, \cdot \rangle_{\mathbb{R}^n_H} = \langle \cdot, r^* \rangle_{\mathbb{R}^n_H} \). From (3.43) we infer the identity

\[
    r = p^*.
\]

In practice (and theory), the relation (3.45) often serves as the “rule of thumb” for the choice of restriction. In our example and speaking in the terms of matrices, it is obvious that \( r = p^T \). The identity \( r = p^T \) is not necessarily true if one defines inner products in \( \mathbb{R}^n_H \) and \( \mathbb{R}^n_s \) differently from the Euclidean ones. The restriction is illustrated in Figure 3.4.

If in \( x = 0 \) the Dirichlet boundary condition is prescribed, then the restriction \( r \) is not defined for \( x = 0 \).

3.2.4 Coarse grid operator

There are two common approaches to defining the coarse grid operator \( A_H \):

1. Direct discretization. Within this approach, equation (3.3) is approximated directly on the coarse grid, i.e., using the subspace \( U_H \). For the example of the diffusion problem, the matrix \( A_H \) is defined through the relation

\[
    \langle A_H z, y \rangle = \int_0^1 k(x) u_H' v_H' \, dx
\]

for all \( z, y \in \mathbb{R}_H, u_H = P_H z, v_H = P_H y \).

2. Galerkin approach: Define \( A_H \) as the superposition of three operators:

\[
\]
For the canonical choice of $r$, $p$ and the finite element method from subsection 3.1.3, both approaches (3.46) and (3.47) produce the same matrix $A_H$. However, there are situations when two approaches for building the coarse grid correction matrix may give different matrices. This may happen if one uses inexact quadrature rules to compute entries of the stiffness finite element matrices ($A$, $A_H$, etc.). This is typically the case, for example, for equations with varying coefficients; see Example 5.3. Equality (3.47) may not be valid if some mesh-dependent terms are added to a finite element formulation. An example of such a finite element method can be found in subsection 5.1.10. Also equality (3.47) may fail to be true if, to discretize a differential problem, one uses a finite difference or a finite volume method rather than a finite element method.

### 3.2.5 Iteration matrix

All necessary components are now defined, and we are prepared to consider the two-grid method as an iteration: For a given iterate $z^k$, the next one, $z^{k+1}$, is computed in two steps. First, one performs a few smoothing steps ($\nu$ iterations) by some simple iterative method, say, the Jacobi method. Thanks to (1.13), the new intermediate iterate $\tilde{z}^k$ satisfies the relation

$$z^k = S^\nu z^k + g^\nu,$$

where $S$ is the iteration matrix for the smoothing step. For the Jacobi method and our example of $A$, it holds that $S = I - \frac{h^2}{2} A$. The vector $g^\nu$ depends on the right-hand side $b$ only. Second, $z^{k+1}$ is computed as a result of the coarse grid correction applied to $\tilde{z}^k$. Due to (3.42), this step can be written as

$$z^{k+1} = \tilde{z}^k - p A_H^{-1} r (A \tilde{z}^k - b).$$

The superposition with (3.48) yields

$$z^{k+1} = M z^k + \tilde{M} g^\nu - p A_H^{-1} r b = \bar{M} z^k + \bar{g}^\nu,$$

where $\bar{M}$ is the iteration matrix of the two-grid method:

$$\bar{M} = (I - p A_H^{-1} r A) S^\nu.$$  (3.49)

Here $\nu$ smoothing iterations are performed before the coarse grid correction.

Therefore, from the general theory of iterative methods, it follows that the two-grid iteration converges iff $|\lambda(\bar{M})| < 1$. A sufficient condition is given by the estimate $||\bar{M}|| < 1$ in some operator norm.

### 3.2.6 Convergence analysis for the model problem

For the given model problem and $r$, $p$ defined above, the convergence of the two-grid method can be verified by direct calculations and with the help of Fourier analysis. We will check that every pair of eigenvectors $\phi^k$ and $\phi^{N-k}$ forms an invariant subspace with respect to the action of $\bar{M}$. Therefore, in the basis of its eigenvectors, $\bar{M}$ has the block-diagonal form, with $2 \times 2$ blocks. Further in this section, we will check that the blocks $\bar{M}^{(k)}$, $k = 1, \ldots, \frac{N}{2} - 1$, have the form

$$\bar{M}^{(k)} = \begin{bmatrix} s_k^2 & c_k^2 \\ c_k^2 & s_k^2 \end{bmatrix} \begin{bmatrix} c_2^\nu & 0 \\ 0 & s_2^\nu \end{bmatrix}, \quad \bar{M}_b^{(k)} = 2^{-\nu},$$  (3.50)
where \( s_k^2 = \sin^2 \pi k |x|, \) \( c_k^2 = \cos^2 \pi k |x| \). The action of smoothing iterations and the coarse grid correction on low- and high-frequency parts of decomposition (3.37) is well seen from (3.50). The left factor in (3.50) corresponds to the coarse grid correction. By applying this matrix, “smooth” (slowly varying) eigenvectors \( \psi_k^s \), such that \( k \leq \frac{N}{2} \), are scaled with the small coefficient \( s_k^2 \). The situation is different with the right factor in (3.50). It corresponds to the smoothing iterations: highly oscillatory (“non-smooth”) eigenvectors are scaled with \( s_k^{2\nu} \).

One can obtain more precise bounds. Direct computations give

\[
\lambda(\tilde{M}(k)) \in [0, s_k^2 c_k^2 + c_k^2 s_k^{2\nu}], \; k = 1, \ldots, \frac{N}{2} - 1, \quad \lambda(\tilde{M}(\psi)) = 2^{-\nu}.
\]

Note that \( s_k^2 \in [0, 0.5] \) for \( k = 1, \ldots, \frac{N}{2} - 1 \). Therefore,

\[
|\lambda(\tilde{M})| \leq \max_{x \in [0, 0.5]} (x(1-x)^{\nu} + (1-x)x^{\nu}) \leq \max_{x \in [0, 0.5]} (x(1-x)^{\nu}) + \max_{x \in [0, 0.5]} ((1-x)x^{\nu}) \leq \frac{1}{\nu+1} \left(1 - \frac{1}{\nu+1}\right)^{\nu} + \left(\frac{1}{2}\right)^{\nu+1}.
\]

Thus, we get

\[
|\lambda(\tilde{M})| \leq \frac{1}{2} \quad \text{and} \quad |\lambda(\tilde{M})| \sim \frac{1}{e^{\frac{1}{\nu}}} \quad \text{for} \; \nu \to \infty.
\]

To estimate the spectral norm of the iteration matrix, we apply the Holders inequality (1.51):

\[
||\tilde{M}||_2 \leq ||\tilde{M}||_\infty ||\tilde{M}||_1 \leq \max_k ((s_k^2 c_k^2 + c_k^2 s_k^{2\nu}) 2 \max_k (s_k^2 c_k^2 + c_k^2 s_k^{2\nu})).
\]

Similar to the eigenvalue estimates above, we get

\[
||\tilde{M}||_2 \leq \frac{1}{2} \quad \text{and} \quad ||\tilde{M}||_2 \sim \frac{\sqrt{\frac{1}{\nu} \nu}}{e^{\nu}} \quad \text{for} \; \nu \to \infty.
\]

Therefore, for this model example, the norm of the iteration matrix for the two-grid iteration is estimated independently of \( h \). This is the major difference between the Jacobi method and other simple (basic) iterations. The dependence of the norm of the multigrid iteration matrix on the number of smoothing iterations of the form \( O(\nu^{-1}) \) is typical.

**The proof of (3.50)**

Since \( \psi_k^s \) and \( \psi^{N-k}_s \) are the eigenvectors of \( A \) (cf. (3.35)), and the iteration matrix of smoothing iterations \( S \) is \( I - \frac{1}{\nu} A \), we conclude that

\[
S^r \psi_k^s = (1 - s_k^2)^{\nu} \psi_k^s, \quad S^r \psi^{N-k}_s = (1 - c_k^2)^{\nu} \psi^{N-k}_s.
\]

Hence, the right factor in the decomposition (3.50) corresponds to the application of \( S^r \) to \( \{\psi_k^s, \psi^{N-k}_s\} \).
Denote by $\psi^k_H$ the $k$th eigenvector of the course matrix $A_H$,

$$\psi^k_H := \{\sin(\pi k m 2h)\}_{m=1}^{N-1}$$

Consider now the action of the coarse grid correction on the eigenvector $\psi^k$ for arbitrary $k = 1, \ldots, N-1$. Simple trigonometric identities yield

$$A\psi^k = \frac{4}{b^2} s^2_k \psi^k,$$

$$r \psi^k = 2c^2_k \psi^k,$$

$$A^{-1}_H \psi^k_H = \frac{b}{8} c^2_k s^{-2} \psi^k_H,$$

$$[p \psi^k_H]_j = \begin{cases} \psi^k_j & \text{if } j \text{ is even}, \\ (c^2_k - s^2_k)\psi^k_j & \text{if } j \text{ is odd.} \end{cases}$$

For the coarse grid correction operator, we get

$$[(I - pA^{-1}_H r A)\psi^k]_j = \begin{cases} 0 & \text{if } j \text{ is even,} \\ (2s^2_k)\psi^k_j & \text{if } j \text{ is odd.} \end{cases}$$

Note the following identities:

$$\psi^k_j = \begin{cases} -\psi^{N-k}_j & \text{if } j \text{ is even}, \\ \psi^{N-k}_j & \text{if } j \text{ is odd,} \end{cases}$$

$$j = 1, \ldots, N-1.$$ 

Thus we can write

$$(I - pA^{-1}_H r A)\psi^k = s^2_k \psi^k + s^2_k \psi^{N-k}.$$ 

Since $s^2_k = c^2_{N-k}$, then

$$(I - pA^{-1}_H r A)\psi^{N-k} = c^2_k \psi^k + c^2_k \psi^{N-k}.$$ 

From the last two equalities we conclude that the eigenvectors $\psi^k$ and $\psi^{N-k}$ indeed form an invariant subspace for the operator $(I - pA^{-1}_H r A)$, and thus for $M$. Taking $(\psi^k, \psi^{N-k})$ as the basis in this two-dimensional subspace, we can write

$$(I - pA^{-1}_H r A)(\psi^k, \psi^{N-k})^T = \begin{bmatrix} s^2_k & c^2_k \\ c^2_k & s^2_k \end{bmatrix} \begin{bmatrix} \psi^k \\ \psi^{N-k} \end{bmatrix}.$$ 

We obtain the left factor in the decomposition (3.50). In the special case of $k = N/2$ we note that $c^2_k = s^2_k = \frac{1}{2}$. Hence it holds that $M(\frac{1}{2}) = 2^{-\gamma}$.

### 3.2.7 Two-grid preconditioner

We already discussed in subsection 1.4.13 that any linear iteration for solving the system $Ax = b$ defines a preconditioner for $A$. Since the two-grid method is a linear iteration for solving (3.32), it defines a preconditioner $B_{TG}$. The crucial observation is that instead of providing the matrix $B_{TG}$ explicitly, the two-grid method defines the matrix vector product with matrix $B_{TG}$ and arbitrary vector $x \in \mathbb{R}_b$. To see this, assume we are given restriction, prolongation, and coarse grid operators $r, p, A_H$, and fix a matrix $W$ (e.g., $W = \omega^{-1}\text{diag}(A)$) and a number $\nu$ of smoothing iterations. Letting $x \in \mathbb{R}_b$ be an arbitrary vector, one computes $y = B_{TG}x$ in several steps, as defined in the next algorithm.
**Algorithm 3.1. Two-grid preconditioner.**

Input: Matrices $A$ and $A_H$, preconditioner $W$ for $A$, prolongation and restriction matrices $p$ and $r$, vector $x \in \mathbb{R}_{h}^n$.

1. Set $z^0 = 0$ (Initialization);
2. For $i = 0, \ldots, \nu - 1$ do
   
   $z^{i+1} = z^i - W^{-1}(A z^i - x)$  
   ($\nu$ smoothing steps);

3. Compute residual and restrict to the coarse grid:
   
   $r_H = r(A z^\nu - x)$;

4. Solve the reduced equation for the error on the coarse grid:
   
   $e_H = A^{-1}_H r_H$;

5. Correct $z^\nu$ to get the resulting vector
   
   $y = z^\nu - p e_H$.

Output: $y \in \mathbb{R}_h$ ($y = B_{TC} x$).

If a norm of the iteration matrix for the two-grid method is bounded by some $q < 1$, and $q$ is independent of the matrix size, then the condition number of the preconditioned matrix $B_{TC} A$ is also bounded independently of the matrix size; see subsection 1.4.13. This is certainly the case for the model discrete problem analyzed in this section and remains true for much more general situations, as we shall see later.

---

**Exercises**

3.2.1. Check (3.35).

3.2.2. Show that for the canonical choice of $r, p$ and the finite element method from subsection 3.1.3, both approaches (3.46) and (3.47) produce the same matrix $A_H$.

3.2.3. Verify that for the canonical choice (3.45) of $r, p$, the coarse grid correction operator $K_h = I - p A_H^{-1} r A$ is a projection, $K_h = K_h^2$. Hence it does not make sense to perform more than one consecutive coarse grid correction step.

3.2.4. What can be said about smoothing property of the steepest descent method from subsection 1.1.10?

**3.3 Multigrid iteration**

We are interested in solving a system of linear algebraic equations

$$A z = f,$$  
(3.51)

which arises from the discretization of partial differential equations by a finite element method. This can be diffusion or the Poisson equations from the previous section. We shall consider other examples in Chapter 5.
In the previous chapter, we considered a two-grid method. In the two-grid method, after a few basic smoothing iterations for system (3.51), we build a coarse grid problem for the unknown error vector \( A_H e_H = d_H \). The coarse grid problem is solved exactly. Further, we compute an approximation \( \hat{e} = p e_H \) to the unknown error \( e \) on the fine grid and correct the current iterate (approximation to \( z \)). The most computationally expensive step is solving the coarse grid problem. We note that \( \hat{e} \) is only an approximation to \( e \), and thus there are no definite reasons to compute \( e_H \) exactly. We can restrict ourselves only to finding an approximation \( \tilde{e}_H \) to \( e_H \). Therefore, to find an approximate solution of the coarse grid problem, let us apply a number (\( \gamma \) in total) of two-grid iterations. This step would involve calculations on an even coarser grid. A multigrid method consists of repeating this procedure recursively until a grid with the fewest number of nodes (the coarsest grid) is reached. On this coarsest grid, the system is often solved exactly, say, using the Gauss elimination.

Below we formalize the idea of the multigrid method. To this end, assume that a hierarchy of finite element spaces is given:

\[
\mathbb{U}_0 \subset \mathbb{U}_1 \subset \cdots \subset \mathbb{U}_k \subset \cdots \subset \mathbb{U}_l = \mathbb{U}_h. \tag{3.52}
\]

Subspace \( \mathbb{U}_0 \) corresponds to the coarsest grid. Operators of prolongation and restriction are acting between the corresponding spaces of coefficients:

\[
\begin{array}{ccccccc}
R_0 & R_1 & \cdots & R_k & R_{k+1} & \cdots & R_l \ \\
\downarrow \hspace{1cm} & \downarrow \hspace{1cm} & \cdots & \downarrow \hspace{1cm} & \downarrow \hspace{1cm} & \cdots & \downarrow \hspace{1cm} \\
p_1 & p_2 & \cdots & p_k & p_{k+1} & \cdots & p_l \\
\end{array}
\]

Assume that all grid level matrices \( A_k = A \) and \( A_k, k = 0, \ldots, l - 1 \), are given or computed as was explained in subsection 3.2.3.

For some given natural \( \nu \) (a number of smoothing iterations on every grid level) and \( \gamma \) (a number of recursive calls of the method on every grid level), one iteration of a multigrid method for solving the system \( A z = f \) is defined by the recursive procedure \( z^{new} = MGM(l, z^{old}, f) \), where \( z^{old} \) is an initial approximation to \( z \), and \( z^{new} \) is a computed new approximation. The procedure is formally defined below.

\[
\tilde{z} = MGM(k, z, b)
\]

\[
\begin{array}{c}
\text{If } k = 0, \text{ then} \\
\text{0. } \tilde{z} = A^{-1}_0 b \text{ (exact solution on the coarsest grid); Exit from the procedure.}
\end{array}
\]

\[
\begin{array}{c}
\text{Else } (k > 0) \\
\text{1. } z^0 = z, z^{i+1} = z^i - W_k^{-1}(A_k z^i - b), i = 0, \ldots, \nu - 1 \text{ (\( \nu \) basic iterations, smoothings);} \\
\text{2. } d = r_k(A_k z^\nu - b) \text{ (the restriction of residual to a coarse grid);} \\
\text{3. } e^0 = 0 \text{ (initial guess for an error on the coarse grid);} \\
\text{4. Execute } e^{i+1} = MGM(k - 1, e^i, d) \text{ for } i = 0, \ldots, \gamma - 1 \text{ (\( \gamma \) iterations of the multigrid on the coarse grid);} \\
\text{5. } \tilde{z} = z^\nu - p_k e^\nu \text{ (the coarse grid correction of } z^\nu) \\
\end{array}
\]

One call of the multigrid procedure \( z^{new} = MGM(l, z^{old}, f) \) constitutes one iteration of the multigrid method. If the multigrid method converges, then one may iterate until a residual norm \( ||A z^{new} - f|| \) becomes less than a given tolerance \( \epsilon \). Much more commonly the multigrid method is used to build an implicit preconditioner for matrix \( A \) from (3.51).
Multigrid preconditioner \( B_{MG} \) is built similarly to the two-grid preconditioner, see subsection 3.2.7. Indeed, the procedure \( MGM(l,0,f) \) is linear for \( f \), and \( B_{MG} \) is defined through the matrix-vector multiplication, \( y = B_{MG}x \): Fix a number of MG “cycles” \( m \); then for a given \( x \in \mathbb{R}_h \), vector \( y \in \mathbb{R}_h \) is found in several steps as defined in the next algorithm.

**Algorithm 3.2. Multigrid preconditioner.**

Input: Matrices \( A_k \), \( k = 0, \ldots, l \), prolongation and restriction matrices for every grid level, vector \( x \in \mathbb{R}_h \), integer parameters \( \gamma > 0, m > 0 \).

1. Set \( z^0 = 0 \) (Initialization);
2. For \( i = 0, \ldots, m - 1 \) do
   \[ z^{i+1} = MGM(l, z^i, x), \]
   End

Output: \( y := z^m \in \mathbb{R}_h \) (\( y = B_{MG}x \)).

If \( \gamma \) is not large, then in many cases (see Exercise 3.3.3), the computational complexity of one call to \( MGM(l, z^{old}, b) \) is \( O(N_l) \) elementary arithmetic and logical operations, where \( N_l \) is the number of unknowns for the finest grid. A constant multiplying \( N_l \) in \( O(N_l) \) depends on the number of smoothing steps. In the following subsections, we shall demonstrate that for many problems multigrid methods retain the remarkable property established earlier for the two-grid method: they give the significant error reduction on every iteration. More precisely, it holds that

\[
\|z - z^{new}\| \leq q \|z - z^{old}\| \quad \text{for} \quad z^{new} = MGM(l, z^{old}, f),
\]

with some \( q < 1 \) independent of the mesh size \( h_l \) or, in other words, independent on the number of unknowns for the finest grid. Moreover, by performing a sufficiently large number of smoothing iterations (but not too many in practice), we may ensure that the convergence factor is small, say, \( q \leq 0.1 \). This implies (see subsection 1.4.13), that \( \text{cond}(B_{MG}A) \leq c \), where the constant \( c \) is independent of \( h_l \) and can be done arbitrarily close to 1. Then, to solve the system (3.51) up to the desired (relative) accuracy \( \epsilon \) with a Krylov subspace method and multigrid preconditioner, one needs only \( O(M|\ln \epsilon|) \) arithmetic operations, where \( M \) is the complexity of the matrix-vector multiplication with matrix \( A \). This is the optimal complexity for an iterative method. In many circumstances, the multigrid preconditioner is the only approach known to deliver the optimal complexity for solving partial differential equations numerically.

### 3.3.1 V-, W- and F-cycles

A computational experience shows that in step 4 of the \( MGM \) procedure it is reasonable to choose \( \gamma = 1 \) or \( \gamma = 2 \). There are also theoretical reasons for this, which should become clear later. For \( \gamma = 1 \), one iteration of the method was named the \( V \)-cycle, and for \( \gamma = 2 \) it was named the \( W \)-cycle. Figures 3.5 and 3.6 illustrate these names. Arrows in the figures show a data transfer from one grid level to another. Horizontal arrows mean that computations are performed on a given grid level.

Together with V- and W-cycles the following F-cycle iteration became popular in applications. One can think of the F-cycle as a compromise between V- and W-cycles.
For solving a coarse grid problem, the F-cycle calls itself one time and the V-cycle one time; see Figure 3.7:

\[ \hat{z} = MGMF(k, z, b) \]

\{  
0. \( \hat{z} = A_{\gamma}^{-1} b \) (exact solution on the coarsest grid), exit from the procedure;

Else (\( k > 0 \))
\[ z^0 = z, z^{i+1} = z^{i} - W_{k}^{-1}(A_{k} z^{i} - b), i = 0, \ldots, \nu - 1 \) (\( \nu \) smoothing iterations);
1. \( d = r_{k}(A_{k} z^{\nu} - b) \) (the restriction of residual to a coarse grid);
2. \( e^0 = 0 \) (initial guess for an error on the coarse grid);
3. \( e^1 = MGMF(k - 1, e^0, d) \) (F-cycle on the coarse grid),
4. \( e^2 = MGM(k - 1, e^1, d) \) for \( \gamma = 1 \) (V-cycle on the coarse grid);
5. \( \hat{z} = z^{\nu} - p_{k} e^2 \) (the coarse grid correction).
\}

### 3.3.2 Complexity of a multigrid preconditioner

As discussed in subsection 3.1.7, the number of nonzero entries in \( A \) is typically proportional to the number of unknowns \( N_l \). Assume that the same property holds on every grid level \( k \):

\[ \sum_{i,j=1}^{N_k} |\text{sign}(A_k)_{i,j}| = O(N_k) \].

This is a plausible assumption if \( A_k \) are built directly by applying a finite element method on coarser grids or if \( A_{k-1} = r_k A_k p_k \), with a canonical choice of prolongation and restriction.
The computational complexity of an iterative algorithm for solving (3.51) is optimal if the number of arithmetic operations required to find the solution, with a given accuracy, linearly depends on the number of unknowns. Below we study the complexity of one multigrid cycle. The complexity of a multigrid preconditioner $B_{MG}$ equals the complexity of one cycle multiplied by a number of cycles used to define $B_{MG}$.

Denote by $N_k$ the number of unknowns on the $k$th grid level. Let

$$\Upsilon_n := \inf_{k=1,\ldots,l} \frac{N_k}{N_{k-1}}.$$  \hspace{1cm} (3.53)

Since the number of nonzero entries in $A_k$ is proportional to the number of unknowns $N_k$ on the corresponding grid level, then a reasonable smoothing iteration such as Jacobi or Gauss–Seidel iteration requires $O(N_k)$ arithmetic operations. The number of smoothing iterations is assumed to be some fixed (not too large) number, independent of $k$. Projection and restriction are local operations. Hence on the $k$th grid level each of these operations requires $O(N_k)$ arithmetic operations.

Denote by $N_{mg}$ the complexity of one multigrid cycle. It is easy to check the following proposition, which we leave to the reader as the exercise.

**Proposition 3.3.** Denote $x = \gamma \Upsilon_n^{-1}$. It holds that

(a) If $x < 1$, then $N_{mg} = O(N_l)$.

(b) If $x = 1$, then $N_{mg} = O(l N_l)$.

(c) If $x > 1$, then $N_{mg} = O(x^l N_l)$.

For the model example of the diffusion equation in one space dimension from Section 3.1, one gets $\Upsilon_n = 2$. Hence the V-cycle falls in (a), and the W-cycle falls in (b). For the Poisson problem and uniform grid refinement both cycles are optimal.

### 3.3.3 Iteration matrix

The smoothing step 1 of the multigrid procedure $MGM$ from the previous sections is commonly called *presmoothing*, since it is performed before the coarse grid correction step. Smoothing iterations also can be performed after the coarse grid correction; then they are called *postsmoothing*. In a multigrid method, it is common to perform both presmoothing and postsmoothing iterations.
Example 3.4. The multigrid V-cycle with one presmoothing and one postsmoothing step can be defined by the execution of the following recursive procedure, $z^{new} = MGMV(l, z^{old}, f)$:

$$
\tilde{z} = MGMV(k, z, b)
\begin{cases}
\quad \text{If } k = 0, \text{ then}
\quad \tilde{z} = A_0^{-1}b; \text{ exit the procedure;}
\quad \text{Else (} k > 0 \text{)}
\quad 1. \ z^0 = z, \ z^1 = z^0 - W_{k-1}^{-1}(A_k z^0 - b) \text{ (one presmoothing iteration)};
\quad 2. \ d = r_k(A_k z^1 - b) \text{ (restriction of the residual on a coarser grid)};
\quad 3. \ e^0 = \tilde{e} \text{ (initial guess for the error)};
\quad 4. \ e^1 = MGMV(k - 1, e^0, d) \text{ (the recursive call of the same procedure on the coarser grid)};
\quad 5. \ z^2 = z^1 - p_k e^1 \text{ (the coarse grid correction)};
\quad 6. \ \tilde{z} = z^2 - W_{k-1}^{-1}(A_k z^2 - b) \text{ (one postsmoothing iteration)};
\end{cases}
\} \quad □

Assume that on every grid level we perform $v_1$ presmoothing iterations (we call them “presmoothing”) and $v_2$ postsmoothing iterations (“postsmoothing”).

Clearly, the multigrid method as defined above is a linear iteration. Therefore, the properties of the corresponding multigrid preconditioner can be understood by looking at the corresponding iteration matrix. Denote by $M_l(v_1, v_2)$ the iteration matrix of one cycle: $z^{new} = MGMV(l, z^{old}, f)$, with $v_1$ presmoothings and $v_2$ postsmoothings. The following lemma gives a useful representation of $M_l(v_1, v_2)$.

Lemma 3.5. The iteration matrix $M_l(v_1, v_2)$ can be defined recursively:

$$
\begin{align*}
M_0(v_1, v_2) &= 0, \quad (3.54) \\
M_k(v_1, v_2) &= S_k^{v_2} \left( I - p_k (I - M_{k-1}^{v_1} A_k^{-1} r_k A_k) S_k^{v_1} \right), \quad k = 1, \ldots, l. \quad (3.55)
\end{align*}
$$

Proof. Let us check the relations (3.54) and (3.55) using the induction in $k$. We recall that the iteration matrix is the matrix responsible for the error evolution on every iteration step. For $k = 0$, the identity (3.54) is valid, since on the coarsest grid the system is solved exactly. We check (3.55) for arbitrary $k > 0$. $S_k$ is an iteration matrix for smoothing steps. The terms $S_k^{v_1}$ and $S_k^{v_2}$ clearly correspond to $v_1$ smoothing iterations before the coarse grid correction and $v_2$ smoothing iterations after it.

Now we find the coarse grid correction matrix. Let $d_{k-1} = r_k(A_k z_k^{v_1} - b_k)$ be the residual restriction to grid level $k - 1$, where $z_k^{v_1}$ is an approximation to the solution of the system $A_k z_k = b_k$ after $v_1$ smoothing iterations. On grid level $k - 1$ we solve approximately the system

$$
A_{k-1} z_{k-1} = d_{k-1}. \quad (3.56)
$$

For solving (3.56), we perform $\gamma$ iterations (cycles) of a multigrid method on level $k - 1$ with zero initial guess. By induction we know that $M_{k-1}$ from (3.54) is the iteration matrix
of these iterations. We denote the new approximation to \( z_{k-1} \) by \( z_{k-1}^\gamma \). From (1.158) we have

\[
    z_{k-1}^\gamma = (I - M_{k-1}^\gamma)A_{k-1}^{-1}d_{k-1}.
\]

(3.57)

The new approximation \( z_{k-1}^\gamma \) is prolonged to level \( k \), and the new iterate \( z_k^{i+1} \) is computed from

\[
    z_k^{i+1} = z_k^i - p_k z_{k-1}^\gamma.
\]

From this and (3.57), we get

\[
    z_k^{i+1} = z_k^i - p_k (I - M_{k-1}^\gamma)A_{k-1}^{-1}r_k (A_k z_k^i - b_k).
\]

Therefore, the iteration matrix for the coarse grid correction is

\[
    I - p_k (I - M_{k-1}^\gamma)A_{k-1}^{-1} r_k A_k.
\]

Superposition with presmoothing and postsmoothing matrices gives (3.55).

We know that one way to prove the convergence of an iterative method or analyze a preconditioner is to estimate a norm of the corresponding iteration matrix. For the multigrid method, this task is significantly simplified by using the following relation between the iteration matrices of multigrid and two-grid methods.

Let \( \bar{M}_k(v_1, v_2) \) be an iteration matrix of the two-grid method, with \( v_1 \) presmootherings and \( v_2 \) postsmootherings. The identities

\[
    M_k(v_1, v_2) = \bar{M}_k(v_1, v_2),
\]

(3.58)

\[
    M_k(v_1, v_2) = \bar{M}_k(v_1, v_2) + S_k^v p_k M_{k-1}^\gamma (v_1, v_2)A_{k-1}^{-1} r_k A_k S_k^v \]

(3.59)

hold for \( k > 1 \).

The relations (3.58) and (3.59) follow from (3.55) and the representation (3.49) for the two-grid iteration matrix

\[
    \bar{M}_k(v_1, v_2) = S_k^v (I - p_k A_{k-1}^{-1} r_k A_k) S_k^v.
\]

### 3.3.4 Convergence of the multigrid iteration

In the remainder of Section 3.3, we prove some estimates on the convergence factor of the multigrid iteration. We follow the framework of smoothing and approximation properties introduced in [96]. These estimates shows that \( W \)- and \( V \)-cycles are useful both as stand-alone solvers and as efficient preconditioners. First, we consider a more general case, and further, the results are refined for the case of symmetric positive definite systems. Concerning the existing convergence theory of multigrid methods, we note the following. If the finite element stiffness matrix is symmetric and positive definite, then \( V \)- and \( W \)-cycles for solving the corresponding linear algebraic system would be convergent iterations already with one smoothing step on every grid level. In a more general case, if a matrix is not necessarily symmetric or positive definite, we prove the convergence of the \( W \)-cycle if the number of smoothing steps on each grid level is sufficiently large, though independent of a level number. This restriction in the general case is hard (or even impossible) to avoid. We start the analysis by looking at this more general case.
3.3.5 Convergence of the W-cycle

In this subsection, $\| \cdot \|$ is a generic vector norm and corresponding operator norm for matrices. Assume the following conditions on prolongation and smoothing iterations. For all $k = 1, \ldots, l$ and all $\nu > 0$, it holds that

$$c_p \|x\| \leq \|p_k x\| \leq C_p \|x\|, \quad (3.60)$$
$$\|S_k^\nu\| \leq C_s, \quad (3.61)$$

with constants $c_p$, $C_p$, and $C_s$ independent of $k$. Suppose that $P_k : \mathbb{R}_k \to \mathbb{U}_k$ is an isomorphism such that the ratio $C_{\text{max}}/\epsilon_{\text{min}}$ for constants from (3.13) is independent of $k$; then the inequalities (3.60) are valid. Assumption (3.61) is true, for example, for all convergent smoothing iterations.

The statement of the theorem below can be conventionally formulated as “the convergence of the two-grid iteration implies the convergence of the W-cycle.” First, we consider the case of no postsmoothers: $\nu_2 = 0$.

**Theorem 3.6.** Assume the conditions (3.60) and (3.61) hold and the norm of the two-grid iteration matrix satisfies the estimate

$$\|\tilde{M}_k(\nu,0)\| \leq \eta(\nu) < 1,$$

with a function $\eta(\nu)$ independent of $k$ such that $\eta(\nu) \to 0$ for $\nu \to 0$. Then there exists $\tilde{\nu} > 0$ such that for all $\nu \geq \tilde{\nu}$ the multigrid method with $\gamma \geq 2$ converges, and

$$\|M_k(\nu,0)\| \leq \frac{\gamma}{\gamma - 1} \eta(\nu). \quad (3.62)$$

**Proof.** We prove the theorem for the case $\gamma = 2$ (W-cycle). The proof is based on the recursive relation (3.59). We start by noting the relations

$$\|S_k^\nu\| = 1, \text{ since } \nu_2 = 0, \quad (3.63)$$
$$\|p\|_{\mathbb{R}_k^{l-1} \to \mathbb{R}_k} \leq C \text{ due to (3.60),} \quad (3.64)$$
$$pA_{k-1}^{-1} r A_k S_k^\nu = S_k^\nu - (I - pA_{k-1}^{-1} r A_k) S_k^\nu = S_k^\nu - \tilde{M}_k(\nu,0). \quad (3.65)$$

From (3.60), (3.61), and (3.65) it follows that

$$\|A_{k-1}^{-1} r A_k S_k^\nu\| \leq C \|pA_{k-1}^{-1} r A_k S_k^\nu\| = C \|S_k^\nu - \tilde{M}_k\|$$
$$\leq C (\|S_k^\nu\| + \|\tilde{M}_k\|) \leq C (C_s + 1). \quad (3.66)$$

For the last inequality, we used the fact that $\|\tilde{M}_k\| \leq 1$ for sufficiently large $\nu$.

Denote $\xi_k := \|M_k(\nu,0)\|$. To complete the proof, we need to show the estimate (3.62), that is $\xi_k \leq 2\eta(\nu)$ for all $k$. We check it with the help of induction in $k$. The basis of induction is the case $k = 1$. Indeed, the equality (3.58) and the theorem assumption for matrices $M_k$ give

$$\xi_1 \leq \eta(\nu).$$

Thanks to (3.59), we conclude for $k > 1$ that

$$\xi_k \leq \|\tilde{M}_k(\nu,0)\| + \|p M_{k-1}^2(\nu,0) A_{k-1}^{-1} r A_k S_k^\nu\|$$
$$\leq \eta(\nu) + \|p\| \|M_{k-1}^2(\nu,0)\| \|A_{k-1}^{-1} r A_k S_k^\nu\|.$$
3.3. Multigrid Iteration

Now (3.58), (3.59) and (3.63), (3.64) and (3.66) give the recursive relation
\[ \xi_1 \leq \eta(v), \quad \xi_k \leq \eta(v) + C^* \xi_{k-1}^2, \quad k = 2, 3, \ldots, l. \]  
(3.67)

We take \( \tilde{v} \) such that \( \eta(v) \leq \frac{1}{4} \) for \( v \geq \tilde{v} \); then \( \xi_k \leq 2\eta(v) \). The latter follows from (3.67) and the inequality \( \eta(v) + C^* 4\eta^2(v) \leq 2\eta(v) \), which holds for \( \eta(v) \leq \frac{1}{4} \). \hfill \Box

The convergence for \( \nu_2 \neq 0 \) can be proved with similar arguments.

3.3.6 Smoothing and approximation properties

Theorem 3.6 is an important (though simple) result in the analysis of multigrid methods. Two other important ingredients, which together with Theorem 3.6 constitute our framework, are the following two properties (smoothing and approximation). Verifying these (or similar) properties for a particular class of matrices is the major part of the analysis of multigrid preconditioners for a concrete problem.

1. Smoothing property: There exists a function \( \eta(v) : \mathbb{R}_+ \to \mathbb{R}_+ \) such that \( \eta(v) \to 0 \), \( v \to \infty \), and
\[ \|A_k S_{\nu}^k\| \leq \eta(v)\|A_k\| \quad \forall k \geq 0. \]  
(3.68)

The function \( \eta(v) \) does not depend on \( k \).

2. Approximation property: For some constant \( C_A > 0 \), it holds that
\[ \|A_k^{-1} - p_k A_{k-1}^{-1} r_k A_{k-1}^{-1} r_k^T\| \leq C_A\|A_k\|^{-1} \quad \forall k \geq 1. \]  
(3.69)

\( C_A \) does not depend on \( k \).

The following theorem about the convergence of the two-grid method holds.

**Theorem 3.7.** Assume the approximation and smoothing properties; then there exists \( \tilde{v} > 0 \) such that for all \( v \geq \tilde{v} \) the two-grid method converges, and
\[ \|\tilde{M}_k(v,0)\| \leq C_A \eta(v). \]  
(3.70)

**Proof.** The proof follows directly from the equalities
\[ \tilde{M}_k(v,0) = (I - p_k A_{k-1}^{-1} r_k A_k) S_{\nu}^k = (A_k^{-1} - p_k A_{k-1}^{-1} r_k) (A_k S_{\nu}^k). \]

The norm on the right-hand side is estimated by the product of the norms of the terms in brackets. We apply (3.68) and (3.69) and obtain (3.70). Thanks to the smoothing property for \( v \) large enough, the right-hand side in (3.70) is arbitrarily small, in particular, less than 1. The latter guarantees the convergence of the two-grid iteration. \hfill \Box

The smoothing and approximation properties, and assumptions (3.60) and (3.61), ensure a nontrivial estimate on a norm of the iteration matrix of W-cycles. The estimate is independent of the total number of grid levels. This implies the following optimality result: If the arithmetic complexity of one iteration (one cycle) of the method is optimal, i.e., proportional to the number of unknowns on the finest mesh and \( \kappa < 1 \), where \( \kappa \) is defined as in Proposition 3.3.3, then the total arithmetic complexity of a Krylov subspace method with a multigrid preconditioner is also optimal.

**Remark 3.1.** The appearance of the terms \( \|A_k\| \) and \( \|A_k\|^{-1} \) in the right-hand sides of (3.68) and (3.69) is not necessary, although it has methodological sense.
A less formal explanation of these properties can be done as follows. $A_k$ is commonly a matrix arising from discretization of a second order differential operator. Hence the multiplication of $A_k$ and a vector $y$ should “emphasize” the contribution of high-frequency harmonics in $y$. Similar, the computing of the second derivative of

$$f(x) = \sum_{m=\pm}^{\infty} \{a_m \sin(mx) + b_m \cos(mx)\}$$

increases coefficients in front of the $m$th harmonic in $m^2$ times. An estimate on $A_kS^\nu_k y$ would mean that after $\nu$ smoothing iterations we have only a small contribution of high-frequency harmonics in $S^\nu_k y$. Concerning the function $\eta(\nu)$ in the smoothing property (3.69), we note that it is typical to prove the asymptotic of the form

$$\eta(\nu) = O\left(\frac{1}{\nu^\alpha}\right), \quad \alpha > 0.$$  

(3.71)

An estimate on $A_k^{-1} - p_kA_{k-1}^{-1}r_k$ in the approximation property tells us that the operators on two grid levels $k$ and $k-1$ are close. In other words, the coarse grid problem on level $k-1$ is a good approximation to the problem on level $k$.

### 3.3.7 The case of a self-adjoint problem

Assume that the matrix $A$ from (3.51) is symmetric positive definite. This is usually the case if a finite element method is applied to discretize differential problems with self-adjoint operators. Example of such problems are the diffusion equation and the Poisson problem. In this subsection, we analyze the convergence of the multigrid V-cycle.

Thus, for the finest grid matrix $A_l = A$ we assume $A_l = A^*_l > 0$. An estimate on the iteration matrix of the V-cycle is proved in the “energy” $A$-norm. Restriction, prolongation, and a coarse grid operator are defined canonically as

$$r_k = p_k^*, \quad A_{k-1} = r_k A_k p_k.$$  

(3.72)

It will be convenient to assume smoothing and approximation properties in a slightly different form.

**Smoothing property:** For the iteration matrix on every grid level assume that it can be written as $S_k = I - W_k^{-1}A_k$, with a preconditioner $W_k$ such that

$$W_k = W_k^T \geq A_k.$$  

(3.73)

**Approximation property:** This will be assumed in the form (compare with (3.69))

$$\|A_k^{-1} - p_kA_{k-1}^{-1}r_k\|_2 \leq C_A\|W_k\|^{-1}.$$  

(3.74)

It can be verified that (3.73) and (3.74) imply the assertion of Theorem 3.7 about the two-grid method convergence (we leave this to the reader as exercise). This result shows that the smoothing property (3.73) and approximation property (3.74) complement each other.

**Example 3.8.** For the damped Jacobi method, one has $W_k = \frac{1}{w}D_k$, and the condition (3.73) can always be satisfied by choosing the relaxation parameter $w$ small enough. The constant $C_A$ from (3.74) generally depends on $w$.  


We recall that for a symmetric positive definite matrix $A$ the bilinear form $\langle Ax, y \rangle$ defines the inner product $\langle x, y \rangle_A$ and the norm. We use the notion $\| \cdot \|_A$ also for the corresponding matrix norm:

$$\|M\|_A := \sup_{0 \neq x \in \mathbb{R}^n} \frac{\|Mx\|_A}{\|x\|_A} \quad \text{for } M \in \mathbb{R}^{n \times n}.$$ 

Now we are ready to prove the main theorem about the V-cycle convergence.

**Theorem 3.9.** Assume (3.72)-(3.74). The V-cycle, with $\nu_1 = \nu_2 = \nu/2$, $\nu = 2, 4, \ldots$, converges. Moreover,

$$\left\| M_{\nu} \left( \begin{array}{c} \nu \\ \frac{\nu}{2} \end{array} \right) \right\|_A \leq \frac{C_A}{C_A + \nu}, \quad (3.75)$$

where $C_A$ is the constant from (3.74).

**Proof.** To prove the theorem, and further in the text, we need a number of simple propositions, which we collect in one lemma below.

**Lemma 3.10.** Let $A, A_1, A_2, B, M, W$ be $n \times n$ real matrices. Moreover, let $A = A^T > 0$; then

(a) $\|M\|_A = \|A^{1/2} MA^{-1} \|_2$;

(b) $\|A\|_2 \leq c$ is equivalent to $A \leq c I$;

(c) $A_1 \leq A_2$ implies $BA_1B^* \leq BA_2B^*$;

(d) sp$(AB) = sp(BA)$;

(e) assume also $W = W^T > 0$; then $\|A\|_2 \leq \|W\|^{-1}_2$ yields $A \leq W^{-1}$.

**Proof.** Let us check item (e). Since the matrices $A$ and $W$ are symmetric, and thus $W^{-1}$ is symmetric, the inequality $\|A\|_2 \leq \|W\|^{-1}_2$ is equivalent to the inequality between maximum eigenvalues (the positive definiteness of $A$ and $W$ implies that the eigenvalues are positive):

$$\lambda_{\max}(A) \leq (\lambda_{\max}(W))^{-1}.$$ 

Since $W = W^T > 0$, it holds that (check this) $\lambda_{\max}(W)^{-1} = \lambda_{\min}(W^{-1})$. We get

$$\lambda_{\max}(A) \leq \lambda_{\min}(W^{-1}). \quad (3.76)$$

The straightforward consequence of the Rayleigh relations is

$$\langle Ax, x \rangle \leq \lambda_{\max}(A) \|x\|_2^2, \quad \lambda_{\min}(W^{-1}) \|x\|^2_2 \leq \langle W^{-1}x, x \rangle.$$

From this and (3.76) we conclude that

$$\langle Ax, x \rangle \leq \langle W^{-1}x, x \rangle \quad \forall \, x \in \mathbb{R}^n. \quad (3.77)$$
We continue to prove the theorem. Thanks to (3.72), \( A_l = A_l^T \) implies \( A_{l-1} = A_{l-1}^T \), and repeating the argument, we get \( A_k = A_k^T \) for all \( k = 0, \ldots, l \). Item (a) of Lemma 3.10 then gives
\[
\| M_k \|_{A_k} = \| A_k^T M_k A_k^{-1} \|_2.
\]
We will check below that the matrices
\[
M' := A_k^T M_k \begin{pmatrix} 
\nu & \nu \\
2 & 2 
\end{pmatrix} A_k^{-1}
\]
are self-adjoint. Due to item (b) of Lemma 3.10, to prove (3.75) it is sufficient to check that
\[
0 \leq M'_k = A_k^T M_k A_k^{-1} \leq \frac{C_A}{C_A + \nu} I. \tag{3.78}
\]
Let us prove (3.78) by induction in \( k \). Towards this end, we verify the following statement: Let
\[
0 \leq M'_{k-1} \leq \xi_{k-1} I < I;
\]
then
\[
0 \leq M'_k \leq \max_{\xi \in [0, 1]} \{ (1 - \xi)' (\xi_{k-1} + (1 - \xi_{k-1}) C_A \xi) \} I. \tag{3.79}
\]
Assume for a moment that this proposition is true; then we put \( \xi_{k-1} = \xi_{k-1}^{(r)} \). The maximum in the right-hand side of (3.80) is attained for \( \xi = 0 \) and equals \( \frac{C_A}{C_A + \nu} \). This proves (3.78) by induction with \( k = 0 \) in the basis of induction (we recall that \( \| M'_{0} \| = 0 \), and thus (3.79) trivially holds).

It remains to check that (3.79) implies (3.80). We compute, by using the definition (3.55),
\[
M'_k = A_k^T \chi^T_k (I - p_k (I - M_{k-1}) A_{k-1}^{-1} r_k A_k) S_k^T A_k^{-1}
\]
\[
= (A_k^T S_k^T A_k^{-1} ) (I - A_k^T p_k (I - M_{k-1}) A_{k-1}^{-1} r_k A_k) (A_k^T S_k^T A_k^{-1} ).
\]
Matrices inside the brackets can be manipulated as
\[
A_k^T S_k^T A_k^{-1} = (I - A_k^T W_k^{-1} A_k^T)^T.
\]
and
\[
A_k^T p_k (I - M_{k-1}) A_{k-1}^{-1} r_k A_k = (A_k^T p_k A_k^{-1} ) (I - A_k^T A_{k-1}^{-1} M_{k-1} A_{k-1}^{-1} ) (A_k^T r_k A_k^{-1} ).
\]
We get
\[
M'_k = (I - X_k)^T (I - p' (I - M'_{k-1}) r' )(I - X_k)^T, \tag{3.81}
\]
where
\[
X_k = A_k^T W_k^{-1} A_k^T, \quad p' = A_k^T p_k A_k^{-1}, \quad r' = A_{k-1}^{-1} r_k A_k^{-1}.
\]
Since \( p' = (r')^T \), from (3.81) we see that \( M'_{k-1} = (M'_{k-1})^T \) yields \( M'_{k} = (M'_{k})^T \). By the same arguments, we check that \( M'_k (\frac{1}{2}, \frac{1}{2}) \) are symmetric for all \( k \).
From (3.81) and (3.79) we conclude that (the inequality below follows from item (c) of the lemma with the matrix \( B = (I - X_k)^{\frac{\gamma}{2}} p' \))

\[
M'_k \leq (I - X_k)^{\frac{\gamma}{2}} (I - (1 - \xi_{k-1}) p' r') (I - X_k)^{\frac{\gamma}{2}}
\]

\[
= (I - X_k)^{\frac{\gamma}{2}} ((1 - \xi_{k-1})(I - p' r') + \xi_{k-1} I) (I - X_k)^{\frac{\gamma}{2}}
\]

\[
= (I - X_k)^{\frac{\gamma}{2}} ((1 - \xi_{k-1}) Q_k + \xi_{k-1} I) (I - X_k)^{\frac{\gamma}{2}},
\]

where

\[
Q_k = A_k^\frac{\gamma}{2} (A_k^{-1} - \rho A_{k-1}^{-1} r) A_k^\frac{\gamma}{2} \leq C A_k^\frac{\gamma}{2} W_k^{-1} A_k^\frac{\gamma}{2} = C A_k.
\]

The last inequality follows from the approximation property (3.74) and item (d). Therefore

\[
M'_k \leq (I - X_k)^{\frac{\gamma}{2}} ((1 - \xi_{k-1}) C A_k + \xi_{k-1} I) (I - X_k)^{\frac{\gamma}{2}}.
\]

From (3.81) it follows that \( sp(W_k^{-1} A_k) \in [0, 1] \), and from item (e) we have

\[
sp(W_k^{-1} A_k) = sp(A_k^\frac{\gamma}{2} W_k^{-1} A_k^\frac{\gamma}{2}) = sp(X_k) \in [0, 1].
\]

This together with (3.82) proves (3.80).

**Exercises**

3.3.1. Check the multigrid cycle complexity dependence in (3.3).

3.3.2. Estimate the complexity of the F-cycle for the model example of the one-dimensional diffusion equation.

3.3.3. For the iteration matrix of the two-grid method, show the identity

\[
\rho(\tilde{M}_k(v_1, v_2)) = \rho(\tilde{M}_k(\mu_1, \mu_2)) \quad \text{if } v_1 + v_2 = \mu_1 + \mu_2.
\]

We recall that \( \rho(\cdot) \) denotes the spectral radius of a matrix. What prevents us from generalizing this result to the multigrid iteration matrices?

3.3.4. Prove Theorem 3.6 for the case \( \gamma > 2 \).

3.3.5. Demonstrate that (3.73) and (3.74) imply the assertion of Theorem 3.7 about the two-grid method convergence.

3.3.6. Show that for the example of the diffusion equation with \( k(x) = 1 \) and linear finite elements from Section 3.1, the Jacobi method satisfies (3.73) for any \( w \in (0, \frac{1}{T}) \).

3.3.7. Check the assertions (a)–(d) of Lemma 3.10.

### 3.4 Convergence analysis

We apply the results from the previous sections to prove the convergence of V- and W-cycles of the multigrid method for finite element discretization of the Poisson problem:

\[
-\Delta u = f \quad \text{in } \Omega, \quad u|_{\partial \Omega} = 0, \quad \Omega \subset \mathbb{R}^d, \quad d = 2, 3,
\]

where \( \Delta := \sum_{i=1}^d \frac{\partial^2}{\partial x_i^2} \), and \( \Omega \) is a bounded domain. To avoid technical details, we assume that \( \Omega \) is a polygon or polyhedron.
The weak formulation of the problem reads as follows: Find $u \in H_0^1(\Omega)$ satisfying the equality
\[
a(u, v) = (f, v) \quad \forall \, v \in H_0^1(\Omega),
\]
where
\[
a(u, v) := \int_\Omega \sum_{i=1}^2 \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i} \, dx, \quad (f, v) := \int_\Omega f \, v \, dx \quad \forall \, u, v \in H_0^1(\Omega).
\]

Assume that a hierarchy of embedded finite element spaces is given by
\[
U_0 \subset U_1 \subset \cdots \subset U_k \subset \cdots \subset U_l \subset H_0^1(\Omega),
\]
and the corresponding grids are produced using a sequential refinement of a coarsest mesh in such a way that the constant $\beta$ from (3.20) is bounded uniformly with respect to $k$. We call such a refinement procedure regular. Assume that finite element spaces consist of continuous piecewise polynomial functions of a degree not greater than $d$. Moreover, assume that the triangulations are quasi-uniform: The discretization parameter $h$ from (3.20) satisfies the estimates
\[
c_0 2^{-k} \leq h \leq c_1 2^{-k},
\]
$c_0, c_1$ do not depend on $k$, the domain $\Omega$ is such that for any $f \in L^2(\Omega)$ the solution $u$ belongs to $H^2(\Omega)$, and it holds that
\[
\|u\|_{H^2} \leq c(\Omega) \|f\|_{L^2}. \quad (3.86)
\]
The latter assumption is known as the $H^2$-regularity assumption.

The above assumptions are not always satisfied. The embedding $U_{k-1} \subset U_k$ may fail to hold for the case of nonconforming finite elements. If $\Omega$ is a polygon, then the $H^2$-regularity assumption holds only in the case when $\Omega$ is a convex polygon. The condition (3.85) may be violated for grids refined locally in a part of the domain. Nevertheless, in all these cases, multigrid methods can be successfully used, although the analysis of the methods becomes more complicated. Subsections 3.4.4–3.4.7 of this chapter discuss in detail such “more complicated” cases.

We shall often use the $\|\cdot\|_1$-norm defined in (3.2) for functions from $H^1_0(\Omega)$. The norm is equivalent to the standard $H^1$-norm. The first interpolation property in (3.25) can be written as follows: For arbitrary $v \in H^1_0 \cap H^2(\Omega)$ there exists $v_k \in U_k$ such that
\[
\|v - v_k\|_1 \leq c \, h_k \|v\|_{H^1},
\]
\[
(3.87)
\]
Consider the finite element method for the Poisson problem (3.83): Given $f \in L^2(\Omega)$ find $u_k \in U_k$ solving
\[
a(u_k, v_k) = (f, v_k) \quad \forall \, v_k \in U_k.
\]
(3.88)
The error estimates (3.26) and (3.27) imply
\[
\|u - u_k\|_{L^2} \leq c \, h_k^2 \|u\|_{H^1}, \quad \|u - u_k\|_1 \leq c \, h_k \|u\|_{H^1},
\]
\[
(3.89)
\]
Due to (3.86), we get
\[
\|u - u_k\|_{L^2} + h_k \|u - u_k\|_1 \leq c \, h_k^2 \|f\|_{L^2},
\]
\[
(3.90)
\]
where $u$ is the solution to (3.83).
3.4. Convergence analysis

On the very fine grid, the stiffness matrix $A_I$ and the right-hand side $b_I$ of the system are defined from the relation

$$\langle A_I z, y \rangle = a(P_I z, P_I y) \quad \forall \, z, y \in \mathbb{R}_I, \quad \langle b_I, y \rangle = (f, P_I y) \quad \forall \, y \in \mathbb{R}_I. \quad (3.91)$$

We recall that $P_k : \mathbb{R}_k \to \mathbb{U}_k$ is the natural isomorphism between the subspace of coefficients $\mathbb{R}_k$ and the space of the finite element functions $\mathbb{U}_k$. The main property of this isomorphism is the following norm equivalence:

$$c_{\min} b_k \| z \|_2 \leq \| P_k z \|_{L^2} \leq C_{\max} b_k \| z \|_2. \quad (3.92)$$

Our assumptions on the grid and refinement process ensure that the constants $c_{\min}$ and $C_{\max}$ are independent of $k$.

Prolongation and restriction are defined from

$$p_k = P_k^{-1} p_{k-1}, \quad r_k = p_k^s. \quad (3.93)$$

As an example, consider piecewise linear elements on a uniform triangulation of a square as in subsection 3.1.3. For this example, to compute a value of $r_k u_h$ in an internal node of a coarse grid, one needs to sum up the values of $u_h$ in this and neighboring nodes of a finer grid, with weights as shown in Figure 3.8. Here we use uniform triangulation, with the fine mesh size $b_k$ and the coarse mesh size $b_{k-1} = 2b_k$. Since the finite element functions are linear, the values of $p_k u_H$ are computed as follows: If a node belongs to both coarse and fine grids, then one keeps the same nodal values; otherwise one performs the linear interpolation using the given values of $u_H$.

In the next two sections we verify approximation and smoothing properties which are fundamental for the analysis of multigrid iterations and preconditioners. In an abstract setting, the properties were introduced in subsection 3.3.6.

### 3.4.1 The Poisson problem: Approximation property

Our plan for verifying the approximation property is the following. First, the desired property is reformulated in the space of finite element functions $\mathbb{U}_k$. In finite element
spaces, we will apply an approximation result for the solution of the finite element problem. Then we go back to the space of coefficients $\mathbb{R}_k$.

Using the definition of the operator norm, we rewrite the approximation property (3.69) in the form

$$
\frac{\| (A_k^{-1} - p_k A_{k-1}^{-1} r_k) b_k \|_2}{\| b_k \|_2} \leq C_A \| A_k \|^{-1}_2
$$

or

$$
\| z_k - p z_{k-1} \|_2 \leq C_A \| A_k \|^{-1}_2 \| b_k \|_2, \tag{3.94}
$$

where $A_k z_k = b_k, A_{k-1} z_{k-1} = r b_k$, and $b_k$ is an arbitrary element from $\mathbb{R}_k$. Since we are given the natural isomorphism $P_k$ between $\mathbb{R}_k$ and the finite element subspace $U_k$, it is convenient to reduce the proof of (3.94) to the estimate of finite element solutions difference on two subsequent grid levels. This can be done as follows. Consider the adjoint $P^*_k : U_k \to \mathbb{R}_k$, and let $F_k = (P^*_k)^{-1} b_k, F_k \in U_k$, and $u_k \in U_k, u_{k-1} \in U_{k-1}$ be solutions to the finite element problems

$$
a(u_k, \phi) = (F_k, \phi) \quad \forall \phi \in U_k, $$

$$
a(u_{k-1}, \phi) = (F_k, \phi) \quad \forall \phi \in U_{k-1}.
$$

Then the following equalities hold:

$$
z_k = P_k^{-1} u_k, \quad p_k z_{k-1} = P_{k-1}^{-1} u_{k-1}. \tag{3.95}
$$

Let us check this statement. Thanks to the result from Exercise 3.2.2 in Section 3.2 we know that for any $k \leq l$ the matrix $A_{k-1} := r_k A_k p_k$ can be equivalently defined by (3.91) for $z, y \in \mathbb{R}_k$. Hence the first and second equalities in (3.95) follow directly from (3.91).

Thanks to (3.92) the estimate (3.94) would follow from the estimate

$$
\| u_k - u_{k-1} \|_2 \leq C^{-1}_\max \epsilon_{\min} b^2_k C_k \| A_k \|^{-1}_2 \| F_k \|_2. \tag{3.96}
$$

Let us show (3.96). The norm of the difference $u_k - u_{k-1}$ is estimated by exploiting the fact that both finite element solutions $u_k$ and $u_{k-1}$ are “close” to the solution of the differential problem

$$
\| u_k - u_{k-1} \|_2 \leq \| u - u_k \|_2 + \| u - u_{k-1} \|_2,
$$

where $u$ is the solution to the problem

$$
a(u, \phi) = (F_k, \phi) \quad \forall \phi \in H^1_0(\Omega).
$$

We apply the estimate (3.90) in both $U_k$ and $U_{k-1}$ to get

$$
\| u - u_k \|_2 \leq c b^2_k \| F_k \|_2, \quad \| u - u_{k-1} \|_2 \leq c b^2_{k-1} \| F_k \|_2.
$$

Since $b_{k-1} \leq c b_k$, it holds that

$$
\| u_k - u_{k-1} \|_2 \leq c b^2_k \| F_k \|_2. \tag{3.97}
$$

It remains to check the inequality $c \leq \| A_k \|^{-1}_2$. To do this, we use the so-called “inverse” inequality for the finite element functions

$$
\| v_k \|_1 \leq c b^{-1}_k \| v_k \|_2 \quad \forall v_k \in U_k. \tag{3.98}
$$
3.4. Convergence analysis

Constant $c$ in (3.98) depends on particular finite elements (e.g., degree of polynomial functions), but not on $k$. For arbitrary $y \in \mathbb{R}_k$, we get

$$\langle A_k y, y \rangle = a(P_k y, P_k y) = ||P_k y||^2_2 \leq c h^{-2} ||P_k y||^2_L \leq c_0 ||y||^2_2.$$  

Since the matrix $A_k$ is symmetric, the inequality

$$\frac{\langle A_k y, y \rangle}{||y||^2_2} \leq c_0$$

and relations (1.50) yield the estimate on $\lambda_{\text{max}}(A_k)$ and thus on the spectral norm of $A_k$ (all $\lambda(A_k)$ are positive):

$$||A_k||_2 \leq c_0.$$ (3.99)

Thus we can rewrite (3.97) as

$$||u_k - u_{k-1}||_{L^2} \leq c h^2 ||A_k||^{-1}_2 ||F_k||_{L^2}.$$ (3.102)

Therefore we checked inequality (3.92), and the approximation property (3.69) is proved with $C_A = c C_{\text{max}}^{-1}$.

To prove the convergence of the V-cycle, one needs the approximation property in the form (3.74). If one considers the relaxed Jacobi method as smoothing iterations, then $W_k = \frac{1}{\nu} D_k$. To prove (3.74), it is sufficient to check that $||W_k||_2 \leq c ||A_k||_2$. If the relaxation parameter $\nu$ is independent of $k$, this follows from $||D_k||_2 \leq c ||A_k||_2$. Since $D_k$ is a diagonal part of $A_k$, it always holds that $||D_k||_2 \leq ||A_k||_2$. At the end of the next subsection, we also check the bound $||W_k||_2 \leq c ||A_k||_2$ for the symmetric Gauss-Seidel iterations.

3.4.2 The Poisson problem: Smoothing property

In the symmetric case, the standard tool for checking the smoothing property is the following simple lemma.

Lemma 3.11. Assume the matrix $B$ is such that $0 \leq B = B^T \leq I$; then

$$||B(I-B)^\nu||_2 \leq \frac{1}{1+\nu}, \quad \nu \geq 0.$$ (3.100)

Proof. Since $B = B^T$, it holds that

$$||B(I-B)^\nu||_2 = \max_{\lambda} \{\lambda(1-\lambda)^\nu : \lambda \in \text{sp}(B)\}.$$  

From the assumption $\lambda(B) \in [0,1]$, we get

$$||B(I-B)^\nu||_2 \leq \max_{x \in [0,1]} x(1-x)^\nu.$$  

The function $f(x) = x(1-x)^\nu$ attains its maximum in $x_0 = \frac{1}{1+\nu}$. It is easy to check that $f(x_0) \leq \frac{1}{1+\nu}$.  

Let us check the smoothing property of the relaxed Jacobi method and the symmetric Gauss-Seidel method. Let $\phi^{(k)}_i$ be the $i$th local basis function from $U_k$; then the diagonal elements of matrix $A_k$ can be estimated using the definition of matrix $A_k$:

$$(D_k)_{ii} = (A_k)_{ii} = a(\phi^{(k)}_i, \phi^{(k)}_i) = ||\phi^{(k)}_i||^2_2 \geq c_1.$$ (3.101)
Identity \( \| D_k^{-1} \|_2 = \max (D_k)_{ii}^{-1} \leq c_1^{-1} \) and (3.99) yield, for \( c_2 = c_2 c_1^{-1} \),
\[
\| A_k \|_2 \leq c_2 \| D_k^{-1} \|_2^{-1}.
\]
We put the relaxation parameter \( w \) in the Jacobi method as \( w = 1/c_2 \). Since \( W_k = \frac{1}{w} D_k \), statement (e) of Lemma 3.10 yields
\[
0 \leq W_k^{-1} A_k \leq I. \tag{3.102}
\]
In the case of the symmetric Gauss–Seidel method as a smoothing iteration (see (1.52)), it holds that
\[
W_k = (L_k + D_k)D_k^{-1}(D_k + L_k^T) = A_k + L_k D_k^{-1} I_k^T \geq A_k.
\]
This proves (3.102) for the symmetric Gauss–Seidel method as well.

Estimate (3.102) is the necessary smoothing property for the V-cycle convergence analysis.

Now we check the inequality
\[
\| W_k \|_2 \leq c \| A_k \|_2. \tag{3.103}
\]
For the Jacobi method, (3.103) was shown in the previous subsection. For the symmetric Gauss–Seidel method, the inequality (3.103) is verified by the following arguments. The support of any basis function from \( U_k \) has a nonempty intersection with a finite number of other basis functions supports. This number is independent of the level \( k \) and coincides with the number of nonzero entries in the corresponding row of the matrix \( A_k \). Hence it holds that
\[
\| L_k \|_2^2 \leq \| L_k \|_1 \| L_k \|_\infty = \left( \max_{j} \left\{ n_k \sum_{i=j+1}^{n_k} |(A_k)_{ij}| \right\} \right) \left( \max_{i,j} \left\{ |(A_k)_{ij}| \right\} \right)
\leq c \max_{i,j} (A_k)_{ij}^2 \leq c \| A_k \|_2^2. \tag{3.104}
\]
Therefore,
\[
\| W_k \|_2 = \| A_k + L_k D_k^{-1} I_k^T \|_2 \leq \| A_k \|_2 + \| L_k \|_2 \| D_k \|^{-1}_2 \leq c \| A_k \|_2.
\]

Below we demonstrate the necessary smoothing property for the W-cycle. More precisely, we show the estimate
\[
\| A_k S_k^w \|_2 \leq \frac{c}{\nu + 1} \| A_k \|_2. \tag{3.105}
\]
Denote \( B = W_k^{-\frac{1}{2}} A_k W_k^{-\frac{1}{2}} \), then \( \text{sp}(B) = \text{sp}(W_k^{-1} A_k) \subset [0, 1] \). Moreover, one can reorganize matrices in the product as
\[
A_k (I - W_k^{-1} A_k) = W_k^{\frac{1}{2}} (W_k^{-\frac{1}{2}} A_k W_k^{-\frac{1}{2}}) (W_k^{\frac{1}{2}} (I - W_k^{-1} A_k)^{\nu} W_k^{\frac{1}{\nu}}) W_k^{\frac{1}{2}}
\]
\[
= W_k^{\frac{1}{2}} B (I - W_k^{-\frac{1}{2}} A_k W_k^{\frac{1}{2}})^{\nu} W_k^{\frac{1}{\nu}}
\]
\[
= W_k^{\frac{1}{2}} B (I - B)^{\nu} W_k^{\frac{1}{\nu}}.
\]
Since \( S_k = I - W_k^{-1} A_k \), it holds that
\[
\| A_k S_k^w \|_2 = \| W_k^{\frac{1}{2}} B (I - B)^{\nu} W_k^{\frac{1}{\nu}} \|_2 \leq \| W_k \|_2 \| B (I - B)^{\nu} \|_2
\leq \frac{1}{1 + \nu} \| W_k \|_2 \leq \frac{c}{1 + \nu} \| A_k \|_2. \tag{3.106}
\]
3.4. Convergence analysis

We have checked the necessary conditions for Theorems 3.6–3.9. Therefore, for the finite element approximation of the Poisson problem, we prove the convergence of the multigrid V-cycle iteration, with an equal number of pre- and postsmoothings, and the convergence of the multigrid W-cycle iteration, with a sufficiently large number of pre- or postsmoothings. As we know from subsection 1.4.13, this also implies that the corresponding multigrid preconditioners deliver a bound of the condition number for preconditioned systems, which is independent of the number of multigrid levels \( k \), and thus of the number of unknowns.

**Smoothing property: Nonsymmetric case**

If \( A_k \) or \( W_k \) is a nonsymmetric matrix, then Lemma 3.11 is not an appropriate tool for checking the smoothing property. Such a situation occurs if the Gauss–Seidel or the SOR method is used for smoothing iterations. Also \( A_k \) can be nonsymmetric if the corresponding differential problem operator is not self-adjoint. One example of such a problem is the convection-diffusion problem from subsection 5.1.8. Moreover, one can hardly use Lemma 3.11 if the smoothing property is checked in a norm different from the spectral norm.

In the nonsymmetric case, the appropriate tool can be the following lemma and corollary.

**Lemma 3.12 (Reusken).** Assume that the matrix \( B \in \mathbb{R}^{n \times n} \) satisfies \( \| B \| \leq 1 \) in some operator norm. Then in the same norm, it holds that

\[
\| (I - B)(I + B)^{\nu} \| \leq 2^{\nu+1} \sqrt{\frac{2}{\pi \nu}}, \quad \nu = 1, 2, 3, \ldots \tag{3.107}
\]

The proof of the lemma can be found in [97]. The immediate consequence is a smoothing property for iteration methods with iteration matrices

\[
S_k = I - W_k^{-1} A_k.
\]

Indeed, in Lemma 3.12 we set \( B = I - 2W_k^{-1} A_k \) and note that with this definition of \( B \) it holds that

\[
\| A_k S_k \| = \left( \frac{1}{2} \right)^{\nu+1} \| W_k (I - B)(I + B)^{\nu} \|.
\]

We get the following result.

**Theorem 3.13.** Assume that in some norm it holds that

\[
\| I - 2W_k^{-1} A_k \| \leq 1 \quad \text{and} \quad \| W_k \| \leq C \| A_k \|,
\]

and then the following smoothing property is valid:

\[
\| A_k S_k \| \leq C \sqrt{\frac{2}{\pi \nu} \| A_k \|}.
\]

The theorem is used below to prove the smoothing property of the SOR method with a relaxation parameter applied for the discrete Poisson problem. As before, we consider linear finite elements. The estimates (3.99) and (3.101) imply

\[
\langle A_k^2 x, x \rangle \leq c \langle D_k^2 A_k x, x \rangle \quad \forall x \in \mathbb{R}_k. \tag{3.108}
\]
Similarly, from (3.104) and (3.101) it follows that
\[ \langle L_k A_k x, x \rangle \leq \epsilon \langle D_k A_k x, x \rangle \quad \forall x \in \mathbb{R}_k. \quad (3.109) \]

Consider the SOR method with the parameter \( \omega > 0 \). The iteration matrix on grid level \( k \) takes the form
\[ S_k = I - \omega (\omega L_k + D_k)^{-1} A_k. \]
To apply Theorem 3.13, we shall check the bound
\[ \| I - 2\omega (\omega L_k + D_k)^{-1} A_k \|_2 \leq 1. \quad (3.110) \]
To do this, we repeat the arguments from the proof of Theorem 1.3 and get the following equality for arbitrary vectors \( x \in \mathbb{R}_k \) and \( v = (\omega L_k + D_k)^{-1} A_k x \):
\[ \| (I - 2\omega (\omega L_k + D_k)^{-1} A_k) x \|^2_2 = \| x \|^2_2 - 2\omega \langle (A_k^{-1} (\omega L_k + D_k) - 2\omega I) v, v \rangle. \]
We want to prove that the second term is positive for sufficiently small, but independent of \( k \), values of \( \omega \). After another substitution \( v = Ay \), we see that it is sufficient to show the inequality
\[ \langle ((\omega L_k + D_k) A_k^{-1} - 2\omega A_k^2) y, y \rangle \geq 0. \]
This inequality follows from (3.109) and (3.108) for sufficiently small \( \omega \). Thus, we have proved
\[ \| (I - 2\omega (\omega L_k + D_k)^{-1} A_k) x \|^2_2 \leq \| x \|^2_2. \]
Hence, (3.110) follows by the definition of the spectral norm.

The second estimate necessary for the application of Theorem 3.13 is
\[ \| (\omega L_k + D_k) \|_2 \leq C \| A_k \|_2. \]
This estimate is trivial to show, since \( \| A_k \|_2 \geq \| D_k \|_2 \geq c_1 \| L_k \|_2 \). Now the theorem implies the smoothing property of the SOR iteration in the spectral norm.

### 3.4.3 Less standard cases

In practice, multigrid methods are successfully used to solve more complicated problems than the one resulting from conforming finite elements applied to the Poisson equations in a polygon domain. If one considers a more general elliptic self-adjoint problem with smooth bounded coefficients, then the arguments from the previous sections can be repeated without any significant alteration. Naturally, “grid level independent” constants from approximation and smoothing properties would depend on estimates for problem coefficients and their derivatives. We will not consider such an extension of the analysis here, since it does not bring many new ideas or technical difficulties. In this section, we review what is known about efficiency of multigrid methods in the following two interesting (and practically important) cases:

(i) An elliptic differential equation does not possess the full regularity property, i.e., its solution is not in \( H^2(\Omega) \), if the right-hand side of this equation is in \( L^2(\Omega) \).

(ii) The finite element spaces \( U_k \) do not necessarily form a hierarchy of embedded spaces, \( U_{k-1} \not\subset U_k \).

We will continue to consider the Poisson problem as the model elliptic equation. However, it can be posed in a geometrically different domain and discretized by non-conforming finite elements. Several examples of equations different from the Poisson problem will be considered in Chapter 5.
3.4. Convergence analysis

3.4.4 The case of less than full regularity

One may distinguish two types of difficulties while extending multigrid methods for solving more complex problems. In the first case, one encounters difficulties in analysis, while in practice a multigrid method with simple smoothing iterations and canonical prolongation-restriction operators (as defined in Section 3.2) demonstrates fast convergence with a factor independent of the number of grid levels. Only proofs of convergence estimates may become more involved. However, there are situations when one needs nonstandard ingredients such as: special smoothers, noncanonical projections and restrictions, or coarse grid operators, to make a multigrid method to converge as fast as in a “simple” case.

The problem studied in this section leads to the first type of difficulties, i.e., the difficulties of analysis. We are interested in multigrid preconditioners for domains with entering corners. The examples are L-shape domain \( \Omega_a = (0,1)^2 \setminus (\frac{1}{2},1)^2 \) or the domain with a cut, \( \Omega_b = (0,1)^2 \setminus (\frac{1}{2},1) \times (\frac{1}{2},\frac{3}{2}) \) (cf. Figure 3.9).

Consider the Poisson equations in domains \( \Omega_a \) or \( \Omega_b \). One easily builds a hierarchy of triangulations for both domains, and proceeds by using the conformal finite element method to discretize the equations. Finally, to solve resulting discrete systems one may consider a Krylov subspace iterations with a multigrid preconditioner. The multigrid preconditioner uses relaxed Jacobi smoothing iterations and canonical prolongation and restriction operators. Does this result in a convergent iterative method with the convergence factor bounded away from 1 and independent of the number of grid levels? We cannot apply the analysis of the previous sections, since in domains with entering corners the \( H^2 \)-regularity property does not hold, and the estimate (3.86) is no longer valid. This estimate played an important role in the previous analysis. Nevertheless, the answer to the question is positive and a proof was given, for example, in [34]. In that paper, the convergence of multigrid methods was proved for variational problems (finite elements are the particular case) under rather general assumptions; see (3.119)–(3.121) below. We will sketch the proof of the W-cycle convergence with a sufficiently large, yet independent of the number of grid levels, number of smoothing steps. Similar assumptions on the number of unknowns were used in formulations of Theorems 3.6 and 3.7, although a more abstract analysis from [34] guarantees that only one smoothing step on every grid level is enough.

While the \( H^2 \)-regularity estimate (3.86) for the Poisson problem is no longer valid, the following weaker regularity estimates hold (one may consider more general elliptic problems with smooth coefficients):

\[
\| \mu \|_{1+\alpha} \leq C(\Omega)\| f \|_{-1+\alpha},
\]  
(3.111)
with some \( \alpha \in (0, 1) \). The norm \( \| \cdot \|_{1+\alpha} \) can be defined through the relation
\[
\|u\|_{1+\alpha}^2 := \|u\|_1^2 + \|\nabla u\|_2^2,
\]
with \( \|u\|_1 = \int_\Omega \int_\Omega \frac{(u(x) - u(y))^2}{|x-y|^{2+2\alpha}} \, dx \, dy \).

\( \| \cdot \|_{-1+\alpha} \) is the dual norm for \( \| \cdot \|_{1-\alpha} \) with respect to the \( L^2 \)-inner product:
\[
\|f\|_{-1+\alpha} := \sup_{u \in C^0(\Omega)} \frac{(f, u)}{\|u\|_{1-\alpha}}.
\]

It is well known (see [94]) that for a domain with internal angles, it holds that \( \alpha \in (\frac{1}{2}, 1) \), and for a \( \Omega_k \)-type domain with a cut it, \( \alpha \in (0, \frac{1}{2}) \) holds.

The estimate (3.111) is sufficient for proving the convergence of the finite element solution \( u_h \) to the solution \( u \) of the differential problem. However, the convergence order in the \( L^2 \)-norm in this case is lower compared to \( O(h^2) \) for the case of \( H^2 \)-regularity. The following convergence result is known to be true (compare to (3.90)):
\[
\|u - u_h\|_{1-\alpha} \leq c \, h^{2\alpha} \|f\|_{-1+\alpha}.
\]  \( \text{(3.112)} \)

We recall that \( A_k \) denotes the stiffness matrix on the grid level \( k \), and \( P_k : \mathbb{R}_k \rightarrow \mathbb{U}_k \) is the natural isomorphism between the space of coefficients \( \mathbb{R}_k \) and the space of vector functions \( \mathbb{U}_k \). The next proposition is proved in [13]. There exist constants \( c_2, c_1 \), independent of \( k \), such that for any \( \alpha \in [0, 1] \) it holds that
\[
c_2 h_{k}^{1-\alpha} \| A_{k}^{1/2} z \|_2 \leq \| P_k z \|_2 \leq c_1 h_{k}^{1-\alpha} \| A_{k}^{1/2} z \|_2 \quad \forall \, z \in \mathbb{R}_k.
\]  \( \text{(3.113)} \)

Now we estimate the norm of the two-grid method iteration matrix \( \tilde{M}_k(\nu, 0) \). It is convenient to consider the matrix norm induced by the vector norm \( \| \cdot \|_{1-\alpha} := \| A_{k}^{1/\alpha} \cdot \| \) (instead of the spectral norm). Consider the following decomposition and estimate:
\[
\| \tilde{M}_k(\nu, 0) \|_{1-\alpha} = \| A_k^{1/\alpha} (I - P_k A_{k-1}^{-1} r_k A_k) S_k A_k^{1/\alpha} \|_2
\leq \| A_k^{1/\alpha} (I - P_k A_{k-1}^{-1} r_k) A_k^{1/\alpha} \|_2 \| A_k^{1/\alpha} S_k A_k^{1/\alpha} \|_2.
\]  \( \text{(3.114)} \)

Let us check the approximation property of the form
\[
\| A_k^{1/\alpha} (I - P_k A_{k-1}^{-1} r_k) A_k^{1/\alpha} \|_2 \leq C_4 \| A_k \|_2^{-\alpha}
\]  \( \text{(3.115)} \)
and the smoothing property
\[
\| A_k^{1/\alpha} S_k A_k^{1/\alpha} \|_2 \leq C_5 \| A_k \|_2^\alpha \left( \frac{1}{\nu} \right)^{\alpha}.
\]  \( \text{(3.116)} \)

For a large enough number of smoothing steps \( \nu \), the decomposition (3.114) and smoothing and approximation properties (3.115)–(3.116) immediately yield the estimate
\[
\| \tilde{M}_k(\nu, 0) \|_{1-\alpha} \leq \xi < 1,
\]
with an arbitrary small \( \xi \) independent of \( k \). The convergence of the W-cycle follows from Theorem 3.6 and the efficiency of the multigrid preconditioner follows from Theorem 1.33.
The required approximation property is proved as follows. The estimate (3.115) is equivalent to
\[ \| A_k (A_k^{-1} - p A_{k-1}^{-1} r) b_k \|_2 \leq C_k \| A_k \|_{\alpha,2}^{-1} \| A_k^{\frac{3}{2}} b_k \|_2 \quad \forall \ b_k \in \mathbb{R}_k. \] (3.117)
We already showed (see (3.99)) the inequality
\[ \| A_k \|_{\alpha,2}^{-1} \geq c. \]
Repeating the arguments from subsection 3.4.1 and using the upper bound in (3.113), we conclude that (3.117) follows from
\[ \| u - u_k \|_{1-z} \leq c b_k^2 \| A_k^{\frac{3}{2}} b_k \|_2, \]
where \( u \) and \( u_k \) are the solutions to the differential and finite element problems with the right-hand side \( F_k = (P_k)^{-1} b_k \) from \( \mathbb{U}_k \). Due to the inequality (3.112), it remains to check that the norm \( \| F_k \|_{1-z} \) is estimated from above by \( c b_k^2 \| A_k^{\frac{3}{2}} b_k \|_2 \). This follows from the lower estimate in (3.113).

Assume that our smoothing steps are some simple iterations with a symmetric pre-conditioner \( W_k \), for example, the damped Jacobi method. The smoothing property (3.116) is verified by arguments similar to those in subsection 3.4.2. The only difference is that instead of (3.100) one uses the estimate
\[ \| B^x (I - B)^y \|_2 \leq \left( \frac{1}{1 + \nu} \right)^x \quad \forall \nu \geq 0, \ x \in [0,1] \] (3.118)
for arbitrary matrix \( B \) such that \( 0 \leq B = B^T \leq I \). The proof of (3.118) is similar to the proof of Lemma 3.11.

The analysis of the V-cycle convergence is more involved; see [44].

In practice, the performance of multigrid methods for problems posed in domains with re-entering corners and cuts can slightly deteriorate. In [96] one finds the following estimates for convergence factors of a multigrid method for the Poisson problem. The estimates were found from numerical experiments: \( q \in [0.044,0.052] \) if \( \Omega \) is a square, \( q \in [0.073,0.085] \) if \( \Omega \) is the L-shaped domain (see Figure 3.9(a)), \( q \in [0.096,0.116] \) if \( \Omega \) has a corner of \( 3 \pi/2 \), and \( q \in [0.113,0.147] \) if \( \Omega \) is the domain with a cut as in Figure 3.9(b).

### 3.4.5 Variational framework

Finally, we briefly consider another approach to analyzing multigrid methods. It does not use the regularity estimate (3.111) for solutions to elliptic equations. In this approach, conditions sufficient for the convergence of multigrid methods are formulated in terms of properties of the bilinear form \( a(\cdot,\cdot) \) from the finite element method setup; see (3.84) and (3.88). We formulate the assumptions below. Denote by \( (\cdot,\cdot) \) some inner products on \( \mathbb{R}_k \), and assume the bilinear form \( a(\cdot,\cdot) \) is symmetric positive definite. On every grid level the matrices \( A_k \) are defined through \( (A_k p_k^{-1} u_k, p_k^{-1} v_k) = a(u_k, v_k) \), and \( \lambda_k \) is the largest eigenvalue of \( A_k \). Assume that there exist linear projection operators \( q_k : \mathbb{U}_l \rightarrow \mathbb{U}_k \) such that for arbitrary \( u_l \in \mathbb{U}_l \) it holds that
\[ \|(q_k - q_{k-1}) u_l \|_2^2 \leq c_1 \lambda_k^{-1} a(u_l, u_l), \quad k = 1,\ldots,l, \] (3.119)
\[ a(q_k u_l, q_k u_l) \leq c_2 a(u_l, u_l), \quad k = 0,\ldots,l-1, \] (3.120)
with constants $c_1$, $c_2$ independent of $k$. Assumptions (3.119)–(3.120) are complemented with the smoothing property (3.73). The preconditioner from smoothing iterations $W_k$ is self-adjoint positive definite with respect to the scalar product $(\cdot, \cdot)_k$ and $W_k \geq A_k$. Moreover, we assume the estimates

$$
\|u_k\|_k^2 \leq c_3 A_k (W_k^{-1} u_k, u_k)_k, \quad k = 1, \ldots, l, \quad u_k \in U_k.
$$

(3.121)

Assumptions (3.119)–(3.121) are sufficient for the following upper bound on the norm of V- and W-cycles iteration matrices to be valid [34]:

$$
\|M_l(v_1, v_2)\|_A \leq 1 - (c I)^{-1}.
$$

Generally speaking, this estimate is weaker compared to those proved earlier in this chapter. It does not account for the number of smoothing iterations and depends on the number of grid levels. The advantage of this variational approach is weaker requirements on the corresponding differential problem. In particular, one can prove the convergence of multigrid methods applied to solve elliptic problems with discontinuous coefficients. Moreover, it appears that it is sufficient to require the property (3.121) only for $u_l \in \text{range}(q_k - q_{k+1})$ rather than for all functions from $U_k$. The latter helps to prove the convergence of multigrid methods in those cases when grid refinement is done only locally (not in a whole domain $\Omega$), and thus smoothing iterations on a grid level $k$ are performed only in those parts of a domain where new basis functions are added while passing from level $k - 1$ to level $k$. We will consider local grid refinement in subsection 4.2.4.

### 3.4.6 Nonconforming finite elements and $U_{k-1} \not\subseteq U_k$

In the previous sections, we considered applications of multigrid methods when the underlying finite element spaces form the hierarchy of embedded spaces, $U_{k-1} \subseteq U_k$. In practice, this is not always the case. One may encounter the situation of $U_{k-1} \not\subseteq U_k$ using

nonconforming finite elements, or approximating curvilinear boundaries of a domain, or using certain mesh refinement strategies or in some other cases. We recall that the embedding of $U_{k-1}$ into $U_k$ naturally induces the prolongation operator $p : \mathbb{R}_{k-1} \to \mathbb{R}_k$. The projection operator $r : \mathbb{R}_k \to \mathbb{R}_{k-1}$ was taken as adjoint to $p$. In the case of nonembedded spaces $U_{k-1}$ and $U_k$, one faces the question of defining the operators $p$ and $r$. It appears that a proper choice of $p$ and $r$ is important for the convergence and efficiency of multigrid methods. This operators should satisfy certain conditions.

Further in this section, we discuss nonconforming finite elements. A finite element method is called nonconforming if the space of finite element functions $U_k$ is not a subspace of $H$, a functional space for weak solutions of a differential problem. In the case of second order elliptic equations, this means typically $U_k \not\subseteq H^1(\Omega)$. Such discretizations are widely used in computational mechanics. In particular, simple stable finite element pairs for the Stokes problem (see details in subsection 5.2.1) include nonconforming elements. In this section, we consider a few examples of nonconforming finite elements and further study multigrid methods for solving the resulting systems of equations.

As the simplest nonconforming finite element, one may consider the Crouzeix–Raviart element, also known as the nonconforming $P_1$ element. To define the space $U_k$ of the Crouzeix–Raviart elements, we assume a triangulation $\mathcal{T}$ of a domain $\Omega$. On each $T \in \mathcal{T}$, a function $v \in U_k$ is linear; however, instead of the global continuity condition in $\Omega$ (which was assumed for conforming elements) one assumes another condition for $v \in U_k$. To formulate it, consider two triangles $T_1$ and $T_2$ from $\mathcal{T}$ sharing an edge, with $E$ denoting the center of this edge (see Figure 3.10(a)), then one assumes $v|_{T_1} = v|_{T_2}$.
3.4. Convergence analysis

\( v|_{T_1} \) denotes a function \( v \) on \( T_1 \) extended by continuity to \( \partial T_1 \); similarly, \( v|_{T_2} \) is defined for the second triangle. If an edge of \( T \in \mathcal{T} \) belongs to \( \partial \Omega \), then one sets \( v|_{T} \) in the middle point of the edge equal to the Dirichlet boundary condition or does not prescribe any condition for the Neumann case. A slightly different way to define the Crouzeix–Raviart element is to require the equality of integral averages of \( v|_{T_1} \) and \( v|_{T_2} \) over an edge shared by \( T_1 \) and \( T_2 \). One easily verifies that both definitions lead to the same space of finite element functions \( \mathbb{U}_b \); see Exercise 3.4.7. In the three-dimensional case, the Crouzeix–Raviart element is defined in a similar way with the only difference being that instead of matching finite element functions values in the middle points of edges, one requires the continuity in the barycenters of faces shared by tetrahedrons.

The rectangle analogue of the Crouzeix–Raviart element is the nonconforming \( Q_2 \) element. On each rectangle, a function \( v \) from \( \mathbb{U}_b \) is the linear combination of \( \{1, x_1, y_1, x_1^2 - y_1^2\} \), where the coordinate axes \( x_1 \) and \( y_1 \) are parallel to the rectangle midlines. If an edge is shared by two rectangles, then conditions similar to the "triangle case" from above are imposed. We note that now the point and integral matching conditions define different spaces \( \mathbb{U}_b \). As the Crouzeix–Raviart element, the nonconforming \( Q_2 \) element was initially introduced for the needs of computational fluid dynamics [166].

One example of a higher order nonconforming finite element is the Wilson element. In the two-dimensional case, the domain \( \Omega \) is divided into rectangles. On each \( T \in \mathcal{T} \), a finite element function from \( \mathbb{U}_b \) is a polynomial of degree 2. If two rectangles \( T_1 \) and \( T_2 \) share a vertex, then the values of \( v|_{T_1} \) and \( v|_{T_2} \) coincide in this vertex. If a vertex of \( T \) lies on the boundary of domain, then \( v|_{T} \) takes a boundary value in the vertex or an arbitrary value in the case of the Neumann conditions. To fix the remainder degrees of freedom for \( v \) on each element \( T \), one prescribes values for \( \int_T v_{xx} \, dx \, dy \) and \( \int_T v_{yy} \, dx \, dy \); in Figure 3.11 this degree of freedom is denoted by crosses. One can define the Wilson element in the three-dimensional case as well.

We consider the use of nonconforming finite elements for the model example of the Poisson problem with Dirichlet boundary conditions. Recall the weak formulation of the problem: Find \( u \in H^1_0(\Omega) \), satisfying

\[
\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f \, v \, dx \quad \forall \ v \in H^1_0(\Omega).
\]

The bilinear form on the left side of the equality is not well-suited for nonconforming finite elements, since functions from \( \mathbb{U}_b \) are discontinuous across the edges of triangulation. This difficulty is overcome by considering the identity

\[
\int_\Omega \nabla u \cdot \nabla v \, dx = \sum_{i \in \mathcal{T}} \int_{T_i} \nabla u \cdot \nabla v \, dx \quad \text{for} \ u, v \in H^1(\Omega)
\]
and for a given triangulation \( \mathcal{T} \). The bilinear form on the right-hand side of the equality makes sense if \( u \) and \( v \) are smooth only in the interior of every element \( t \). This bilinear form is used for the definition of nonconforming finite element methods. For any given triangulation \( \mathcal{T} \), the finite element solution \( u \in \mathbb{U}_h \) is defined by the identity

\[
\sum_{t \in \mathcal{T}} \int_t \nabla u_h \cdot \nabla v_h \, dx = \int_{\Omega} f \, v_h \, dx \quad \forall \, v_h \in \mathbb{U}_h.
\]  

(3.122)

If certain regularity assumptions hold for the domain \( \Omega \), then the error of the method obeys the estimate similar to the one in the conforming case

\[
\|u - u_h\| + b \left( \sum_{t \in \mathcal{T}} \|\nabla (u - u_h)\|_{L^2(t)} \right)^{\frac{1}{2}} \leq c \, b^2 \|u\|_{H^2(\Omega)}.
\]  

(3.123)

### 3.4.7 Multigrid for nonconforming elements

Now we shall discuss multigrid methods for systems resulting from nonconforming discretizations. Assume we are given a hierarchy of grids \( \mathcal{T}_k \), which was obtained by a sequential application of a mesh refining procedure; see Figures 3.10 and 3.11. This defines a hierarchy of, in general, nonembedded spaces \( \mathbb{U}_k \). Similarly to what we have done before, \( P_k : \mathbb{R}_k \rightarrow \mathbb{U}_k \) is the natural isomorphism between the space of finite element functions \( \mathbb{U}_k \) and the space of coefficients of these functions decompositions with respect to a (nodal) basis. Every triangulation \( \mathcal{T}_k \) defines the bilinear form

\[
a_k(u_k, v_k) = \sum_{t \in \mathcal{T}_k} \int_t \nabla u_k \cdot \nabla v_k \, dx, \quad u_k, v_k \in \mathbb{U}_k.
\]

The stiffness matrix is defined on every grid level by the identity

\[
\langle A_k z, y \rangle = a_k(P_k z, P_k y) \quad \forall \, z, y \in \mathbb{R}_k.
\]

To set up an iteration of a multigrid method for the solution of a linear algebraic system with matrix \( A_k \), we need to define a prolongation operator \( p \) from \( \mathbb{R}_{k-1} \) to \( \mathbb{R}_k \) and a projection operator \( r : R_k \rightarrow R_{k-1} \). In a conforming case the prolongation is simply induced by the embedding \( \mathbb{U}_{k-1} \subset \mathbb{U}_k \), i.e., one defines \( p = P_k^{-1} P_{k-1} \). In a nonconforming
3.4. Convergence analysis

Let the lack of such an embedding requires some extension of this approach to define the prolongation. For this purpose, we define a special mapping \( \varphi : \mathbb{U}_{k-1} \to \mathbb{U}_k \) and set

\[
p = P^{-1}_k \varphi P_{k-1}.
\]

Similarly to the conforming case, the restriction is the adjoint of the prolongation, \( r = p^* \). We define \( \phi \), following ideas from [32]. To this end, suppose that a space \( \mathbb{V}_k \subset L^2(\Omega) \) is such that \( \mathbb{U}_k + \mathbb{U}_{k-1} \subset \mathbb{V}_k \). Assume a mapping \( \varphi : \mathbb{V}_k \to \mathbb{U}_k \) is defined such that

\[
||\varphi||_{1,\mathbb{V}_k} \leq C_{\varphi}, \quad \varphi|_{\mathbb{U}_k} = id,
\]

where \( id \) is the identity mapping and the constant \( C_{\varphi} \) is independent of \( k \). The above assumptions for \( A_k, A_{k-1}, p, \) and \( r \) are sufficient for the approximation property to hold. Complemented by appropriate smoothing iterations, this property implies the convergence of the W-cycle. The proof of the approximation property relies, similarly to the conforming case, on the finite element method error estimate (3.123). Indeed, with the same arguments as in subsection 3.4.1, one verifies that the approximation property, that is, the estimate

\[
\frac{||(A^{-1}_k - pA^{-1}_{k-1})b_k||_2}{||b_k||_2} \leq C_{A}||A_k||_2^{-1} \quad \forall \ b_k \in \mathbb{R}_k,
\]

is equivalent to the estimate

\[
||P^{-1}_k(u_k - \varphi u_{k-1})||_2 \leq C_{A}||A_k||_2^{-1}||b_k||_2,
\]

where \( u_k \in \mathbb{U}_k \) and \( u_{k-1} \in \mathbb{U}_{k-1} \) are the solutions to the problems

\[
as_k(u_k, \xi) = (F_k, \xi) \quad \forall \ \xi \in \mathbb{U}_k,
\]

\[
as_{k-1}(u_{k-1}, \xi) = (\varphi F_k, \xi) \quad \forall \ \xi \in \mathbb{U}_{k-1},
\]

with \( F_k = (P_k^*)^{-1}b_k \). Now we use that \( \varphi \varphi_k = \varphi_k \) for any \( \varphi_k \in \mathbb{U}_k \) to obtain

\[
||P^{-1}_k(u_k - \varphi u_{k-1})||_2 = ||P^{-1}_k \varphi (u_k - u_{k-1})||_2 \leq C_{\min}C_{\varphi}||u_k - u_{k-1}||_{L^2};
\]

and

\[
(F_k, \xi) = (F_k, \varphi \xi) = (\varphi F_k, \xi).
\]

Therefore, the proof of the approximation property is reduced to verifying the estimate between the difference of two finite element solutions \( u_k \) and \( u_{k-1} \), with the right-hand side source term \( \varphi F_k \in L^2(\Omega) \). Note that we can estimate the \( L^2 \) norm of the right-hand side,

\[
||\varphi F_k||_{L^2} \leq C_{\varphi}||F_k||_{L^2} = C_{\varphi}||P^{-1}_k b_k||_{L^2} \leq C_{\max}C_{\varphi}||b_k||_2.
\]

It remains to estimate the norm of the difference \( u_k - u_{k-1} \) via the finite element errors of \( u_k \) or \( u_{k-1} \) to the solution of the differential problem, with the same right-hand side source term,

\[
||u_k - u_{k-1}||_{L^2} \leq ||u - u_{k-1}||_{L^2} + ||u - u_{k-1}||_{L^2}.
\]

Applying the error estimate (3.123),

\[
||u - u_{k-1}||_{L^2} \leq c b_k^2 ||\varphi F_k||_{L^2},
\]

\[
\frac{||(A^{-1}_k - pA^{-1}_{k-1})b_k||_2}{||b_k||_2} \leq C_{A}||A_k||_2^{-1} \quad \forall \ b_k \in \mathbb{R}_k,
\]

is equivalent to the estimate

\[
||P^{-1}_k(u_k - \varphi u_{k-1})||_2 \leq C_{A}||A_k||_2^{-1}||b_k||_2,
\]

where \( u_k \in \mathbb{U}_k \) and \( u_{k-1} \in \mathbb{U}_{k-1} \) are the solutions to the problems

\[
as_k(u_k, \xi) = (F_k, \xi) \quad \forall \ \xi \in \mathbb{U}_k,
\]

\[
as_{k-1}(u_{k-1}, \xi) = (\varphi F_k, \xi) \quad \forall \ \xi \in \mathbb{U}_{k-1},
\]

with \( F_k = (P_k^*)^{-1}b_k \). Now we use that \( \varphi \varphi_k = \varphi_k \) for any \( \varphi_k \in \mathbb{U}_k \) to obtain

\[
||P^{-1}_k(u_k - \varphi u_{k-1})||_2 = ||P^{-1}_k \varphi (u_k - u_{k-1})||_2 \leq C_{\min}C_{\varphi}||u_k - u_{k-1}||_{L^2};
\]

and

\[
(F_k, \xi) = (F_k, \varphi \xi) = (\varphi F_k, \xi).
\]

Therefore, the proof of the approximation property is reduced to verifying the estimate between the difference of two finite element solutions \( u_k \) and \( u_{k-1} \), with the right-hand side source term \( \varphi F_k \in L^2(\Omega) \). Note that we can estimate the \( L^2 \)-norm of the right-hand side,

\[
||\varphi F_k||_{L^2} \leq C_{\varphi}||F_k||_{L^2} = C_{\varphi}||P^{-1}_k b_k||_{L^2} \leq C_{\max}C_{\varphi}||b_k||_2.
\]

It remains to estimate the norm of the difference \( u_k - u_{k-1} \) via the finite element errors of \( u_k \) or \( u_{k-1} \) to the solution of the differential problem, with the same right-hand side source term,

\[
||u_k - u_{k-1}||_{L^2} \leq ||u - u_{k-1}||_{L^2} + ||u - u_{k-1}||_{L^2}.
\]

Applying the error estimate (3.123),

\[
||u - u_{k-1}||_{L^2} \leq c b_k^2 ||\varphi F_k||_{L^2},
\]

\[
||u_k - u_{k-1}||_{L^2} \leq ||u - u_{k-1}||_{L^2} + ||u - u_{k-1}||_{L^2}.
\]

Applying the error estimate (3.123),

\[
||u - u_{k-1}||_{L^2} \leq c b_k^2 ||\varphi F_k||_{L^2},
\]
and, noting that \( h_k \leq \epsilon_\delta \), we get

\[
\|u - u_{k-1}\|_{L^2} \leq c \epsilon_\delta h_k^2 \|\phi^* T_k\|_{L^2}.
\]

Finally, similar to the conforming case, we check

\[
\|A_k\|_2 = \epsilon_\delta h_k^{-2}.
\]

Summarizing the above results, we prove the approximation property (3.125) with the constant \( C_\delta = \epsilon_\delta^{-1} c_\delta (c_\delta + 1) c_{\min}^{-1} C_{\max} C_\phi^2 \).

Now we show how to define the required space \( \mathcal{V}_k \) and the mapping \( \phi \), satisfying the condition (3.124). We note that the proof of V-cycle convergence by the existing theory requires a bound of \( \phi \) in the energy norm, which is harder to show. Moreover, if a problem is not \( H^2 \)-regular, then the convergence theory requires a bound in an \( H^2 \)-norm, \( \alpha \in (0, 1) \); see [43].

As an example we consider the Crouzeix-Raviart element and homogeneous Dirichlet boundary conditions. Assume that the triangulation \( \mathcal{T}_k \) on level \( k \) is built from \( \mathcal{T}_{k-1} \) by connecting the midpoints of all triangles; cf. Figure 3.10. For the space \( \mathcal{V}_k \), we consider the space of functions linear on every triangle from \( \mathcal{T}_k \). No restrictions on \( \phi \in \mathcal{V}_k \) are imposed on the edges of the triangulation. Clearly, \( \mathcal{U}_k \subset \mathcal{V}_k \) and \( \mathcal{U}_{k-1} \subset \mathcal{V}_k \). The mapping \( \phi : \mathcal{V}_k \to \mathcal{U}_k \) is defined as follows. For any two triangles \( t_1, t_2 \in \mathcal{T}_k \) sharing a common edge, denote by \( a \) the middle point of this edge. Then for any fixed \( \psi_k \in \mathcal{V}_k \) the function \( \phi \psi_k \) is the piecewise linear function such that

\[
(\phi \psi_k)(a) = \frac{1}{2} (\psi_k|_{t_1}(a) + \psi_k|_{t_2}(a)) \quad (3.126)
\]

for all \( t_1, t_2 \in \mathcal{T}_k \) sharing a common edge. If \( t \in \mathcal{T}_k \) has an edge lying on \( \partial \Omega \), then \( \psi_k|_t \) vanishes in the middle point of this edge. Conditions (3.124) are easy to verify.

It is easy to compute the prolongation \( \rho : \mathbb{R}_{k-1} \to \mathbb{R}_k \) induced by this mapping \( \phi \). In the notation of Figure 3.10, one gets

\[
\begin{align*}
u_k(b) & = \frac{1}{2} (\nu_{k-1}(E) + \nu_{k-1}(C)), \\
u_k(a) & = \frac{1}{2(|t_1| + |t_2|)} \left(|t_1| (2\nu_{k-1}(E) + \nu_{k-1}(C) - \nu_{k-1}(B))ight. \\
 & \left. + |t_2| (2\nu_{k-1}(E) + \nu_{k-1}(D) - \nu_{k-1}(A))\right),
\end{align*}
\]

where formally \( \nu_{k-1} = P_{k-1} z \), \( \nu_k = P_k \rho z \) for arbitrary \( z \in \mathbb{R}_{k-1} \); \( |t_i| = \text{mes}(t_i) \).

If one uses integral, rather than local, edge conditions to define \( \mathcal{U}_k \), then one defines the mapping \( \phi \) in a similar way. The only difference is that in (3.126), values in the edge midpoints are replaced by the integral mean values.

The construction and analysis of a multigrid method for the Wilson element differs only slightly from the case of lower order nonconforming elements. To define \( \mathcal{V}_k \), one may consider the space of all functions, which are second order polynomial on every rectangle from \( \mathcal{T} \). Then for any fixed \( \psi_k \in \mathcal{V}_k \), one defines \( \phi \psi_k \) as follows: On every rectangle from \( \mathcal{T} \) the function \( \phi \psi_k \) is the second order polynomial such that the identities

\[
(\phi \psi_k)(a) = \frac{1}{n} \sum_{i=1}^{n} \psi_k|_{t_i}(a)
\]
hold in every vertex $a$, which is shared by $n$ rectangles $t_i$, and

$$
\int_t (\varphi v_k)_{xx} \, dx = \int_t (v_k)_{xx} \, dx, \quad \int_t (\varphi v_k)_{yy} \, dx = \int_t (v_k)_{yy} \, dx \quad \forall \ t \in \mathcal{T}.
$$

The function $\varphi v_k$ is uniquely defined subject to boundary conditions. It is not difficult to check the property (3.124).

### Exercises

3.4.1. Prove that the matrix from the approximation property is nonnegative definite:

$$
A^{-1}_k - pA^{-1}_{k-1} r \geq 0.
$$

3.4.2. Show that the simple iteration possesses a smoothing property, if assumptions for $A$ and $w$ from Exercise 1.1.5 hold true.

3.4.3. (Exercise from [97]) Assume that the matrix $A_k$ is normal, i.e., $A_k A_k^T = A_k^T A_k$. As a smoothing method consider the simple iteration with a relaxation parameter $w$. Prove the estimate

$$
||A_k S_k|| \leq \frac{1}{w} \max_{x \in \mathcal{Z}} (1 - |z|^r),
$$

where $\mathcal{Z}$ is a subdomain of $C$ such that $\text{sp}(w A_k) \subseteq \mathcal{Z}$.

3.4.4. Assume that the hierarchy $U_0 \subset U_1 \subset \cdots \subset U_k \subset \cdots U_\ell$ of finite element spaces is built based on a regular global refinement procedure applied to the coarsest grid in a bounded planar domain. Denote by $N_V$ and $N_W$ the complexity of $V$- and $W$-cycles with one pre- and one postsmoothing step for solving the Poisson problem. Compute the limit $\lim_{l \to \infty} N_V / N_W$. Let $q_V$ and $q_W$ be the corresponding convergence factors. How small should $q_W$ be compared to $q_V$ to justify using the $W$-cycle with enough grid levels?

3.4.5. Consider the domain

$$
\Omega_\varepsilon = \{ x = r e^{i \phi} : \ r \in (0, 1), \ \phi \in (0, \pi / \omega) \},
$$

$\omega \in (\frac{1}{2}, 1)$, i.e., the unit disk without a segment. Show that the function $u(r, \phi) = (r^{\omega} - r^2) \sin(\omega \phi)$, defined in the polar coordinates, solves the Poisson problem (3.83) with the right-hand side $f = (4 - \omega^2) \sin(\omega \phi)$. Check that $f$ is in $L^2(\Omega)$, but $u$ does not belong to $H^2(\Omega)$.

3.4.6. Prove (3.118) for arbitrary matrix $B$, $0 \leq B = B^T \leq I$.

3.4.7. Assume we should solve an elliptic problem with Neumann or homogeneous Dirichlet boundary conditions with the $P_1$ nonconforming finite element. Show that

$$
|v|_{t_1}(E) = |v|_{t_2}(E) \quad \text{and} \quad \int_{t_1 \cap t_2} v \, dx = \int_{t_1 \cap t_2} v \, dx
$$

defines the same finite element space $U_h$.

3.4.8. Check that the space of conforming $P_1$ finite elements is a subspace of the Crouzeix–Raviart finite elements.
3.4.9. Check that for the Crouzeix–Raviart element the constructed mapping \( \varphi \) satisfies conditions (3.124). Build the appropriate \( \mathbb{V}_k \) and \( \varphi \) for \( Q_2 \) nonconforming elements.

3.4.10. Find the stiffness matrix for the Crouzeix–Raviart element discretization of the Poisson problem, using the uniform triangulation of the domain in Figure 3.2. What can be the value of the relaxation parameter in the Jacobi method for the smoothing property to hold?
Chapter 4
Preconditioners by Space Decomposition

An efficient approach to building preconditioners is based on a decomposition of vector space $\mathbb{R}^N$ on subspaces of smaller dimensions and approximate solution of $Ax = b$ on the subspaces. The choice of subspaces is often driven by an application where the system originates and by properties of matrix $A$. Depending on decomposition principles, a field of application, and algorithmic details, such preconditioners are known in the literature as domain decomposition methods, additive and multiplicative Schwartz methods, subspace correction, and block preconditioners. Many other known names may refer to particular methods from these largely overlapping classes of algorithms. As we shall see, simple multigrid methods on a hierarchy of embedded spaces also fall into the category of space decomposition preconditioners. It is commonly believed that methods based on space decomposition are especially well suited for parallel computer architecture, and this often makes them methods of choice for large-scale scientific computing.

4.1 Space decomposition framework

Space decomposition is a rather general framework for building preconditioners to symmetric positive definite matrices. It also provides a new look at some classical linear iterative methods. Presumably first analysis of preconditioners based on space decompositions was done by Matsokin and Nepomnyaschikh [145] in the context of what is currently known as the additive Schwarz method, and many ideas can be also traced back to the series of papers of Lions [133, 134, 135]: We refer the reader to [193] for a more expanded historical overview. An extended and unified presentation of the approach was given in [221] and [91]. The framework and abstract eigenvalue estimates in this section are largely due to the review article [221]. We shall see that these abstract estimates are very useful when applied to particular practical methods.

4.1.1 Space decomposition and preconditioning

Consider a system of linear algebraic equations

$$Az = b,$$

where $b$ and $z$ belong to a space $\mathbb{V} = \mathbb{R}^N$ and $A$ is a symmetric positive definite matrix. Assume there is a decomposition of $\mathbb{V}$ into the sum of subspaces (not necessarily
a direct sum):

\[ \mathbb{V} = \sum_{i=0}^{l} \mathbb{V}_i, \quad \mathbb{V}_i \subset \mathbb{V}. \]  

(4.2)

In other words, for any \( v \in \mathbb{V} \) there exists (not necessarily the unique) decomposition

\[ v = \sum_{i=0}^{l} v_i, \quad v_i \in \mathbb{V}_i. \]

Our intermediate goal is to build an algorithm for solving (4.1) based on the solutions of several smaller algebraic systems on the subspaces. The plan is first to project the system (4.1) onto the subspaces, and then to solve systems of smaller dimensions. When done, information from all subspaces \( \mathbb{V}_i \) is collected to produce a new approximation to the solution on \( \mathbb{V} \).

Let us define orthogonal projector \( Q_i : \mathbb{V} \rightarrow \mathbb{V}_i \), A-orthogonal projector \( Q_i^A : \mathbb{V} \rightarrow \mathbb{V}_i \), and the subspace operator \( A_i : \mathbb{V}_i \rightarrow \mathbb{V}_i \) through the relations

\[ \langle Q_i v, v_i \rangle = \langle v, v_i \rangle, \quad \langle Q_i^A v, v_i \rangle_A = \langle v, v_i \rangle_A \quad \forall v \in \mathbb{V}, \quad v_i \in \mathbb{V}_i, \]

\[ \langle A_i z, v_i \rangle = \langle A z, v_i \rangle \quad \forall z, v_i \in \mathbb{V}_i. \]

The operator \( A_i \) can be observed as a projection of \( A \) into the subspace \( \mathbb{V}_i \). It directly follows from the definition of \( A_i, Q_i^A, \) \( Q_i \) that for any \( v \in \mathbb{V} \) and \( v_i \in \mathbb{V}_i \) it holds that

\[ \langle A_i Q_i^A v, v_i \rangle = \langle A Q_i^A v, v_i \rangle = \langle Q_i^A v, v_i \rangle_A = \langle v, v_i \rangle_A = \langle A v, v_i \rangle = \langle Q_i A v, v_i \rangle. \]

Since \( v \) and \( v_i \) are arbitrary, we get

\[ A_i Q_i^A = Q_i A. \]  

(4.3)

The direct consequence of (4.3) and (4.1) is the equality

\[ A_i z_i = b_i, \]  

(4.4)

where \( z_i = Q_i^A z, \quad b_i = Q_i b \). Equation (4.4) is the subspace equation.

Assume that one solves approximately the system (4.1) using an iterative method. Let \( z^0 \) be a given approximation to \( z \), let \( r^0 = A z^0 - b \) be residual, and let \( e = z^0 - z \) be unknown error. The error vector satisfies the equation

\[ A e = r^0. \]

If one finds \( e \), the exact solution of the system is recovered from \( z = z_0 - e \). We will look for an approximation to \( e \). For this purpose, we consider subspace equations for the error in \( \mathbb{V}_i \):

\[ A_i e_i = Q_i r^0, \quad \text{where} \quad e_i = P_i e, \quad i = 0, \ldots, l. \]

Generally, it may happen that an efficient exact method for solving these subspace equations is not available. Assume that we can find an approximation \( \hat{e}_i \) to \( e_i \) from

\[ \hat{e}_i = R_i Q_i r^0, \]

where \( R_i \) is an approximate inverse of \( A_i \). In the familiar terms, we may say that \( R_i \) is a preconditioner to \( A_i \). Collecting subspace error approximations, we find the new approximation to the solution in the form

\[ z^1 = z^0 - \sum_{i=0}^{l} \hat{e}_i. \]
4.1. Space decomposition framework

The latter can be written as one preconditioned iteration

\[ z^1 = z^0 - B(Az^0 - b) \]

with the preconditioner

\[ B = \sum_{i=0}^l R_i Q_i. \]

**Lemma 4.1.** If all \( R_i \) are positive definite and self-adjoint, then the operator \( B \) from (4.6) is positive definite and self-adjoint.

**Proof.** To show that \( B \) is self-adjoint, we observe the following equalities for arbitrary \( y, v \) from \( \mathbb{V} \):

\[ \langle By, v \rangle = \left( \sum_{i=0}^l R_i Q_i y, v \right) = \sum_{i=0}^l \langle R_i Q_i y, v \rangle. \]

The assumptions on \( R_i \) yield \( \langle R_i Q_i y, v \rangle = \langle Q_i y, R_i Q_i v \rangle \), and \( B \) is self-adjoint by the following definition: \( \langle By, v \rangle = \langle y, Bv \rangle \). Setting \( y = v \) in (4.7), we get that \( B \) is nonnegative: \( \langle Bv, v \rangle \geq 0 \).

To prove positive definiteness, it remains to check that \( \langle Bv, v \rangle = 0 \) implies \( v = 0 \). Indeed, let \( \langle Bv, v \rangle = 0 \); then from (4.7) for \( y = v \) and the positive definiteness of \( R_i \) it follows that \( Q_i v = 0 \) for any \( i \). Consider arbitrary decomposition \( v = \sum_{i=0}^l v_i, \ v_i \in \mathbb{V}_i \). We get

\[ \langle v, v \rangle = \sum_{i=0}^l \langle v, v_i \rangle = \sum_{i=0}^l \langle Q_i v, v_i \rangle = 0. \]

Therefore, if \( A = A^T > 0 \) and \( R_i = R_i^T > 0 \), then it looks reasonable to solve (4.1) numerically by applying the conjugate gradient (CG) method with the preconditioner \( B = \sum_{i=0}^l R_i Q_i \).

**Example 4.2.** Consider \( \mathbb{V} = \mathbb{R}^N \) and the decomposition of \( \mathbb{R}^N \) into the sum of orthogonal one-dimensional subspaces \( \mathbb{R}^N = \sum_{i=1}^N \text{span} \{ \hat{e}_i \} \); here \( \hat{e}_i \) are the unit basis vectors. Then

\[ Q_i v = v_i \hat{e}_i, \]

where \( v_i \) is the \( i \)th component of the vector \( v \). For a matrix \( A \) we get \( \langle A \hat{e}_i, \hat{e}_j \rangle = a_{i,j} \delta_{i,j} \), and hence

\[ A_{ij} = a_{i,j}. \]

If we set \( R_i = A_i^{-1} = a_{i,i}^{-1} \), then the linear preconditioned iteration (4.5) is equivalent to the Jacobi method.

### 4.1.2 Sequential version

In the iteration (4.5)–(4.6), computations on every subspace are independent of each other and can be done in parallel. Conversely, one may consider a strategy where the subspaces are visited sequentially and the approximation to the solution is updated every time.
Starting with $V$, one may proceed as follows: Let $z^{old} \in V$ be given; then the next iterate $z^{new}$ is found in $l + 1$ steps of the algorithm: Set $z^0 = z^{old}$ and compute
\[ z^{i+1} = z^i - R_i Q_i (A z^i - b), \quad i = 0, \ldots, l. \] (4.8)

Now set $z^{new} = z^{l+1}$.

Denote $T_i = R_i Q_i A$. Thanks to (4.3) it holds that $T_i = R_i A_i Q_i^2$. Iterations (4.8) are performed sequentially, and thus the iteration matrix of the sequential subspace decomposition method is given by
\[ M_l = (I - T_l)(I - T_{l-1}) \ldots (I - T_0). \] (4.9)

For the decomposition of $V = \mathbb{R}^n$ as in Example 4.2, the sequential subspace iterative method is equivalent to the Gauss-Seidel iteration (check it).

Of course, the sequential subspace iterative method, as any linear iteration, defines an implicit preconditioner for matrix $A$; see subsection 1.4.13.

### 4.1.3 The strengthened Cauchy–Schwarz inequality

The strengthened Cauchy–Schwarz inequality holds for subspaces $H_1 \subset H$ and $H_2 \subset H$, of a Hilbert or Euclidean space $H$, if
\[ |(u_1, u_2)_H| \leq \gamma ||u_1||_H ||u_2||_H \quad \forall u_1 \in H_1, u_2 \in H_2, \] (4.10)
with some $\gamma \in [0, 1)$. Note that the Cauchy inequality ensures (4.10) with $\gamma = 1$. However, it may happen that the optimal $\gamma$ is less than one. For example, for orthogonal subspaces $\gamma = 0$, the value of arccos $\gamma_0$ is sometimes called the angle between subspaces $H_1$ and $H_2$, where $\gamma_0 := \sup_{u_1 \in H_1, u_2 \in H_2} |(u_1, u_2)_H| / (||u_1||_H ||u_2||_H)$.

The strengthened Cauchy–Schwarz inequality for subspaces $V_i$ and $V_j$ plays an important role in the analysis of preconditioners based on space decompositions. Assume that for any $y_i \in V_i$ and $v_j \in V_j$ it holds that
\[ \langle y_i, v_j \rangle_A \leq \gamma_{ij} ||y_i||_A ||v_j||_A, \] (4.11)
where $\gamma_{ij}$ does not depend on the choice of particular elements from $V_i$ and $V_j$. All constants $\gamma_{ij}$ form a symmetric matrix $G = \{ \gamma_{ij} \} \in \mathbb{R}^{(l+1) \times (l+1)}$. Denote by $\rho(G)$ the spectral radius of $G$; then $||G||_2 = \rho(G)$.

For any two vectors $a, b \in \mathbb{R}^{l+1}$ we have
\[ \sum_{i,j=0}^l \gamma_{ij} a_i b_j = \langle (G - \operatorname{diag}(G)) a, b \rangle \leq ||G||_2 ||a||_2 ||b||_2 = \rho(G) ||a||_2 ||b||_2, \] (4.12)
\[ \sum_{i=0}^l \sum_{j=0}^{l-1} \gamma_{ij} a_i b_j = \frac{1}{2} \langle (G - \operatorname{diag}(G)) a, b \rangle \leq \frac{1}{2} ||G||_2 ||a||_2 ||b||_2 = \frac{1}{2} \rho(G) ||a||_2 ||b||_2. \] (4.13)

We remark that to show the inequality in (4.13) one can use the fact that entries of $G$ are nonnegative, and hence there exists the eigenvector $v$ with nonnegative entries such that $G v = \rho(G) v$; see, e.g., [146]. From this, the estimate $||G - \operatorname{diag}(G)||_2 \leq ||G||_2$ follows (prove this).

Now consider an arbitrary set of vectors $y_i \in V_i, \; v_j \in V_j, \; i, j = 0, \ldots, l$. Summing up (4.11) and applying (4.12), (4.13) with $a_i = ||y_i||_A$ and $b_j = ||v_j||_A$, we get the following
two estimates useful for further analysis:
\[
\sum_{i,j=0}^{l} \langle y_i, v_j \rangle_A \leq \rho(G) \left( \sum_{i=0}^{l} \| y_i \|_A^2 \right)^{1/2} \left( \sum_{j=0}^{l} \| v_j \|_A^2 \right)^{1/2},
\]
(4.14)
\[
\sum_{i=0}^{l} \sum_{j=0}^{l-1} \langle y_i, v_j \rangle_A \leq \frac{1}{2} \rho(G) \left( \sum_{i=0}^{l} \| y_i \|_A^2 \right)^{1/2} \left( \sum_{j=0}^{l} \| v_j \|_A^2 \right)^{1/2}.
\]
(4.15)

Since it always holds that \( \gamma_{ij} \leq 1 \), the trivial bound for \( \rho(G) \) is
\[
\rho(G) \leq l + 1.
\]
(4.16)

### 4.1.4 Eigenvalue analysis of space decomposition preconditioners

An estimate for the convergence factor of the preconditioned CG method, with a preconditioner \( B \), can be written in terms of the condition number of the matrix \( AB \); see subsection 1.4.13. For the space decomposition preconditioner (4.6), an estimate on \( \text{cond}(AB) \) is given by Theorem 4.3 below. We shall apply this theorem to prove a mesh-independent convergence of the CG method, with an additive multigrid, hierarchical basis, and additive Schwarz preconditioners. In Theorem 4.4, we also show an estimate on a norm of an iteration matrix for the sequential version of subspace iteration (4.8). Thanks to Theorem 1.33, it also implies an estimate on \( \text{cond}(AB) \) for the corresponding space decomposition preconditioner. We shall apply this result to multiplicative Schwarz preconditioners.

**Theorem 4.3.** Assume that \( A \) and \( R_i \) are symmetric positive definite matrices. Assume also that for any \( v \in \mathcal{V} \) there exists decomposition \( v = \sum_{i=0}^{l} v_i \), \( v_i \in \mathcal{V}_i \), such that
\[
\sum_{i=0}^{l} \langle R_i^{-1} v_i, v_i \rangle \leq K_0 \langle Av, v \rangle,
\]
(4.17)
with a constant \( K_0 \) independent of \( v \).

Let \( G \) be a matrix defined by (4.11) and
\[
\alpha = \max_{i=0,\ldots,l} \lambda_{\max}(R_i A_i).
\]
(4.18)

Then for the parallel space decomposition preconditioner defined in (4.6), it holds that
\[
\text{sp}(BA) \in [K_0^{-1} \alpha \rho(G)].
\]
(4.19)

**Proof.** Let \( T_i = R_i Q_i A = R_i A_i Q_i^T \) and denote \( T = BA = \sum_i T_i \). From Lemma 4.1 we know that the operator \( T \) is self-adjoint and positive definite in the scalar product \( \langle \cdot, \cdot \rangle_A \), and so all eigenvalues of \( T \) are real and positive.

First, we check the estimate \( \lambda_{\max}(BA) \leq \alpha \| G \| \). From (4.14) and (4.18), we get, for an arbitrary set of vectors \( y_i \in \mathcal{V}_i \), \( v_j \in \mathcal{V}_j \), \( i, j = 0, \ldots, l \),
\[
\sum_{i,j=0}^{l} \langle y_i, v_j \rangle_A \leq \rho(G) \left( \sum_{i=0}^{l} \langle A_i y_i, y_i \rangle \right)^{1/2} \left( \sum_{j=0}^{l} \langle A_j v_j, v_j \rangle \right)^{1/2}
\]
\[
\leq \alpha \rho(G) \left( \sum_{i=0}^{l} \langle R_i^{-1} y_i, y_i \rangle \right)^{1/2} \left( \sum_{j=0}^{l} \langle R_j^{-1} v_j, v_j \rangle \right)^{1/2}.
\]
Chapter 4. Preconditioners by Space Decomposition

For any two vectors $y, v \in \mathbb{V}$, setting $y_i = T_i y$ and $v_i = T_i v$ yields

$$
\sum_{i,j=0}^l \langle T_i y, T_j v \rangle_A \leq \alpha \rho(G) \left( \sum_{i=0}^l \langle R_i^{-1} T_i y, T_i y \rangle \right)^{1/2} \left( \sum_{i=0}^l \langle R_i^{-1} T_i v, T_i v \rangle \right)^{1/2}.
$$

(4.20)

Using the definitions of the matrices $T_i$ and $Q_i$, one checks that

$$
\langle T_i y, y \rangle_A = \langle AT_i y, y \rangle = \langle AR_i Q_i A y, y \rangle = \langle R_i Q_i A y, y \rangle
$$

$$
= \langle R_i Q_i A y, Q_i A y \rangle = \langle R_i^{-1} R_i Q_i A y, R_i Q_i A y \rangle = \langle R_i^{-1} T_i y, T_i y \rangle.
$$

Therefore, the estimate (4.20) yields

$$
\sum_{i,j=0}^l \langle T_i y, T_j v \rangle_A \leq \alpha \rho(G) \langle T y, y \rangle_A^{1/2} \langle T v, v \rangle_A^{1/2}.
$$

(4.21)

The choice of $y = v$ in the above inequality implies

$$
\|T v\|^2_A = \sum_{i,j=0}^l \langle T_i v, T_j v \rangle_A \leq \alpha \rho(G) \langle T v, v \rangle_A \leq \alpha \rho(G) \|T v\|_A \|v\|_A.
$$

We conclude that

$$
|\lambda(BA)| = |\lambda(T)| \leq \|T\|_A = \sup_{0 \neq v \in \mathbb{V}} \frac{\|T v\|_A}{\|v\|_A} \leq \alpha \rho(G).
$$

We used the fact that the spectral radius of a matrix does not exceed any operator matrix norm.

Now we check the estimate $\lambda_{\text{min}}(BA) \geq K^{-1}_\alpha$. The theorem implies that for any $v \in \mathbb{V}$ there exists decomposition $v = \sum_i v_i$, $v_i \in \mathbb{V}_i$, such that the estimate (4.17) is valid. We make use of this fact in the following relations:

$$
\langle v, v \rangle_A = \sum_{i=0}^l \langle v_i, v_i \rangle_A = \sum_{i=0}^l \langle v_i, Av \rangle = \sum_{i=0}^l \langle v_i, Q_i Av \rangle.
$$

Further, we estimate

$$
\sum_{i=0}^l \langle v_i, Q_i Av \rangle = \sum_{i=0}^l \langle R_i^{-1/2} v_i, R_i^{1/2} Q_i Av \rangle
$$

$$
\leq \sum_{i=0}^l \langle R_i^{-1/2} v_i, v_i \rangle^{1/2} \langle R_i Q_i Av, Q_i Av \rangle^{1/2}
$$

$$
= \sum_{i=0}^l \langle R_i^{-1} v_i, v_i \rangle^{1/2} \langle R_i Q_i Av, v \rangle^{1/2} = \sum_{i=0}^l \langle R_i^{-1} v_i, v_i \rangle^{1/2} \langle T_i v, v \rangle_A^{1/2}
$$

$$
\leq \left( \sum_{i=0}^l \langle R_i^{-1} v_i, v_i \rangle \right)^{1/2} \left( \sum_{i=0}^l \langle T_i v, v \rangle_A \right)^{1/2} \leq \sqrt{K_\alpha} \|v\|_A \langle T v, v \rangle_A^{1/2}.
$$

We get $\|v\|^2_A \leq K_\alpha \langle T v, v \rangle_A$, which implies the lower bound for $\lambda_{\text{min}}(T)$, if we set $v$ equal to the corresponding eigenvector.

\[ \square \]

**Theorem 4.4.** Given the same assumptions as in Theorem 4.3 and $\alpha < 2$, the iteration matrix of the sequential subspace iterative method (4.8) satisfies the estimate
4.1. Space decomposition framework

\[ ||M_l||_A^2 \leq 1 - \frac{2 - \alpha}{K_0(1 + \frac{1}{2} \alpha \rho(G))}. \] (4.22)

**Proof.** We repeat the arguments from the proof of Theorem 4.3, but now we use (4.15) instead of (4.14) and set \( y_i = T_i w_i \) for some \( w_i \in V \) defined later. We get (compare to (4.21))

\[ \sum_{i=0}^{l-1} \sum_{j=0}^{l-1} \langle T_i w_i, T_j v \rangle_A \leq \frac{1}{2} \alpha \rho(G) \left( \sum_{i=0}^{l} \langle T_i w_i, w_i \rangle_A \right)^{1/2} \langle T v, v \rangle_A^{1/2} \] (4.23)

for an arbitrary \( v \in V \). Also, the Cauchy inequality gives

\[ \sum_{i=0}^{l} \langle T_i w_i, v \rangle_A \leq \left( \sum_{i=0}^{l} \langle T_i w_i, w_i \rangle_A \right)^{1/2} \langle T v, v \rangle_A^{1/2}. \] (4.24)

Denote \( M_i = (I - T_i) \ldots (I - T_0), i = 0, \ldots, I, \) and set \( M_{l-1} = I \). From the identity \( M_{l-1} - M_i = T_i M_{l-1-i} \) it follows that \( I - M_i = \sum_{j=0}^{l} T_j M_{l-1-i} \). We use the last identity and \( T = \sum_{i=0}^{l} T_i \) to obtain

\[ \langle T v, v \rangle_A = \sum_{i=0}^{l} \langle T_i v, M_{l-1-i} v \rangle_A + \sum_{i=0}^{l} \langle T_i v, (I - M_{l-1-i}) v \rangle_A \]

\[ = \sum_{i=0}^{l} \langle T_i v, M_{l-1-i} v \rangle_A + \sum_{i=0}^{l} \sum_{j=0}^{l-1} \langle T_i v, T_j M_{l-1-i} v \rangle_A. \]

Now we use (4.23) and (4.24), with \( w_i = M_{l-1-i} v \), to estimate two terms on the right-hand side. This leads to

\[ \langle T v, v \rangle_A \leq \left( 1 + \frac{1}{2} \alpha \rho(G) \right)^2 \sum_{i=0}^{l} \langle T_i M_{l-1-i} v, M_{l-1-i} v \rangle_A. \] (4.25)

Note that for the left-hand side term, the spectral bound (4.19) implies

\[ K_0^{-1} ||v||_A^2 \leq \langle T v, v \rangle_A. \] (4.26)

To estimate the right-hand side of (4.25), we note that \( M_{l-1-i} - M_i = T_i M_{l-1-i} \) yields

\[ ||M_{l-1-i} v||_A^2 - ||M_i v||_A^2 = \langle T_i M_{l-1-i} v, v \rangle_A + 2 \langle T_i M_{l-1-i} v, M_i v \rangle_A \]

\[ = \langle T_i M_{l-1-i} v, M_i v \rangle_A + 2 \langle T_i (I - T_i) M_{l-1-i} v, M_i v \rangle_A \]

\[ = \langle (2I - T_i) T_i M_{l-1-i} v, M_i v \rangle_A \geq (2 - \alpha) \langle T_i M_{l-1-i} v, M_i v \rangle_A. \]

Summing up the inequalities for \( i = 0, \ldots, l \) gives

\[ \sum_{i=0}^{l} \langle T_i M_{l-1-i} v, M_i v \rangle_A \leq (2 - \alpha)^{-1} (||v||_A^2 - ||M_i v||_A^2). \]

Using this to estimate the right-hand side of (4.25) and applying (4.26) to the left-hand side of (4.25) leads to

\[ K_0^{-1} ||v||_A^2 \leq \frac{(1 + \frac{1}{2} \alpha \rho(G))^2}{(2 - \alpha)} (||v||_A^2 - ||M_i v||_A^2). \]

Performing obvious rearrangements of terms gives \( ||M_i v||_A^2 \leq 1 - \frac{2 - \alpha}{K_0(1 + \frac{1}{2} \alpha \rho(G))} ||v||_A^2. \) This proves the theorem. \( \square \)
4.2 Grid decomposition methods

For linear algebraic systems resulting from partial differential equations discretizations, a large group of vector space decompositions for the purpose of preconditioning is based on hierarchical grid decompositions. This links the approach with classical multigrid techniques described in the previous chapter. To make this connection even more clear, let us introduce another variant of multigrid iteration known as the additive multigrid method.

4.2.1 Additive multigrid method

The multigrid preconditioners considered earlier in the text are sometimes called multiplicative multigrid methods. Consequently, computations in a multiplicative method are performed on different grid levels and cannot be run in parallel. For computers with parallel architecture, it is often more efficient to use an additive variant of a multigrid method. Moreover, additive multigrid methods turn out to be more convenient (also from the implementation point of view) for problems discretized on locally refined grids. Building preconditioners for locally refined grids is discussed at the end of this chapter.

Thus, assume we are looking for a solution to $Az = f$, where $A$ is a finite element stiffness matrix with discrete space $U_h$. Repeating assumptions from Section 3.3, let a hierarchy of finite element spaces be given as

$$U_0 \subset U_1 \subset \cdots \subset U_k \subset \cdots \subset U_l = U_h.$$  \hspace{1cm} (4.27)

Subspace $U_0$ corresponds to the coarsest grid. Canonical isomorphisms between $U_k$ and Euclidean vector spaces of coefficients $\mathbb{R}^k$ lead to the chain of vector spaces,

$$\mathbb{R}_0 \rightarrow \mathbb{R}_1 \rightarrow \cdots \rightarrow \mathbb{R}_k \rightarrow \cdots \rightarrow \mathbb{R}_l = \mathbb{R}_h,$$  \hspace{1cm} (4.28)

with prolongation and restriction operators $p_k$ and $r_k$. Assume that each of the grid level matrices $A_l = A$ and $A_k$, $k = 0, \ldots, l - 1$, is given or computed as was explained in subsection 3.2.3.

We define one iteration of an additive method for solving the system $Az = f$ as the one execution of the procedure $z^{\text{new}} = \text{AMGM}(l, z^{\text{old}}, f)$, where $z^{\text{old}}$ is a given approximation to $z$: $\tilde{z} = \text{AMGM}(k, z, f)$

If $k = 0$, then

1. $d = r_k(A_k z - f)$ (restriction of the residual to a coarser grid);
2. $z^0 = z, z^{i+1} = z^i - W^{-1}_k(A_k z^i - f), i = 0, \ldots, \nu - 1$ (v basic smoothing iterations);
3. $e = 0$ (the initial guess to the error on the coarser grid);
4. call $\tilde{e} = \text{AMGM}(k - 1; e, d)$ (recursive call of the multigrid on the coarser grid);
5. $z = z^\nu - \theta p_k \tilde{e}$ (coarse grid correction of $z^\nu$, $\theta \in \mathbb{R}$ is a parameter),

Note that the recursive call to the procedure in step 4 is independent of the result of smoothing iterations in step 2 and can be performed in parallel. This is in contrast to
4.2. Grid decomposition methods

V- and W-cycles studied previously. Therefore, the smoothing iterations on all grid levels can be performed in parallel. Schematically the additive method is illustrated in Figure 4.1.

The iteration matrix of the additive two-grid method is

\[ M_A = I - (\theta p_h A_h^{-1} r_k + (I - S) A_h^{-1}) A_h, \] (4.29)

where \( S \) is the iteration matrix of the smoothing iterations, and the coarse problem in step 4 is solved exactly. We leave checking (4.29) to the reader for an exercise.

From (4.29) we see that the additive multigrid method does not turn out to be the exact solver if \(|S| < 1\) (smoothing iterations are convergent) and \( \nu \to \infty \). This is in contrast to the multiplicative one. The additive method would turn to an exact solver only if \( \theta \) depends on \( \nu \), and \( \theta(\nu) \to 0 \) for \( \nu \to \infty \). To ensure convergence of the additive method, one has to make a correct choice of the parameter \( \theta \). The optimal value of \( \theta \) is not known a priori. For this reason, the additive method is almost always used as a preconditioner in a Krylov subspace method with \( \theta = 1 \) and rarely as an independent iterator. We study convergence of such iterations further in this chapter.

To see how the additive multigrid method fits into the framework of space decomposition, let us write out the preconditioner \( B_A \) defined by one iteration of AMGM,

\[ B_A = \theta^i p_0 A_0^{-1} r_0 + \sum_{k=1}^l \theta^{i-k} p_k (I - S_k) A_k^{-1} r_k, \]

where \( p_l = I \) and \( p_k = p_{l-1} \ldots p_{k+1} \), \( r_k = p_k \); cf. the diagram (4.28). For the particular case of \( \theta = 1 \) and \( \nu = 1 \), we see that the additive multigrid method defines the preconditioner of a form similar to (4.6):

\[ B_A = p_0 A_0^{-1} r_0 + \sum_{k=1}^l p_k W_k^{-1} r_k. \] (4.30)

A natural question is, Does (4.30) correspond to a space decomposition? If yes, then it can be studied within the developed framework.

### 4.2.2 BPX preconditioner

We take another viewpoint on the additive multigrid method above. Consider a finite element problem with the symmetric elliptic bilinear form \( a(\cdot, \cdot) \) on a space \( \Omega_h \). For example, \( a(\psi, \phi) = (\nabla \psi, \nabla \phi) \) for the Poisson problem. This gives rise to the system of linear algebraic equations

\[ Az = f \quad \text{and} \quad (A \cdot, \cdot) = a(P \cdot, P \cdot). \] (4.31)
Assume we are given the hierarchy of embedded finite element spaces \( \mathbb{U}_0 \subset \cdots \subset \mathbb{U}_I =: \mathbb{U}_h \). Obviously, one has the decomposition
\[
\mathbb{U}_h = \mathbb{U}_0 + \cdots + \mathbb{U}_I.
\] (4.32)

According to the formalism of subsection 4.1.1 this decomposition can be used to define a preconditioner to the operator \( A_h : \mathbb{U}_h \to \mathbb{U}_h \) induced by the bilinear form:
\[
\begin{align*}
(A_h u_h, v_h) = a(u_h, v_h) & \quad \forall u_h, v_h \in \mathbb{U}_h.
\end{align*}
\]

On subspaces \( \mathbb{U}_k, k = 1, \ldots, I \), let us define preconditioners \( R_k : \mathbb{U}_k \to \mathbb{U}_k \), of the Jacobi type
\[
R_h^{(k)} v_h = \left( \sum_{i=1}^{n_k} (v_h, \phi_i^{(k)}) \phi_i^{(k)} \right) \quad \text{for } v_h \in \mathbb{U}_k,
\] (4.33)

where \( \{\phi_i^{(k)}\}_{i=1,...,N_k} \) is a nodal basis in \( \mathbb{U}_k \). On the very coarsest space, our preconditioner is the exact solver:
\[
([R_h^{(0)}]^{-1} u_h, v_h) = a(u_h, v_h) \quad \forall u_h, v_h \in \mathbb{U}_0.
\]

Note that in the denominator of (4.33) we have the diagonal elements of the coarse grid matrices \( a(\phi_i^{(k)}, \phi_j^{(k)}) = (A_k)_{ij} \). The subspace decomposition preconditioner is then
\[
B_h = R_h^{(0)} Q_h^{(0)} + \sum_{k=1}^I R_h^{(k)} Q_h^{(k)} = R_h^{(0)} Q_h^{(0)} + \sum_{k=1}^I R_h^{(k)},
\] (4.34)

where \( Q_h^{(k)} \) is the orthogonal projector on \( \mathbb{U}_k \). The preconditioner is convenient for analysis, since both the machinery of finite element analysis and the abstract results of Section 4.1 can be combined. Now let us study how the operator in (4.34) relates to the additive multigrid preconditioner (4.30). Toward this end, for an operator \( C_h : \mathbb{U}_h \to \mathbb{U}_h \) consider its matrix counterpart \( C_m \in \mathbb{R}^{N \times N} \) defined through the following identity:
\[
(C_m u, v) = (C_h u_h, v_h) \quad \forall u, v \in \mathbb{R}^N \text{ and } u_h = P_I u, \ v_h = P_I v \in \mathbb{U}_h.
\]

Obviously, the matrix counterpart of \( A_h \) is the matrix \( A \) from (4.31). Following the arguments from [222], we show the following equality:
\[
(A_h B_h A_h)_{mn} = A B_m A.
\] (4.35)

To check (4.35), denote by \( M \) a mass matrix for \( \mathbb{U}_h \),
\[
(M)_{ij} = (\phi_i^{(l)}, \phi_j^{(l)}).
\]

Since \( M = P_I^* P_I \), one easily shows the identity
\[
(C_h D_h)_{mn} = C_m M^{-1} D_m
\] (4.36)

for arbitrary operators \( C \) and \( D \) on \( \mathbb{U}_h \). Further, noting the following matrix-vector identity:
\[
[\phi_1^{(k)}, \ldots, \phi_{n_k}^{(k)}]^T = r_k [\phi_1^{(l)}, \ldots, \phi_{n_k}^{(l)}]^T,
\]

one finds for \( k = 1, \ldots, I \)
\[
(R_h^{(k)})_{mn} = M p_k W_k^{-1} r_k, \quad \text{with } W_k = \text{diag}(A_k).
\] (4.37)
4.2. Grid decomposition methods

The equality (4.35) follows from (4.36), (4.37), and the definitions of $B_h$ and $B_h'$. Thanks to (4.35), we have the following spectral equivalence of operators and matrices:

$$c_{\min}(A_h v_h, v_h) \leq \langle A_h B_h A_h v_h, v_h \rangle \leq c_{\max}(A_h v_h, v_h) \quad \forall \, v_h \in U_h$$

(4.38)

with suitable $c_{\min}$ and $c_{\max}$. Since $c_{\min}$ and $c_{\max}$ can be taken the same in (4.38) and (4.39), we may conclude the following for the spectral condition numbers:

$$\text{cond}(B_h A_h) = \text{cond}(B_h A).$$

(4.40)

The equality is important. It tells us that to analyze the properties of the additive multigrid preconditioner or its variants, it is sufficient to study the space decomposition preconditioner defined in the sequential manner in subsection 4.1.2 can be considered to analyze the space decomposition preconditioner.

The space decomposition preconditioner (4.34) was first analyzed by Bramble, Pasciak, and Xu in [39]. Following this contribution, the preconditioner is also known in the literature as the BPX preconditioner.

Finally, we note that by analogue with AMGM and (4.34), the subspace preconditioner defined in the sequential manner in subsection 4.1.2 can be considered to analyze the V-cycle of the multiplicative multigrid method with $v_1 = 0$, $v_2 = v$ (only post-smoothings).

4.2.3 Local grid refinement

As we already know, an additive multigrid cycle can be analyzed as a linear iteration with a finite element subspace decomposition preconditioner,

$$u^{new}_h = u^{old}_h - B_h r_h = u^{old}_h - \sum_{i=0}^{l} R_h^{i} Q_h^{i} r_h,$$

(4.41)

where $r_h$ is the finite element residual $r_h = A_h u^{old}_h - f_h$. The operator $B_h$ defines the BPX preconditioner.

If we use the Jacobi-type smoother, and if the coarse grid problem is solved exactly, then the second term in (4.41) can be rewritten as

$$B_h r_h = A_h^{-1} Q_h^{(0)} r + \sum_{k=1}^{n_b} \sum_{i=1}^{n_{b,k}} \frac{(r_{h,k}, \phi_i^{(k)})}{a(\phi_i^{(k)}, \phi_i^{(k)})} \phi_i^{(k)},$$

(4.42)

where $n_b$ is the number of basis functions on a $k$th grid level, $\phi_i^{(k)}$ are the basis functions on the $k$th level, and $a(\cdot, \cdot)$ is the bilinear form of the finite element problem formulation.

Assume that the fine element space $U_h$ is obtained from $U_{h-1}$ by local grid refinement (the grid is refined locally, not in the whole domain): on the $k$th level one adds grid nodes $x_i$, with indexes $i = n_{k-1} + 1, \ldots, n_k$, and then the following modification of (4.42) is reasonable:

$$B_r = A_h^{-1} Q_h r + P_l^{-1} \bigg( \sum_{k=1}^{l} \sum_{\phi_i^{(k)} \neq \phi_j^{(k-1)}} \frac{(r_{h,k}, \phi_i^{(k)})}{a(\phi_i^{(k)}, \phi_j^{(k)})} \phi_j^{(k)} \bigg),$$

(4.43)
where the following notation is adopted:

\[
\sum_{j=1}^{n_j} \phi_j^{(k)} := \sum_{i=1}^{n_{k-1}} \phi_i^{(k-1)} + \sum_{i=n_{k-1}+1}^{n_k} \phi_i^{(k)}.
\]

For a level \( k \), the right-hand side in (4.43) collects only basis functions corresponding to new (for the level \( k \)) nodes and nodes neighboring the new nodes, i.e., all new basis functions. On every grid level, only nodes in the refined region, rather than in the whole domain \( \Omega \), are involved in calculations. This can reduce the amount of computations considerably. Indeed, consider the following example: In a two-dimensional domain when passing from a level \( k \) to the level \( k + 1 \) one refines the grid only in a subdomain, whose area is \( \frac{1}{4} \) of the area of a subdomain where the refinement was done passing from the level \( k - 1 \) to the level \( k \). Such situations occur in computations when the solution to a differential problem has large variations in a part of the domain, e.g., the solution has a boundary layer. Hence, it is a common practice to use adaptive grids to approximate such a rapidly changing solution. For the given example, one can easily estimate that the number of unknowns on the \( k \)th level is approximately equal to \( (\frac{3}{2}^k - \frac{1}{2}) n_0 \), where \( n_0 \) is the number of unknowns on the coarsest level. For the constant from (3.53) it holds that \( T_n \rightarrow 2 \) if \( l \rightarrow \infty \). Therefore, the multigrid method from Chapter 3 fails to be of optimal computational complexity for \( \gamma \geq 2 \). In contrast, the additive method in the form (4.41) with the operator \( B \) from (4.43) preserves the optimal complexity for any grid refinement strategy. There are also extensions of multiplicative multigrids for computing on adaptive grids; see [171] and the review paper [15] with comparisons of additive and multiplicative methods.

### 4.2.4 Hierarchical basis preconditioner

Assume piecewise linear conforming finite elements.

In the hierarchical basis method, the application of \( B_h \) is further “simplified” and takes the form

\[
B_h r_h = A_0^{-1} Q_{L}^{(0)} r_h + \sum_{k=1}^{l} \sum_{i=n_{k-1}+1}^{n_k} \frac{r_{ik}(k)}{d(\hat{\phi}_i^{(k)}, \hat{\phi}_i^{(k)})} \hat{\phi}_i^{(k)},
\]

which corresponds to the choice of \( \forall_k = U_k - U_{k-1} \) in the space decomposition method. The difference \( U_k - U_{k-1} \) denotes a subspace of \( U_k \) spanned over nodal basis functions, corresponding to nodes added on the new mesh level. For a piecewise linear conforming finite element method, \( U_k - U_{k-1} \) is the \( H^1 \)-orthogonal complement of \( U_{k-1} \) in \( U_k \).

On a new, say \( k \)th, level only new nodes with respect to the previous \((k - 1)\)th level are involved in calculations. In this case, the formula for calculations becomes very simple and convenient for parallel implementation. Basis functions in (4.44) can be reordered, and (4.44) can be rewritten as

\[
B_h r_h = A_0^{-1} Q_{L}^{(0)} r_h + \sum_{i=n_{k-1}+1}^{n_k} \frac{r_{ik}(k)}{d(\hat{\phi}_i^{(k)}, \hat{\phi}_i^{(k)})} \hat{\phi}_i^{(k)}.
\]

For the one-dimensional case and linear finite elements, the structure of the hierarchical basis \( \hat{\phi}_i^{(k)} \) is illustrated in Figure 4.2. Compare it to the nonhierarchical basis in Figure 4.3.

In the next section, using the example of the two-dimensional Poisson problem we show that the hierarchical basis leads to the preconditioner \( B_h \), which is optimal up to a logarithmic factor. More about the hierarchical basis approach can be learned from the review paper [225].
Exercises

4.2.1. Show that the additive two-grid iteration matrix is defined by

\[ M_s = I - (\theta p A_H^{-1} r + (I - S^v) A_b^{-1}) A_h, \]

where \( S \) is the iteration matrix of the smoothing iterations, and the coarse grid problem is solved exactly.

4.2.2. Using the decomposition introduced in subsection 4.2.2, prove that the sequential subspace iteration (4.8) is equivalent to the V-cycle of the multiplicative multigrid with \( v_1 = 0, v_2 = v \).
4.2.3. For $\mathbb{V} = \mathbb{R}^n$ and the decomposition $\mathbb{R}^n = \sum_{i=1}^n \text{span}\{\hat{e}_i\}$, where $\hat{e}_i$ are the unit basis vectors, and $R_i = A_i^{-1}$, check that (4.8) defines the Gauss–Seidel iteration.

4.2.4. The finite element space decomposition (4.32) and the diagram (4.28) induce the decomposition of $\mathbb{V}$: $\mathbb{V} = \mathbb{R}_h$, into subspaces $\mathbb{V}_k = \text{range} p_k$, i.e., $\mathbb{V}_k$ is the range of $p_k$. What are the matrix analogues of $L^2$-orthogonal projection and Jacobi subspace preconditioners in this case?

4.2.5. Describe a hierarchical basis for the uniform triangulation of a unit square (see Figure 3.8) with $b_k = (1/2)^{k+1}$.

4.3 Domain decomposition methods

Similar to grid decomposition, the domain decomposition approach mostly applies to building preconditioners for algebraic systems arising from discretization of partial differential equations. However, now vector space decomposition is based on the division of a domain, where a partial differential equation is posed, into smaller subdomains.

The history of domain decomposition methods can be traced back to the paper of Schwarz [178]. In that paper, the method, nowadays known as the alternating Schwarz method, was introduced for an entirely theoretical purpose: Proving the well-posedness of the Poisson problem in a domain which can be decomposed as the union of two “simpler” domains, where the problem is known to have a unique solution. In 1970s and 1980s the idea of Schwarz was put into a computational framework and led to a variety of numerical techniques known as domain decomposition methods and preconditioners. The approach gained significant attention, and nowadays a vast literature exists on the subject including a dozen of books; see [143, 193, 165, 218, 148] among the recent ones and [127] for one of the first to cover the subject. While here we consider the domain decomposition method as an iterative procedure or a preconditioning technique, in computational mathematics the approach is more general purpose and includes building discretizations in subregions with nonmatching grids (see, e.g., [119, 105]), and approximation of heterogeneous equations by simpler models coupled on interfaces between subregions (see, e.g., [16]).

4.3.1 Alternating Schwarz method

The discussion of several basic domain decomposition techniques begins with recalling the classical alternating Schwarz method.

We are looking for the solution to the Poisson equation

$$-\Delta u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{in } \partial \Omega,$$

where $\Omega$ is a bounded connected domain in $\mathbb{R}^N$, with sufficiently regular (at least Lipschitz) boundary $\partial \Omega$, and $f \in L^2(\Omega)$. Assume that $\Omega$ can be decomposed into two (overlapping) subdomains $\Omega_1$ and $\Omega_2$: $\Omega = \Omega_1 \cup \Omega_2$, and we denote $\Gamma_1 := \partial \Omega_1 \cap \Omega_2$, $\Gamma_2 := \partial \Omega_2 \cap \Omega_1$.

The decomposition and notation are illustrated in Figure 4.4: $\Omega_1$ is a rectangle and $\Omega_2$ is a circle.

The alternating Schwarz method defines a sequence of functions, $u^0, u^1, \ldots$, in $H_0^1(\Omega)$ through the successive solution of the Poisson equations in the subdomains $\Omega_1$ and $\Omega_2$. 
4.3. Domain decomposition methods

Let $u^n \in H^1_0(\Omega)$ be given; then $u^{n+1}$ is found in two steps:

**Step 1:** Solve for $u_1$

$$-\Delta u_1 = f \text{ in } \Omega_1, \quad u_1 = u^n \text{ on } \partial \Omega_1. \quad (4.47)$$

**Step 2:** Solve for $u_2$

$$-\Delta u_2 = f \text{ in } \Omega_2, \quad u_2 = u_1 \text{ on } \partial \Omega_2. \quad (4.48)$$

Define $u^{n+1}$ (explain why $u^{n+1} \in H^1_0(\Omega)$) as

$$u^{n+1} = \begin{cases} u_2 & \text{in } \Omega_2, \\ u_1 & \text{in } \Omega \setminus \Omega_2. \end{cases} \quad (4.49)$$

We shall show that under certain assumptions on how $\Omega_1$ and $\Omega_2$ overlap, the sequence $u^n$ converges to the solution of (4.46) in $H^1_0(\Omega)$. The assumptions will be essentially the same as those made by Sobolev in [185] and Lions in [133].

It is instructive to see how the alternating Schwarz method fits the framework of space decomposition methods from subsection 4.1.1 and so defines a preconditioner.

**4.3.2 Fitting alternating Schwarz to space decomposition framework**

Obviously, we can rewrite (4.47)–(4.49) in terms of residuals and corrections: Given $u^n \in H^1_0(\Omega)$, solve for $e_1 \in H^1_0(\Omega_1)$

$$-\Delta e_1 = f + \Delta u^n \text{ in } \Omega_1, \quad e_1 = 0 \text{ on } \partial \Omega_1. \quad (4.50)$$

Extend $e_1$ by zero on $\Omega$ and set $u_1 = u^n - e_1$. Further, solve for $e_2 \in H^1_0(\Omega_2)$

$$-\Delta e_2 = f + \Delta u^n \text{ in } \Omega_2, \quad e_2 = 0 \text{ on } \partial \Omega_2. \quad (4.51)$$

Extend $e_2$ by zero on $\Omega$ and set $u^{n+1} = u_1 - e_2$.

To have a more formal setup, let $\mathcal{V} = H^1_0(\Omega)$, with $(u, v)_\mathcal{V} := (\nabla u, \nabla v)_{L^2(\Omega)}$, and define two closed subspaces

$$\mathcal{V}_k := \{ v \in \mathcal{V} \mid v = 0 \text{ on } \Omega \setminus \Omega_k \}, \quad k = 1, 2.$$

Equations (4.50)–(4.51) can be written in the weak form:

- **Find** $e_1 \in \mathcal{V}_1$ : $(e_1, v)_\mathcal{V} = (f, v)_{L^2} - (u^n, v)_\mathcal{V} \quad \forall \ v \in \mathcal{V}_1$, 
  $u_1 = u^n - e_1$,

- **Find** $e_2 \in \mathcal{V}_2$ : $(e_2, v)_\mathcal{V} = (f, v)_{L^2} - (u_1, v)_\mathcal{V} \quad \forall \ v \in \mathcal{V}_2$, 
  $u^{n+1} = u_1 - e_2$. 

Figure 4.4. Decomposition of $\Omega$ into two overlapping domains for the alternating Schwarz method.
Denote by $A : \mathcal{V} \to \mathcal{V}'$ and $A_k : \mathcal{V}_k \to \mathcal{V}'_k$ the natural isomorphisms (based on the Riesz representation theorem) between the spaces $\mathcal{V}$, $\mathcal{V}_k$ and their dual spaces. We may consider the chain

$$\mathcal{V} \xrightarrow{A} \mathcal{V}' \subset \mathcal{V}'_k \xrightarrow{A_k^{-1}} \mathcal{V}_k \subset \mathcal{V}. \tag{4.52}$$

Now, using the formalism of (4.52), we rewrite (4.50)–(4.51), and thus the alternating Schwarz method, in the form of subspace iterations (compare with (4.8)): Set $u_0 = u^n$ and compute

$$u_k = u_{k-1} - A_k^{-1}(A u_{k-1} - f), \quad k = 1, 2. \tag{4.53}$$

Set $u_{n+1} = u_n$.

To treat the alternating Schwarz method as the sequential space decomposition iteration, we still need to verify that

$$\mathcal{V} = \mathcal{V}_1 + \mathcal{V}_2. \tag{4.54}$$

One easily verifies (see [133]) that (4.54) follows from the existence of a partition of unity:

$$\chi_1 + \chi_2 = 1 \text{ on } \Omega, \quad \chi_k = 0 \text{ on } \Gamma_k, \quad \chi_k \geq 0, \quad k = 1, 2,$$

such that

$$\chi_k \in W^{1,\infty}_{\text{loc}}(\Omega) \quad \text{and} \quad |\nabla \chi_k(x)| \leq C \left( \text{dist}(x, \partial \Omega) \right)^{-1} \text{ on } \Omega, \quad k = 1, 2. \tag{4.55}$$

Here we used the notion $W^{1,\infty}_{\text{loc}}(\Omega)$ to denote a set of functions $u$ which are from $W^{1,\infty}(\Omega')$ on any compact subset $\Omega'$ of $\Omega$. The required condition (4.55) on a partition of unity imposes certain restrictions on how $\Omega_1$ and $\Omega_2$ may overlap.

In previous sections, we analyzed space decomposition methods in spaces of finite dimension. Without much effort the analysis can be extended to Hilbert spaces, so the convergence of (4.47)–(4.49) can be proved rigorously. We shall not do this here, since ultimately the domain decomposition method will be applied to discrete (and thus finite-dimensional) problems.

The immediate consequences of putting the alternating Schwarz method into the space decomposition framework are the following: (i) one can easily extend the method (and analysis) to the case of multiple domain overlap; (ii) one is prompted to consider the parallel version of the method; (iii) one can use a domain decomposition approach to build a preconditioner.

### 4.3.3 Multiplicative and additive Schwarz preconditioners

To define preconditioners based on a domain decomposition technique, we now return to the discrete setting and consider a finite element method for an elliptic problem. Let $\mathcal{T}_b$ be a regular triangulation of a bounded domain $\Omega$. We assume that $\Omega$ is decomposed in a union of $l$ overlapping domains $\Omega_k$:

$$\Omega = \bigcup_{k=1}^l \Omega_k$$

such that for any $\Omega_k$, there is a subdomain $\Omega_j$, $j \neq k$, such that $\Omega_k \cap \Omega_j \neq \emptyset$. The boundaries of subdomains are assumed to align with the mesh. Hence for a subdomain $k$ we have a separate triangulation $\mathcal{T}_k \subset \mathcal{T}_b$. For the purpose of analysis, we also have to consider a
4.3. Domain decomposition methods

regularity of the decomposition in terms of partition of unity properties: We assume that there exists \( \chi_k \in W_{1,\infty}^1(\Omega), k = 1, \ldots, l, \) such that

\[
\sum_{k=1}^{l} \chi_k = 1 \quad \text{on} \quad \Omega, \quad \chi_k = 0 \quad \text{on} \quad \partial \Omega_k \setminus \partial \Omega, \quad \chi_k \geq 0, \quad k = 1, \ldots, l,
\]

and

\[
|\nabla \chi_k(x)| \leq C_k \left( \text{dist}(x, \partial \Omega) \right)^{-1} \quad \text{on} \quad \Omega, \quad k = 1, \ldots, l, \tag{4.56}
\]

with some finite constants \( C_k. \)

Denote \( C = \max_k C_k. \)

Let \( U_h \subset H^1_0(\Omega) \) be a space of finite element functions (vanishing on \( \partial \Omega \)) with respect to the triangulation \( \mathcal{T}_h, \) and denote by \( U_k \) a subspace of \( U_h \) of functions vanishing on \( \Omega \setminus \Omega_k: \)

\[
U_k := \{ v_k \in U_h | v_k = 0 \quad \text{on} \quad \Omega \setminus \Omega_k \}.
\]

Then we have the decomposition

\[
U_h = U_1 + \cdots + U_l. \tag{4.57}
\]

Indeed, to see that \( U_1 + \cdots + U_l \) spans the whole space \( U_h, \) note that any \( v_h \in U_h \) can be decomposed into the sum

\[
v_h = \sum_{k=1}^{l} v_k, \quad \text{where} \quad v_k = I_h(\chi_k v_h) \in U_k, \tag{4.58}
\]

and \( I_h \) is the nodal interpolation operator.

Due to the canonical linear isomorphism between the space of finite element functions \( U_h \) and the Euclidian space of the coefficients \( \mathbb{R}_h, \) the decomposition (4.57) naturally implies the subspace decomposition

\[
\mathbb{R}_h =: \mathbb{V} = V_1 + \cdots + V_l. \tag{4.59}
\]

The space decomposition (4.59) is defined by the domain decomposition. For this space decomposition, (4.2), the preconditioners defined in subsections 4.1.1 and 4.1.2, are known as additive and multiplicative Schwarz preconditioners. Setting \( R_k = A_k \) for subspace solvers corresponds to the exact solution of the elliptic finite element problem with homogeneous boundary conditions in each subdomain, as was the case for the classical alternating Schwarz method. Otherwise, we may assume that the subgrid problems are solved only approximately, and thus \( R_k \) are local preconditioners for subspace matrices \( A_k. \) In this case, we assume that

\[
c_d R_k^{-1} \leq A_k \leq C_d R_k^{-1} \tag{4.60}
\]

holds, with some positive constants \( c_d, C_d, \) independent of the number of subdomains \( l \) and a mesh size.

4.3.4 Nonoverlapping domain decomposition preconditioners

Sometimes, it is more convenient to decompose a domain \( \Omega \) into the union of nonintersecting subdomains \( \Omega_i. \) Such splitting defines a class of iterations and preconditioners different from the alternating Schwarz method. Yet, a basic ingredient remains the same: one should be able to apply fast solvers for subproblems posed in each subdomain separately. For the sake of presentation, we start with the case of two subdomains. Assume
The complement of the simplest way to complete the decomposition is to define a third subspace as an orthogonal complement of \( V_1 + V_2 \) in \( V \):

\[
V_3 := \{ v \in V \mid (v, w) = 0 \ \forall w \in V_1 + V_2 \}.
\]

It is easy to see that \( V_3 \) is a subspace of all functions from \( H^1(\Omega) \) harmonic in \( \Omega_1 \) and \( \Omega_2 \).

Following the formalism of subsection 4.3.2 (see (4.52)) one may define a space decomposition preconditioner \( B : V' \rightarrow V \) based on the decomposition

\[
V = V_1 + V_2 + V_3.
\] (4.61)

For the Poisson problem this space decomposition preconditioner is an ideal one, i.e., \( B = A^{-1} \) (check this!). There is, however, one problem with applying \( B = \sum_{k=1}^{3} A_k^{-1} \): while subproblems with \( A_k, k = 1, 2 \), reduce to the Poisson equation in each subdomain, handling \( A_3^{-1} \) is not that easy.

Let us study some properties of \( A_3 \). Consider the Poisson equations (4.46) and decompose the solution \( u = u_1 + u_2 + v_3 \) according to (4.61). It holds that

\[
-\Delta u_k = f \ \text{in } \Omega_k, \quad u_k = 0 \ \text{in } \Omega \setminus \Omega_k \quad \text{for } k = 1, 2,
\] (4.62)

and by the definition of the subspace problem, we have

\[
(A_3 u_3, w)_{V'} = (\nabla u_3, \nabla w) = (f, w) \quad \forall w \in V_3.
\] (4.63)

For a function \( f \) defined on \( \Omega \), denote by \( f|_{\Omega_i} \) the restriction of \( f \) on \( \Omega_i \); then \( [f] = (f|_{\Omega_1})_i - (f|_{\Omega_2})_i \) defines a "jump" of \( f \) over interface \( \Gamma \). For continuous functions (and also for elements of \( H^1(\Omega) \)), it obviously holds that \( [f] = 0 \). Using the fact that functions from \( V_3 \) are harmonic in each subdomain and applying integration by parts, we get

\[
(\nabla u_3, \nabla w) = \int_{\Gamma} \left[ \frac{\partial u_3}{\partial n} \right] w \, ds,
\]

\[
(f, w) = -\sum_{k=1}^{2} (\Delta u_k, w) = -\int_{\Gamma} \left[ \frac{\partial (u_1 + u_2)}{\partial n} \right] w \, ds.
\]

Here we assume that \( n \) is a normal vector on \( \Gamma \) pointing from \( \Omega_1 \) to \( \Omega_2 \). This and (4.63) yield the equality

\[
\int_{\Gamma} \left[ \frac{\partial u_3}{\partial n} \right] w \, ds = -\int_{\Gamma} \left[ \frac{\partial (u_1 + u_2)}{\partial n} \right] w \, ds.
\] (4.64)

We see that the subspace equation for \( V_3 \) can be written in terms of interface quantities. This is not a surprise, since any function from \( V_3 \) is completely determined by its values on \( \Gamma \). To define this correspondence formally, consider a space of all traces of functions from \( V = H^1_0(\Omega) \) on \( \Gamma \):

\[
\Lambda = \{ \eta \in L^2(\Gamma) \mid \eta = v|_{\Gamma} \ \text{for some } v \in V \}.
\]
4.3. Domain decomposition methods

With an appropriately defined norm, $\Lambda$ is a Banach space, which can be identified with a fractional order Sobolev space on $\Gamma$; see, e.g., [165]. For every $\eta \in \Lambda$, denote by $\mathcal{H}\eta$ its harmonic extension vanishing on $\partial\Omega$, i.e., $v = \mathcal{H}\eta$ is an element of $\mathbb{V}_1$ such that $\eta = v|_{\Gamma}$. Let $F = -\left[\frac{\partial(u_1+u_2)}{\partial n}\right]$. It can be shown that $F \in \Lambda'$. Now (4.64) can be written as

$$S\eta = F, \quad u_3 = \mathcal{H}\eta, \quad \text{with} \quad S = \left[\frac{\partial(\mathcal{H}\cdot)}{\partial n}\right].$$

(4.65)

Here $S$ is a self-adjoint (with respect to the $L^2$-duality) positive definite operator acting from $\Lambda$ onto $\Lambda'$; $S$ is well known in the literature as the Poincare–Steklov operator; see, e.g., [2].

Summarizing, the solution $u$ to (4.46) posed in the entire domain $\Omega$ can be decomposed as $u = u_1 + u_2 + u_3$, where $u_1$ and $u_2$ are solutions to the Poisson equations in each subdomain in (4.62), and $u_3 = \mathcal{H}\eta$, where $\eta$ solves (4.65). Therefore, to build a preconditioner, one has to define (approximate) solvers for subdomain problems and a preconditioner for $S$. We already discussed several options for building preconditioners for subdomain problems (e.g., based on a multigrid method). Now we shall take a closer look at how to define a preconditioner for $S$. To this end, we first consider the matrix counterpart of the Poincare–Steklov operator and then discuss a few well-known preconditioning strategies.

4.3.5 A matrix form of the Poincare–Steklov operator

Similar to the overlapping Schwarz method in subsection 4.3.3, we consider a conforming triangulation $\mathcal{T}_h$ of $\Omega$, which is consistent with the decomposition of $\Omega$ into two subdomains, i.e., the interface $\Gamma$ is matched by the triangulation. A finite element method for (4.46) leads to the system of linear algebraic equations

$$Au = b, \quad (4.66)$$

with matrix elements and the right-hand side vector

$$a_{ij} = \int_{\Omega} \nabla \varphi_i \cdot \nabla \varphi_j \, dx, \quad b_i = \int_{\Omega} f \varphi_i \, dx$$

defined by the choice of a nodal basis $\varphi_i, i = 1, \ldots, N$.

We decompose all internal nodes of $\mathcal{T}_h$ (nodes not on $\partial\Omega$) into three sets. The sets $\mathcal{T}_k, k = 1, 2$, contain all internal nodes for $\Omega_1$ and $\Omega_2$, respectively, and all nodes lying on $\Gamma$ form the third set $\mathcal{T}_\Gamma$. This induces the corresponding decomposition of the basis functions $\varphi_i$ into three sets. We decompose the finite element space $\mathbb{U}_h$ as

$$\mathbb{U}_h = \mathbb{U}_1 + \mathbb{U}_2 + \mathbb{U}_\Gamma, \quad (4.67)$$

where

$$\mathbb{U}_k = \text{span}\{\varphi_i \mid \varphi_i \in \mathcal{T}_k\}, \quad k = 1, 2,$$

$$\mathbb{U}_\Gamma = \{v \in \mathbb{U}_h \mid (v, w) = 0 \quad \forall \ w \in \mathbb{U}_1 + \mathbb{U}_2\}.$$

This is the discrete analogue of (4.61). If we enumerate the unknowns such that degrees of freedom from $\mathbb{U}_1$ come first, then the interface degrees of freedom, and finally degrees of freedom from $\mathbb{U}_2$, then the matrix from (4.66) has the block structure. In the block form, the system (4.66) is

$$\begin{bmatrix}
A_{11} & A_{1\Gamma} & 0 \\
A_{\Gamma 1} & A_{\Gamma \Gamma} & A_{\Gamma 2} \\
0 & A_{2\Gamma} & A_{22}
\end{bmatrix}
\begin{bmatrix}
u_1 \\
u_{\Gamma} \\
u_2
\end{bmatrix} =
\begin{bmatrix}
f_1 \\
f_{\Gamma} \\
f_2
\end{bmatrix}.$$
Note that the $u_i$ part of the vector of coefficients corresponds only to the mesh nodes lying on the interface $\Gamma$, while finite element functions from $\Omega_1$ have global support in $\Omega$—these are discretely harmonic functions in $\Omega_1$ and $\Omega_2$.

The matrix analogue of the Poincare–Steklov equation (4.65) is given by eliminating unknowns $u_1$ and $u_2$ from the system (4.68) by the block Gauss method. This leads us to the following system for $u_1$:

$$ (A_1 - A_{11}^{-1}A_{11} - A_{12}^{-1}A_{22})u_1 = f_1 - A_{11}^{-1}f_1 - A_{12}^{-1}f_2. $$

(4.69)

The matrix on the left-hand side of (4.69),

$$ S = A_1 - A_{11}^{-1}A_{11} - A_{12}^{-1}A_{22}, $$

is the analogue of the Poincare–Steklov operator and is called the Schur complement matrix for $u_1$.

We can explicitly write in matrix notation the space decomposition preconditioner for $A$ that is based on the decomposition (4.67):

$$ B_A := \begin{bmatrix}
A_{11}^{-1} + A_{11}^{-1}A_{11}S^{-1}A_{11}^{-1} & -A_{11}^{-1}A_{11}S^{-1} & 0 \\
-S^{-1}A_{11}^{-1} & S^{-1} & -S^{-1}A_{22}^{-1} \\
0 & -A_{22}^{-1}S^{-1} & A_{22}^{-1} + A_{22}^{-1}S^{-1}A_{11}^{-1}
\end{bmatrix}. $$

(4.70)

By straightforward computations, one may check that $B_A$ is the ideal preconditioner for $A$, i.e., $B_A = A^{-1}$. This also follows from the fact that the decomposition (4.67) is $V$-orthogonal.

To obtain a practical preconditioner, one should further replace $A_{11}^{-1}$, $A_{22}^{-1}$, and $S^{-1}$ in the definition of $B_A$ by some preconditioners for local problem solvers and the interface Schur complement.

### 4.3.6 A Neumann–Neumann preconditioner

Assume for a moment that solving linear systems with $A_{11}$ and $A_{22}$ is easy. If $u_1$ is found from (4.69), the missing components $u_1$ and $u_2$ of the vector of unknowns are readily computed from (4.68). The matrix $S$ is not a sparse matrix. However, it accounts only for interface degrees of freedom, and so its size is typically much smaller than the size of the original system. Thus, if the size is not too big, then an exact factorization of $S$ is feasible. Otherwise, one should look for a preconditioner to $S$. We shall consider several preconditioning approaches based on solvers for subdomain problems.

Let $\varphi_i \in U_1$ and $\varphi_j \in U_2$ be nodal basis functions for some $i, j$. Then $a_{ij} = \int_{\Omega} \nabla \varphi_i \cdot \nabla \varphi_j \, dx$ is an entry of $A_1$. We split each entry as follows:

$$ a_{ij} = a_{ij}^{(1)} + a_{ij}^{(2)}, \quad \text{with } a_{ij}^{(k)} = \int_{\Omega} \nabla \varphi_i \cdot \nabla \varphi_j \, dx. $$

This implies the splitting of $A_1$ and $S$:

$$ A_1 = A_1^{(1)} + A_1^{(2)}, \quad S = S_1 + S_2, \quad \text{with } S_k = A_1^{(k)} - A_{1k}A_{k1}^{-1}A_{k1}, \ k = 1, 2. $$

One possible choice for $S$ is the matrix $S_1$ (alternatively $S_2$). In subsection 4.4.4 we prove the spectral equivalence of $S$ and $S_1$. We note that applying $S_1^{-1}$ to a vector $v_1$ is equivalent to solving a finite element problem in a subdomain. Indeed, $u_1 = S_1^{-1}v_1$ is the interface part of the solution $\{u_1, u_1\}^T$ to the following system (check this!):

$$ \begin{bmatrix}
A_{11} & A_{11} \\
A_{11} & A_{11}^{(1)}
\end{bmatrix} \begin{bmatrix}
u_1 \\
u_1^{(1)}
\end{bmatrix} = \begin{bmatrix}
0 \\
v_1
\end{bmatrix}. $$

(4.71)
This linear algebraic system corresponds to the finite element discretization of the Poisson problem posed in \( \Omega \), with mixed Neumann boundary conditions on the interface:

\[-\Delta u = 0 \quad \text{in} \ \Omega, \quad u|_{\partial\Omega_1 \setminus \Gamma} = 0, \quad \frac{\partial u}{\partial n}|_{\Gamma} = v.\]

Such a localization property does not hold for the solution of a system with full Schur complement, i.e., finding \( u |_{\Gamma} = S^{-1} v |_{\Gamma} \). The latter cannot be reduced to solving few subdomain problems! Thus the application of this domain decomposition preconditioner reduces to solving decoupled subdomain problems with Dirichlet boundary conditions on \( \Gamma \) to update variables in a subdomain and to solving a subdomain problem with Neumann boundary condition on \( \Gamma \) to update the interface variable.

A straightforward extension of the above preconditioner is done by including both \( S^{-1} \) and \( S^{-1} \) in the Schur complement preconditioner, hence accounting for each subdomain. Such a version is known in the literature as the Neumann–Neumann preconditioner. The Neumann–Neumann preconditioner is obviously extended for the case of \( l > 2 \) subdomains: For subdomain \( \Omega_k \), let \( \tilde{A}_{kl} \) be a submatrix consisting of only the non-zero rows of \( A_{kl} \), i.e., \( \tilde{A}_{kl} \) accounts for only those interface nodes which belong to \( \partial \Omega_k \). Similarly, we define \( \tilde{A}_{k1} \) and \( \tilde{A}_{kl}^{(k)} \). Let \( S_k = \tilde{A}_{k1}^{(k)} - \tilde{A}_{k1}^{-1} \tilde{A}_{kl} \). The Neumann–Neumann preconditioner \( B_S \) for the Schur complement \( S \) is defined by

\[ B_S = \sum_{k=1}^{l} W_k S_k^{-1} W_k, \]

where \( W_k \) are diagonal matrices. For every interface node the diagonal element of \( W_k \) is set as \( 1/m \), where \( m \) is the number of subdomains sharing the node.

### Exercises

4.3.1. Derive a variant of an alternating Schwarz method from subsection 4.3.1 for more than two overlapping subdomains. Does the result depend on the order in which the subdomains are visited?

4.3.2. Consider a preconditioner \( B : \mathcal{V} \to \mathcal{V} \) based on the decomposition (4.61). Show that \( B = A^{-1} \).

4.3.3. Prove that the operator \( S \) in (4.65) is a self-adjoint positive definite operator.

4.3.4. Check that \( B_A \) from (4.70) is the ideal preconditioner, i.e. \( B_A = A^{-1} \).

### 4.4 Convergence analysis for the Poisson problem

We analyze the convergence of the conjugate gradient method with the BPX, hierarchical basis, and domain decomposition preconditioners for the model example of the Poisson equation.

For the sake of notation, in this section we denote by \( (\cdot, \cdot)_{\tau} \) and \( \| \cdot \|_{\tau} \) the \( L^2 \)-scalar product and the \( L^2 \)-norm of a function over subdomain \( \tau \subset \Omega \) and, as before, \( (\cdot, \cdot) = (\cdot, \cdot)_{\Omega} \) and \( \| \cdot \| = \| \cdot \|_{\Omega} \).
4.4.1 BPX preconditioner

Consider the additive multigrid method as introduced in subsection 4.2.1 with $\theta = 1$ and the Jacobi method for smoothing iterations. In subsection 4.2.2 we discussed that the method can be analyzed in the framework of a space decomposition preconditioner for a finite element space, also known as the BPX preconditioner. Thus denote by $A_h$ and $B_h$ the finite element bilinear form and preconditioner operators on $\mathbb{U}_h$ as they are defined in subsection 4.2.2. Because of (4.40), we are interested in an estimate on $\text{cond}(B_hA_h)$. For a general space decomposition preconditioner, such an estimate is given by Theorem 4.3.

Assume that the conditions on the grid are the same as in Section 3.4 and the finite element spaces $\mathbb{U}_h$ consists of piecewise linear continuous finite elements with respect to a triangulation $\mathcal{T}_h$ of a domain $\Omega$.

**BPX preconditioner: $\lambda_{\text{max}}(BA)$**

First, we prove an estimate on the maximum eigenvalue of the preconditioned operator $B_hA_h$. For the Jacobi smoothings method, we have (see Section 3.4)

$$c_1\|v_k\|^2 \leq (\{R_b^{(k)}\}^{-1}v_k, v_k) \leq c_2\|v_k\|^2,$$

(4.72)

$$c_3(A_h v_k, v_k) \leq (\{R_b^{(k)}\}^{-1}v_k, v_k)$$

(4.73)

for any $v_k \in \mathbb{U}_h$, $k = 0, \ldots, l$, with constants $c_0, c_1, c_2$, independent of $k$ and $v_k$.

We need following auxiliary lemma.

**Lemma 4.5.** Let $0 \leq k \leq m \leq l$; then for any $u \in \mathbb{U}_k$, $v \in \mathbb{U}_m$ the following estimate is valid:

$$|\langle \nabla u, \nabla v \rangle| \leq c \sqrt{\frac{b_m}{b_k}} \|\nabla u\| h^{-1}_m \|v\|.$$

(4.74)

**Proof.** For $k = m$ the assertion of the lemma follows from the Cauchy inequality and the inverse inequality for finite element functions (see subsection 3.4.1). Letting $k < m$, we take arbitrary element $\tau$ (triangle) of the triangulation $\mathcal{T}_h$, $\text{diam}(\tau) \sim b_h$. Let us check (4.74) on this triangle: We want to show the estimate

$$|\langle \nabla u, \nabla v \rangle_\tau| \leq c \sqrt{\frac{b_m}{b_k}} \|\nabla u\|_\tau h^{-1}_m \|v\|_\tau.$$

(4.75)

Once proved, the local estimate (4.75) implies the global estimate in (4.74). Indeed, taking a sum of (4.75) for all triangles from $\mathcal{T}_h$, we get

$$|\langle \nabla u, \nabla v \rangle| = \left| \sum_{\tau \in \mathcal{T}_h} \langle \nabla u, \nabla v \rangle_\tau \right| \leq \sum_{\tau \in \mathcal{T}_h} \sqrt{\frac{b_m}{b_k}} \|\nabla u\|_\tau h^{-1}_m \|v\|_\tau \left( \sum_{\tau \in \mathcal{T}_h} \|v\|_\tau^2 \right)^{\frac{1}{2}} \leq \sqrt{\frac{b_m}{b_k}} \|\nabla u\| h^{-1}_m \|v\|.$$

It remains to prove (4.75). Decompose $v$ into the sum $v = v_0 + v_1$, where $v_0$ is a piecewise linear function from $\mathbb{U}_m$ such that $v_0$ coincides with $v$ in all grid nodes of $T_m$.
which belong to $\partial \tau$ (the boundary of the triangle $\tau$); in all other grid nodes of $T_m$ the function $v_0$ equals 0. Then $v_1 = 0$ on $\partial \tau$ and, integrating by parts, we have

$$\langle \nabla u, \nabla v_1 \rangle_\tau = \langle (\Delta u, v_1) \rangle_\tau = 0.$$  

$\Delta u = 0$ on $\tau$, since $u$ is linear on $\tau$. Let $S$ be a (“next-to-boundary”) part of $\tau$, where $v_0$ does not equal zero; then

$$\langle \nabla u, \nabla v_0 \rangle_\tau = \langle (\nabla u, \nabla v_0) \rangle_S \leq \|\nabla u\|_S \|\nabla v_0\|_S.$$  

Since $u$ is linear on $\tau$, the components of $\nabla u$ are constants on $\tau$ and

$$\|\nabla u\|^2_S = \frac{\text{area}(S)}{\text{area}(\tau)} \|\nabla u\|^2_\tau = c \frac{b_m}{b_k} \|\nabla u\|^2_\tau.$$  

Due to the inverse estimate we also get

$$\|\nabla v_0\|_S \leq c h_{m}^{-1} \|v_0\|_S \leq c h_{m}^{-1} \|v\|_\tau.$$  

Therefore, the local estimate (4.75) follows from

$$\langle \nabla u, \nabla v \rangle_\tau = \langle (\nabla u, \nabla v_0) \rangle_\tau \leq c \sqrt{\frac{b_m}{b_k}} \|\nabla u\|_\tau h_{m}^{-1} \|v\|_\tau. \quad \square$$  

In terms of operators, the estimate (4.74) is written

$$(A_k v_k, v_m) \leq c \sqrt{\frac{b_m}{b_k}} (A_k v_k, v_k)^{1/2} h_{m}^{-1} \|v_m\| \quad \forall v_k \in U_k, v_m \in U_m, k \leq m. \quad (4.76)$$  

Combining (4.76) with (4.72) and (4.73), we get

$$(A_k v_k, v_m) \leq c \sqrt{\frac{b_m}{b_k}} ([R_b^{(k)}]^{-1} v_k, v_k)^{1/2} ([R_b^{(m)}]^{-1} v_m, v_m)^{1/2}, \quad k \leq m.$$  

Thus we proved the estimate (4.11) with

$$\gamma_{ij} = c \min \left\{ \sqrt{\frac{b_i}{b_j}}, \sqrt{\frac{b_j}{b_i}} \right\}.$$  

Our assumptions on the triangulation and refinement from (3.85) imply $\frac{b_i}{k} \leq c (\frac{1}{k})^{j-i}$. 

The spectral radius $\rho(G)$ of the matrix $G$ can be estimated by its $\infty$-norm as

$$\rho(G) \leq \|G\|_\infty = \max_i \sum_{j=1}^l |\gamma_{ij}| \leq c \frac{2\sqrt{2}}{\sqrt{2} - 1}.$$  

Since $\lambda_{\max}(BA) \leq c \rho(G)$ (see (4.19)), we proved an upper bound for $\lambda_{\max}(BA)$, independent of $b_l$ and $l$. 

**BPX preconditioner: $\lambda_{\min}(BA)$**

Assume we are given an arbitrary element $v \in U_k$ and we are looking for a decomposition such that (4.17) holds. Let $u_k \in U_k$ be an $H^1_0$-orthogonal projection of $u$ to $U_k$:

$$\langle \nabla u_k, \nabla \psi \rangle = \langle \nabla u, \nabla \psi \rangle, \quad \forall \psi \in U_k. \quad (4.77)$$
Thanks to our assumptions on triangulation, and if the domain $\Omega$ is such that the condition of $H^2$-regularity holds, the Aubin–Nitsche lemma (see the literature on the finite element method, e.g., [61], and our Exercise 3.1.7 for the one-dimensional case) provides us with the estimate

$$\|u - u_k\| \leq c h_k \|\nabla (u - u_k)\|.$$  

(4.78)

Now we consider the following set of functions $v_k \in \mathbb{U}_k$:

$$v_0 = u_0, \quad v_k = u_k - u_{k-1}, \quad k = 1, \ldots, l.$$  

We obtain the decomposition

$$u = \sum_{k=0}^{l} v_k.$$  

(4.79)

With this decomposition we plan to verify (4.17). Since $\mathbb{U}_{k-1} \subset \mathbb{U}_k$, the identity (4.77) implies for all $k = 1, \ldots, l$,

$$(\nabla u_{k-1}, \nabla \psi) = (\nabla u_k, \nabla \psi) \quad \forall \psi \in \mathbb{U}_{k-1}.$$  

(4.80)

Therefore, similarly to (4.78), we obtain

$$\|u_k - u_{k-1}\| \leq c h_{k-1} \|\nabla (u_k - u_{k-1})\| \leq c_1 h_k \|\nabla (u_k - u_{k-1})\|.$$  

(4.81)

Moreover, setting $\psi = v_m$ for $m < k$ in (4.80), we get the orthogonality property

$$(\nabla v_k, \nabla v_m) = 0 \quad \forall \ k \neq m.$$  

(4.82)

The estimate (4.81) can be rewritten as

$$\|v_k\| \leq c h_k \|\nabla v_k\|.$$  

Assume that $h_0 \sim O(1)$; then for $v_0 \in \mathbb{U}_0$ the estimate $\|v_0\| \leq c h_0 \|\nabla v_0\|$ directly follows from the Friedrichs inequality. The last inequality can be written as

$$\|v_k\| \leq c h_k (A_h v_k, v_k).$$  

(4.83)

Due to the orthogonality property (4.82), which takes the form

$$(A_h v_k, v_m) = 0 \quad \forall \ k \neq m,$$

we get

$$(A_h u, u) = \sum_{k=0}^{l} (A_h v_k, v_k).$$

Now we apply (4.83) and obtain

$$(A_h u, u) = \sum_{k=0}^{l} (A_h v_k, v_k) \geq c \sum_{k=0}^{l} h_k^{-2} \|v_k\|^2.$$  

\[10\] This condition can be avoided, if $R_0 = A^{-1}$; i.e., a problem on the coarsest grid is solved exactly.
The last estimate and (4.72) imply

\[(A_h u, u) \geq c \sum_{k=0}^{l} h_k^{-2} ||v_k||^2 \geq c \sum_{k=0}^{l} (R_{h_k}^{-1} v_k, v_k).\]

The estimate (4.17) for \(K_0\), independent of \(h_l\) and \(l\), is proved.

Therefore, if \(A\) is a matrix resulting from a finite element discretization of the Poisson problem and \(B_A\) is the additive multigrid preconditioner from (4.30), then \(\text{cond}(BA) \leq C\) with some constant \(C\) independent of the mesh size and the number of grid levels. The latter ensures the \(h\)-independent convergence of the conjugate gradient method for solving the system \(Az = b\) with preconditioner \(B_A\).

4.4.2 Hierarchical basis preconditioner

Similar to the BPX preconditioner, the analysis of the CG method with the hierarchical basis preconditioner can be done in the framework of the theory of space decomposition methods in this chapter. We find estimates for the constants \(K_0\) and \(K_1\) from Theorem 4.3.

Here we consider the Poisson problem again, and the assumptions on finite element discretization are the same as in Section 4.4. Let \(B_h\) be the preconditioner defined by the hierarchical basis method (the formula (4.45)); then

\[\text{cond}(B_h A_h) \leq K_0 K_1, \quad \text{where } K_0 = O(|\ln h_l|^2), K_1 = O(1).\]  

(4.84)

Compared to the BPX preconditioner, the estimate (4.84) depends on \(h_l\). However, such logarithmic dependence is often deemed to be an acceptable price to pay for the ease and efficacy of computations.

The proof of (4.84) is carried out following the lines of the analysis for the BPX preconditioner in Section 4.4. We write \(B_h\) formally in the form

\[B_h = \sum_{i=1}^{l} R_{h_i}^{(i)} Q_i,\]  

(4.85)

where \(Q_i\) is the \(L^2\)-projection operator on the space \(V_i = \mathbb{U}_i - \mathbb{U}_{i-1}\), \(\mathbb{V}_0 = \mathbb{U}_0\). Now the space \(\mathbb{V}\) can be decomposed as the direct sum

\[\mathbb{V} = \mathbb{V}_0 \oplus \cdots \oplus \mathbb{V}_l.\]  

(4.86)

The constant \(K_1\) is the product of \(\alpha\) from (4.18) and \(\rho(G)\) from (4.14). Consider the strengthened Cauchy–Schwarz inequality (4.11). Earlier, when we analyzed the BPX preconditioner, this estimate was proved for \(\mathbb{V}_i = \mathbb{U}_i\). Note that \(\mathbb{U}_i - \mathbb{U}_{i-1} \subset \mathbb{U}_i\). Therefore, for the hierarchical basis method we have \(\mathbb{V}_j \subset \mathbb{U}_j\), and the estimate (4.11) holds with the same constants \(\gamma_{ij}\) as for the BPX preconditioner. Furthermore, the estimate (4.11) yields (4.14).

Now we find an estimate for the constant \(K_0\) from (4.17). The arguments from subsection 4.4.1 cannot be repeated in a straightforward way, since functions \(\tilde{u}_i\) from the decomposition (4.79) are not necessarily in \(\mathbb{U}_i - \mathbb{U}_{i-1}\). Similar to subsection 4.4.1, we consider the unique decomposition of arbitrary \(v \in \mathbb{V}\) induced by (4.86):

\[v = \sum_{k=0}^{l} v_k, \quad v_k \in \mathbb{V}_k.\]  

(4.87)

Define the projector \(\Pi_k : \mathbb{U}_m \to \mathbb{U}_k\), \(m \geq k\), by imposing the equality \(\Pi_k u = u\) in all grid nodes on the grid level \(k\). \(\Pi_k u \in \mathbb{U}_k\) is often called the Lagrange (nodal) interpolant.
for \( u \in \mathbb{U} \). We need the following property of the nodal interpolant in two dimensions: for any \( \phi \in \mathbb{U}_m \) it holds that
\[
\| \phi - \Pi_k \phi \| \leq c \, b_k \| \phi \|_{1\ell} . \tag{4.88}
\]
For vectors from the decomposition (4.87) we have \( v_k = \Pi_k u - \Pi_{k-1} u \). Note that \( \Pi_{k-1} u_k = 0 \). Applying the property (4.88) with \( \phi = \Pi_k u \), we get
\[
\| v_k \| = \| \Pi_k u - \Pi_{k-1} \Pi_k u \| \\
\leq c_1 \, b_{k-1} \| \nabla \Pi_k u \| \leq c_1 \, b_k \| \nabla \Pi_k u \|. \tag{4.89}
\]
We need the following estimate for finite element functions (it holds in the two-dimensional case):
\[
\| u \|_{L^\infty} \leq C(\Omega) \| u \| \| u \|_{1\ell} \quad \forall u \in \mathbb{U}_h. \tag{4.90}
\]
We want to estimate \( \| v_k \| \) by \( \| u \|_1 \). Due to (4.89) it is sufficient to estimate \( \| \Pi_k u \| \) by \( \| u \|_1 \). We decompose the norm into local terms:
\[
\| \Pi_k u \|_1^2 = \sum_{\tau \in \mathcal{T}_h} \| \nabla (\Pi_k u) \|_1^2 . \tag{4.91}
\]
Since \( \Pi_k u \) is linear on every triangle of the triangulation, \( \tau \in \mathcal{T}_h \), it is straightforward to check that
\[
\| \nabla (\Pi_k u) \| \leq \max_{\tau} \| \Pi_k u \|. \tag{4.92}
\]
Since \( \Pi_k u = u \) in the vertices of \( \tau \), then
\[
\max_{\tau} \| \Pi_k u \| \leq \max_{\tau} \| u \|. \tag{4.93}
\]
We sum up the estimates (4.91)-(4.93) and, using (4.90) on every triangle \( \tau \) and appropriate scaling, we get
\[
\| \nabla \Pi_k u \|_1^2 \leq c \left( \ln \frac{b_1}{b_k} \right) \| u \|_1^2 = c \ln \left( \frac{b_k}{b_1} \right) \| \nabla u \|_1^2 .
\]
Thanks to (4.89), it follows that
\[
\sum_{k=0}^l b^{-2}_k \| u_k \|_1^2 \leq c \left( \ln b_1 \right)^2 \| \nabla u \|_1^2 = c \left( \ln b_1 \right)^2 (A_b, v, v).
\]
For the Poisson problem and the uniform grid we conclude that
\[
\sum_{k=0}^l (K^{(k)}_b)^{-1} v_k, v_k \leq c \left( \ln b_1 \right)^2 (A, u, u).
\]
The estimate (4.17) with \( K_0 = c \left( \ln b_1 \right)^2 \) is now proved.

Summarizing, the preconditioner based on the hierarchical basis approach gives for the two-dimensional Poisson problem the following estimate of the condition number:
\[
\text{cond}(B_h A_b) \leq c \left( \ln b_1 \right)^2.
\]
4.4. Convergence analysis for the Poisson problem

In the three-dimensional case, the dependence of the condition number on \( h \) is more pessimistic: \( \text{cond}(B_h A_h) = O(h_l^{-1}) \).

4.4.3 Multiplicative and additive Schwarz preconditioners

The multiplicative and additive Schwarz preconditioners are the space decomposition preconditioners defined by the decomposition (4.59) in subsection 4.3.3. Therefore, the analysis is based on Theorems 4.3 and 4.4.

We start with the additive version of the method and check the assumptions of Theorem 4.3. Let \( B_{\text{add}} \) be the additive Schwarz preconditioner. First, note that due to the trivial estimate (4.16) and equivalence (4.60), the bound \( \lambda_{\text{max}}(B_{\text{add}} A) \leq l C_d \) holds. This estimate can be improved by noting that for the Schwarz preconditioners, we have \( \gamma_{ij} \neq 0 \) in (4.11) iff \( \Omega_i \) and \( \Omega_j \) have nonzero overlap. Denote by \( l_0 \) the maximum number of subdomains any \( \Omega_k \) may overlap (including itself):

\[
l_0 = \max_{k=1, \ldots, l} \sum_{i=1}^{l} \text{sign}(|\text{area}(\Omega_k \cap \Omega_i)|).
\]

We get

\[
\lambda_{\text{max}}(BA) \leq l_0 C_d.
\]

If the number of subdomains \( l \) is large, then this estimate can be a significant improvement over \( \lambda_{\text{max}}(BA) \leq l C_d \).

It remains to estimate the constant \( K_0 \) from (4.17). Fixing an arbitrary \( v \in V \) and letting \( v_h = P_h v \) be the corresponding finite element function, we get \( \langle Av, v \rangle = \langle \nabla v_h, \nabla v_h \rangle \). Consider the decomposition \( v_h = \sum_{k=1}^{l} v_k \), \( v_k \in U_k \) given by (4.58). Due to (4.60), we have \( \langle R^{-1}_k \chi_k v_k, v_k \rangle \leq c_d^{-1} \langle A_k \chi_k^{-1} v_k, v_k \rangle = \langle \nabla v_k, \nabla v_k \rangle \). Therefore, to satisfy (4.17), we are looking for the estimate

\[
\sum_{k=1}^{l} ||\nabla v_k||^2 \leq C_0 ||\nabla v_h||^2. \tag{4.94}
\]

Thanks to the \( H^1 \)-stability of the nodal interpolation, i.e., \( ||\nabla I_h v|| \leq C ||\nabla v|| \) for all \( v \in U_h \), and (4.56), we have

\[
||\nabla v_k|| = ||\nabla I_h (\chi_k v_k)|| \leq C ||\nabla (\chi_k v_k)|| \leq C (||\chi_k ||d^{-1}v_h|| + ||\chi_k ||v_h||)
\leq C (||\nabla v_h|| + C_d ||d^{-1}v_h||), \quad \text{with} \quad d = \text{dist}(x, \partial \Omega).
\]

We use the Harding inequality to estimate the last term, and we get

\[
||\nabla v_k|| \leq C (1 + C_d) ||\nabla v_h||,
\]

with some constant \( C \) independent of a mesh size parameter and the number of subdomains. Squaring and summing up over all \( k \) proves (4.94), with \( C_0 = C l (1 + C_d)^2 \). Theorem 4.3 gives us the estimate for the minimal eigenvalue of \( BA \) for the additive Schwarz preconditioner:

\[
\lambda_{\text{min}}(B_{\text{add}} A) \geq C c_d l^{-1}(1 + C_d)^{-2}.
\]

Here a constant \( C > 0 \) is independent of a mesh size parameter and the number of subdomains; constants \( C_d \) and \( c_d \) are defined in (4.56) and (4.60), respectively.
4.4.4 Nonoverlapping domain preconditioners

A spectral equivalence of the nonoverlapping domain preconditioners is elementary to show, assuming that \( l = 2 \) (only two subdomains) and that the subproblems in each subdomain are solved exactly; i.e., we consider the exact action of \( A^{-1} \) and \( A_{22}^{-1} \) in the definition of \( B_2 \) from subsection 4.3.5. Instead of \( S^{-1} \) consider the Neumann–Neumann preconditioner; cf. subsection 4.3.6. We start by proving the equivalence

\[
\alpha_k S \leq S_k \leq S, \quad k = 1, 2,
\]

with some positive constants \( \alpha_k \) independent of the mesh size \( h \).

To prove (4.95), we need the following well-known extension results for finite element functions [8, 144, 215]: For any \( v_b \in U_b \), there exists \( \tilde{v}_b \in U_b \) such that

\[
\tilde{v}_b = v_b \quad \text{on} \quad \Omega_b \quad \text{and} \quad \|\nabla \tilde{v}_b\|_{L^2(\Omega_b)} \leq C_E(\Omega_b)\|\nabla v_b\|_{L^2(\Omega_b)}, \tag{4.96}
\]

with constants \( C_E(\Omega_b) \) independent of \( v_b \) and \( b \).

Let \( \psi_k \) be a vector of coefficients for interface mesh nodes, and \( \psi_b \) be a finite element function with values \( \psi_k \) at the interface nodes and is discretely harmonic in \( \Omega_1 \) and \( \Omega_2 \), i.e., \( \psi_b \in U_\Gamma \). Then, using the definition of submatrices in (4.71), one readily checks the identity

\[
\langle S_k \psi_1, \psi_1 \rangle = \|\nabla \psi_b\|_{L^2(\Omega_b)}^2, \quad k = 1, 2, \tag{4.97}
\]

Since \( S = S_1 + S_2 \), the equality (4.97) immediately implies the upper bound in (4.95).

The lower bound in (4.95) follows from the observation that discrete harmonic functions minimize the Dirichlet integral over all finite element functions with the same trace on \( \Gamma \). Utilizing this observation and the extension result in (4.96), one shows that

\[
\langle S \psi_1, \psi_1 \rangle = \|\nabla \psi_b\|_{L^2(\Omega_b)}^2 \leq \|\nabla \psi_b\|_{L^2(\Omega_b)}^2 + \|\nabla \tilde{v}_b\|_{L^2(\Omega_b)}^2 \leq (1 + C_E^2(\Omega_b))\|\nabla \psi_b\|_{L^2(\Omega_b)}^2 = (1 + C_E(\Omega_b))2\langle S_k \psi_1, \psi_1 \rangle
\]

for any \( \psi_1 \).

We proved (4.95), and it remains to apply Theorem 4.3. This yields the spectral bounds (4.19) for the Neumann–Neumann preconditioner with \( K_0 = 1 \), \( \alpha = c_1^{-1} + c_2^{-1} \), and \( \rho(G) = 1 \) (the space decomposition is \( A \)-orthogonal). 

Certain technical difficulties arise in analyzing the multisubdomain case \( l > 2 \). In this case, the spectral bound for the Neumann–Neumann preconditioner appears to be
mildly mesh dependent. Assuming a quasi-uniform triangulation and that all subdomains $\Omega_k$ are of regular shape and diameter $d_k \approx d$, one can show [138] the condition estimate for the Neumann–Neumann preconditioner of the Schur complement matrix:

$$\text{cond}(B_S S) \leq C \left(1 + \log \frac{d}{h}\right) \max \left\{ \frac{1}{d^2}; \log \frac{d}{h} \right\}.$$ 

---

**Exercises**

4.4.1. Adapt the analysis of the BPX preconditioner of the preconditioner based on local grid refinement from subsection 4.2.3.

4.4.2. Show (4.74) for higher order conforming finite elements.

4.4.3. How does the estimate for multiplicative and additive Schwarz preconditioners depend on subdomains overlapping?
Chapter 5

Some Applications

In this final chapter, we discuss iterative methods for linear algebraic problems arising from discretizations of several mathematical models described in terms of partial differential equations. These models are chosen to have different specific properties, which make a straightforward application of Krylov subspace methods and preconditioners, built on simple linear iterations such as a multigrid cycle, not always efficient (this is in contrast to the example of the Poisson problem from the previous chapter). To design suitable iterative solvers for many of these problems, special ingredients are needed, such as block preconditioners or multigrid methods with problem-dependent smoothing, restriction, and prolongation operators. We note that for some of the problems discussed further, developing a thorough convergence analysis of iterative methods is an open problem, although in practice iterations can be very efficient.

5.1 Multigrid preconditioners for singular-perturbed problems

5.1.1 Singular-perturbed partial differential equations

In this section, we consider an important class of elliptic partial differential equations: so-called “singular-perturbed” problems. Assume that a differential operator \( \mathcal{L}(\epsilon) \) depends on the parameter \( \epsilon \in (0, 1] \). We call the differential problem

\[
\mathcal{L}(\epsilon) u = f \quad \text{in} \quad \Omega, \quad \ell(u) = g \quad \text{on} \quad \partial\Omega
\]

(5.1)

singular-perturbed if the operator \( \mathcal{L}(0) \) has a type other than \( \mathcal{L}(\epsilon) \). A limiting problem for \( \epsilon = 0 \) may be ill-posed. Singular-perturbed problems are of interest since they are abundant in applications and often reluctant to building efficient preconditioners. Below we consider a few examples of singular-perturbed elliptic second order equations.

Example 5.1. Consider the reaction-diffusion equation

\[
-\epsilon \Delta u + u = f \quad \text{in} \quad \Omega, \quad u|_{\partial\Omega} = 0.
\]

(5.2)

In this example, \( \mathcal{L}(0) \) is a zero order operator.

The problem (5.2) arises if one applies implicit time stepping schemes for parabolic equations. In this case \( \epsilon \sim \nu \tau \), where \( \nu \) is the diffusion coefficient, \( \tau \) is the time step, and the right-hand side \( f \) stands for source terms and all other terms that are treated explicitly.
Example 5.2. Consider the anisotropic diffusion equation

\[-\varepsilon \frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f \text{ in } \Omega , \quad u|_{\partial \Omega} = 0. \tag{5.3}\]

In this example, \(\mathcal{L}(0)\) is not an elliptic operator.

It is important to have a good preconditioner for an algebraic counterpart of (5.3) if one is interested in numerically solving elliptic equations on anisotropic grids. Let us clarify what we mean. Anisotropic grids are often needed to approximate solutions in boundary layers, (small) parts of a domain next to its boundary, where a partial differential equation solution is changing rapidly in the direction orthogonal to the boundary. The elements of anisotropic grids in such case are elongated in the direction parallel to the boundary. The Poisson equation discretized on anisotropic grids occurs often as an auxiliary problem and has to be solved repeatedly during numerical simulations. Assume a grid is stretched in the \(x\) direction. The discrete Poisson problem on such a stretched grid is equivalent to the discretization of (5.3) on regular equidistant grids. In this case \(\varepsilon \approx \left( \frac{h_y}{h_x} \right)^2\), where \(h_x\) (or \(h_y\)) is an element length in the \(x\) (or \(y\)) direction. Thus equation (5.3) discretized on a regular grid gives us a prototypical algebraic system. \(\blacksquare\)

Example 5.3. Consider the convection-diffusion equation

\[-\varepsilon \Delta u + a_1 \frac{\partial u}{\partial x} + a_2 \frac{\partial u}{\partial y} = f \text{ in } \Omega , \quad u|_{\partial \Omega} = 0. \tag{5.4}\]

In this example, \(\mathcal{L}(0)\) is the first order operator.

Equation (5.4) arises in the simulation of fluid or gas dynamics. It models convection phenomena induced by the temperature variation of a fluid occupying a volume \(\Omega\). In this case \((a_1, a_2)\) is the vector of fluid velocity, \(u\) is the temperature, and \(\varepsilon\) is the coefficient of thermal conductivity. Moreover, (5.4) is the linearization of the elliptic part of the momentum equations of fluid motion and often serves as an auxiliary problem. In this case, \(u\) is a vector function, (5.4) is the system of two equations, and \(\varepsilon\) is the kinematic viscosity (in the three-dimensional case the term \(a_3 \frac{\partial u}{\partial z}\) is added to the left-hand side of (5.4)). \(\blacksquare\)

Example 5.4. Consider the equations of linear elasticity,

\[-\varepsilon \Delta u - \nabla \text{div} u = f \text{ in } \Omega , \quad u|_{\partial \Omega} = 0. \tag{5.5}\]

\(u : \Omega \to \mathbb{R}^d\) and \(f : \Omega \to \mathbb{R}^d\), \(\Omega \subset \mathbb{R}^d\), \(d = 1,2\), are vector functions. In this example, \(\mathcal{L}(0)\) is not an elliptic operator.

Here and in what follows we write in bold (e.g., \(u, f, U, H^1\)) vector functions and spaces of vector functions.

Equations (5.5) are a so-called displacement formulation of linear elasticity equations. These equations are often used to model the behavior of rigid bodies under small deformations such that one may assume the Hooke’s law applies. In fluid mechanics applications, equations (5.5) are considered a part of fluid-structure interaction problems in aerodynamics (flows around elastic structures, such as wings) and hemodynamics (blood flow in elastic vessels). Parameter \(\varepsilon\) depends on material properties described in terms of Young’s modulus and the Poisson ratio. When the Poisson ratio approaches \(\frac{1}{2}\) for
nearly incompressible materials, the (rescaled) displacement formulation takes the form (5.5) with very small \( \varepsilon > 0 \).

We consider the homogeneous Dirichlet boundary conditions in (5.2)–(5.5) as an example. Other boundary conditions (such as the Neumann or the Robin condition) are also suitable for analysis. We shall make necessary comments on the well-posedness of these problems later.

5.1.2 What is a robust method?

We call an iterative method for solving a discretized singular-perturbed problem robust if its convergence factor is bounded by a constant less than one, which is independent of a mesh size, the number of grid levels, and the value of \( \varepsilon \) from \((0,1]\). If a multigrid method is invoked, then the number of arithmetic operations on every grid level should also stay bounded independent of \( \varepsilon \) and be proportional to the number of unknowns on this level. The same definition extends to preconditioners: convergence of a Krylov subspace solver for a system of equations with a robust preconditioner should be “fast” and insensitive to variations of the parameter \( \varepsilon \) (for any right-hand side vector \( b \)). The robustness property of an iterative method for singular-perturbed problems is important in many applications.

In all of the examples in subsection 5.1.1, depending on physical setup the value of \( \varepsilon \) may vary from being of order one to a much smaller value, say, of order \( 10^{-3} - 10^{-6} \) or even smaller. Thus, we expect that an efficient iterative method, such as the multigrid method, or a preconditioner for problems (5.1)–(5.5) satisfies the robustness property.

Speaking formally, a sufficient condition for a linear iterative method to be robust is that the iteration matrix \( M \) of the multigrid cycle satisfy the estimate

\[
\|M\| \leq 1 - c^2, \tag{5.6}
\]

with some \( c > 0 \) independent of \( \varepsilon \) and discrete parameters. In particular, (5.6) implies that the corresponding preconditioner is also robust; see subsection 1.4.13 and the discussion after Theorem 1.33. In this section, we consider multigrid iterations and preconditioners for singular-perturbed problems. Thus, we proceed by discussing some desired properties of such multigrid solvers.

5.1.3 Conditions on smoothing iterations

Assume the problem (5.1) is approximated on two grid levels, which correspond to finite element subspaces \( \mathbb{U}_h \) and \( \mathbb{U}_H \). Let \( u_h \) and \( u_H \) be finite element solutions to the problem (5.1); then, using canonical restriction and prolongation \( r \) and \( p \), the approximation property (3.69) can be written equivalently in the form of the estimate on the difference \( u_h - u_H \) (see subsection 3.4.1):

\[
\|u_h - u_H\|_{L^2} \leq C_A \|A\|^{-1} \|f\|_{L^2}, \quad \forall f \in L^2(\Omega), \tag{5.7}
\]

where \( A \) is the matrix of the system of algebraic equations from the finite element method.

As we already know, a bound on \( \|u_h - u_H\| \) follows from the finite element error estimate and a regularity estimate for the solution to a differential problem:

\[
\|u_h - u\|_{L^2} \leq C_1(\varepsilon) h^2 \|u\|_{H^2},
\]

\[
\|u\|_{H^2} \leq C_2(\varepsilon) \|f\|_{L^2}.
\]

For the example of the anisotropic diffusion equation, it can be shown that \( \|A\|^{-1} = c h^2 \) and \( C_A \sim C_1(\varepsilon) C_2(\varepsilon) \), where both constants \( C_1 \) and \( C_2 \) depend on \( \varepsilon \). Moreover, \( \zeta(\varepsilon) := C_A \to \infty \) for \( \varepsilon \to 0 \).
Let $\eta(\nu, \epsilon)$ be a function from the smoothing property (3.68), where $\nu$ is the number of smoothing iterations. The proof of Theorem 3.7 shows that the norm of the two-grid iteration matrix is estimated by the product of two right-hand side terms from the approximation and smoothing properties. To prove the robust convergence, one has to show that this product is less than one and independent of $\epsilon$ and $h$. Therefore, the definition of the robustness implies the following condition for $\eta(\nu, \epsilon)$. For any given $q \in (0, 1)$, there exists $\bar{\nu} > 0$, independent of $\epsilon$, such that

$$\eta(\nu, \epsilon) \zeta(\epsilon) \leq q < 1 \quad \forall \nu > \bar{\nu}. \quad (5.8)$$

Thus, from the analysis it follows that smoothing iterations should be built to compensate for the degradation of the approximation property as $\epsilon \to 0$. In particular, from (5.8) it follows that $\eta(\nu, \epsilon) \to 0$ for $\epsilon \to 0$; i.e., smoothing iterations should solve the limit problem (almost) exactly. This notion of “robust smoother” was introduced in [96]. It appears, however, that not every singular-perturbed problem and discretization admits a simple way of building smoothing iterations, which are exact solvers for $\epsilon = 0$. This requirement for smoothing iterations can be weakened if the finite element solution satisfies the estimate $\|u_h\|_{L^2} \leq c \|f\|_{L^2}$. In this case, it is sufficient to ensure that for $\epsilon \to 0$ smoothing iterations become an iterative solver with a convergence factor independent of $h$ and $\epsilon$.

From the theoretical point of view, one can try to weaken the dependence of $C_A$ on $\epsilon$ by checking the approximation property in some other norm (different from the spectral norm). However, this would also require the proof of a smoothing property in a new norm. This is not always easy. One can also try to improve $C_A$ by choosing $p$ and $r$ not in the canonical way but in a way that $p$ and $r$ depend on the matrix $A$. Using this idea, robust algorithms were obtained for problem (5.4). They demonstrate satisfactory numerical performance; however, it appears to be hard to analyze such methods.

More flexibility in developing robust multigrid methods gives a potential improvement of smoothing iterations. One can try more sophisticated and problem-oriented iterative methods as smoothers to improve the overall multigrid performance. Common tools here are the incomplete LU (ILU) decomposition and block preconditioners. A special problem-dependent enumeration of grid points (resp., basis functions) can be quite important as well.

### 5.1.4 Bibliographical remarks

Iterative methods for problems (5.2)–(5.5) have been extensively studied in the literature. It is well known that standard BPX and hierarchical basis preconditioners are not robust for the linear reaction-diffusion equation (5.2). Adaptation of these preconditioners to obtain full robustness has been done in [92]. At the same time, geometric multigrid $V$- and $W$-cycles are robust [156]. A robust preconditioner for (5.2) based on hierarchical space decomposition was introduced in [187] and based on domain decomposition techniques in [25, 116].

For the anisotropic diffusion equation (5.3), robust variants of both multiplicative and additive multigrid methods have been developed; see [11, 40, 170, 186, 187, 216, 226]. To achieve robustness, they use ILU, block Gauss-Seidel, or block Jacobi iterations as smoothers. Several studies also suggest semicoarsening strategies for grid refinement, where a grid is allowed to remain fine in one Cartesian direction; see [6, 220]. Robust preconditioners for (5.3) based on domain decomposition techniques were developed in [22, 113, 120, 121]. These domain decomposition preconditioners are built similarly to those in the isotropic case. However, the analysis requires more advanced tools such as anisotropic trace theorems for Sobolev spaces.
5.1. Multigrid preconditioners for singular-perturbed problems

Unlike the previous two equations, the convection-diffusion problem (5.4) results in linear algebraic systems with possibly strongly nonsymmetric matrices. Moreover, it appears that the ability to build an efficient preconditioner for this problem depends on properties of the vector function $a$. Thus, it is not a surprise that analysis of robust preconditioners for (5.4) is far less developed. This said, we note that in general, numerical experiments show good performance of multigrid methods if the functions $a_1(x,y)$ and $a_2(x,y)$ are sufficiently smooth and the finite element or finite difference discretizations are constructed in a special way. For a smoothing iteration one typically considers an ILU preconditioned linear iteration or the Gauss-Seidel method with a special (depending on $a_1,a_2$) numbering of grid nodes. There is a vast literature discussing the application of multigrid methods for solving this problem; see $[23, 71, 73, 98, 150, 168, 219, 224, 228]$.

So far, the robust convergence is mathematically proved only in a few cases; specifically, for constant $a_1,a_2$ and uniform grid. A multigrid method for a the first order finite difference scheme is analyzed in $[169]$ and for a finite element method in $[157, 152]$. Domain decomposition preconditioners for (5.4) have been intensively studied in the literature; see, for example, $[50, 81, 82, 101, 117, 125, 136, 192]$. The performance of these preconditioners is influenced by many factors such as type of discretization, alignment of a grid with flow streamlines, and adaptation of domain partition depending on the flow $a$.

For small values of parameter $\epsilon$, the dominating term in equations (5.5) has a large non-trivial kernel consisting of all solenoidal functions satisfying boundary conditions. For a linear iterative method it is hard to eliminate those components of the error which belong to this kernel (more precisely, to its algebraic counterpart). This makes building robust preconditioners for (5.5) a notoriously difficult task. To obviate this difficulty, it is common to rewrite the elliptic problem (5.5) as a saddle point problem by introducing new variable $p = -\epsilon^{-1} \text{div} u$; see, e.g., $[48]$. After a rescaling, the new system of partial differential equations reads

$$\begin{align*}
-\Delta u + \nabla p &= \epsilon^{-1} f \text{ in } \Omega, \\
-\text{div} u - \epsilon p &= 0 \text{ in } \Omega.
\end{align*}$$

The matrix of the resulting algebraic system is indefinite, but for an appropriate discretization it remains well-conditioned for all $\epsilon \in [0, 1]$ and has a convenient block structure. We shall discuss iterative methods for such indefinite block matrices in Section 5.2. Nevertheless, for some discretizations it is feasible to handle vectors from the kernel of a discrete $\nabla \text{div}$ operator in smoothing and prolongation steps, and so one may try to build a robust iterative method for solving (5.5) directly as the positive definite problem; see $[176, 129]$. Other recent attempts to build efficient preconditioners for (5.5) without resorting to saddle point formulation (5.9) include the domain decomposition method $[68]$, preconditioners based on the inverse Sherman-Morrison algorithm $[100]$, and hierarchical-LU factorization $[27]$.

In the remainder of Section 5.1, we consider several robust multigrid methods for linear algebraic systems resulting from (5.2)-(5.5) and their analysis.

5.1.5 The linear reaction-diffusion problem

In this subsection, we show the robust convergence of V- and W-cycles of the multiplicative multigrid method for the reaction-diffusion equation

$$
-\epsilon \Delta u + u = f \text{ in } \Omega, \quad u|_{\partial \Omega} = 0.
$$

The analysis follows paper $[156]$. 

We assume $\Omega$ is a bounded polyhedral domain in $\mathbb{R}^n$, $n = 2, 3$. Similarly to the examples from the previous chapters, the weak formulation of the problem is obtained by multiplying (5.10) by an arbitrary function $v$ from $H_0^1(\Omega)$ and integrating the resulting equality by parts:

$$
\varepsilon \int_{\Omega} \sum_{i=1}^{n} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i} \, dx + \int_{\Omega} u \, v \, dx = \int_{\Omega} f \, v \, dx.
$$

It is convenient to write the weak formulation in a slightly more abstract form: Find $u$ from $H^1_0(\Omega)$, satisfying the identity

$$
a(u, v) = (f, v) \quad \forall \ v \in H^1_0(\Omega), \tag{5.11}
$$

where the bilinear form $a(\cdot, \cdot)$ is given by

$$
a(u, v) := \varepsilon \int_{\Omega} \sum_{i=1}^{n} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i} \, dx + \int_{\Omega} u \, v \, dx.
$$

In the finite element method, the Hilbert space $H^1_0(\Omega)$ is replaced by a finite element space $U_h \subset H^1_0(\Omega)$. The discrete problem is then to find $u_h \in U_h$ satisfying the equality

$$
a(u_h, v_h) = (f, v_h) \quad \forall \ v_h \in U_h. \tag{5.12}
$$

The bilinear form $a(\cdot, \cdot)$ is symmetric. The finite element stiffness matrix (5.12) corresponding to the symmetric form $a(\cdot, \cdot)$ is also symmetric.

We recall the important orthogonality property of the finite element error. The equality (5.11) holds, in particular, for $v = v_h$, and thus it follows from (5.11) and (5.12) that the finite element error function is $a(\cdot, \cdot)$-orthogonal to every finite element function:

$$
a(u - u_h, v_h) = 0 \quad \forall \ v_h \in U_h. \tag{5.13}
$$

The orthogonality property (5.13) will be frequently used below.

A system of algebraic equations results from (5.12) if one fixes a (finite) basis in $U_h$. The approximation property of a multigrid method relies on the estimate of the finite element error given in Lemma 5.5 below. In this section and further in this chapter, $c, c_1, \ldots$ or $C, C_1, \ldots$ denote some generic constants independent of $\varepsilon$, mesh size, and grid level. We continue to use the notation $\|v\|$ to denote the $L^2$-norm of $v$.

**Lemma 5.5.** Assume the domain $\Omega$ and finite element space satisfy conditions from Section 3.4. Let $u$ and $u_h$ be the continuous and finite element solutions to the problem (5.2); then

$$
\|u - u_h\| \leq c \left( \frac{\varepsilon}{h^2} + 1 \right)^{-1} \|f\|. \tag{5.14}
$$

**Proof.** Consider the error function $e_h = u - u_h$. Thanks to the orthogonality property (5.13) and the symmetry of the bilinear form $a(\cdot, \cdot)$, we get

$$
\|e_h\|^2_{L^2} \leq a(e_h, e_h) = a(u, e_h) = (f, e_h) = \|f\| \|e_h\|.
$$

Therefore,

$$
\|e_h\| \leq \|f\|. \tag{5.15}
$$
At the same time, for arbitrary \( \psi_h \in U_h \) it holds that
\[
\varepsilon \| e_h \|^2_{H^1} + \| e_h \|^2_{L^2} = a(e_h, e_h) = a(u - \psi_h, e_h) = \varepsilon (\nabla (u - \psi_h), \nabla e_h) + (u - \psi_h, e_h) \\
\leq \varepsilon \| u - \psi_h \| \| e_h \|_1 + \| u - \psi_h \| \| e_h \|_1 \\
\leq (\varepsilon \| u - \psi_h \|_1^2 + \| u - \psi_h \|_1^2)^{\frac{1}{2}} (\varepsilon \| e_h \|_1^2 + \| e_h \|_1^2)^{\frac{1}{2}}.
\]
We obtain
\[
\varepsilon \| e_h \|^2_{H^1} + \| e_h \|^2_{L^2} \leq \varepsilon \| u - \psi_h \|^2_1 + \| u - \psi_h \|^2_{L^2}. \tag{5.16}
\]
For \( \psi_h \) we take a function from \( \mathbb{U}_h \) such that the standard approximation estimates \( \| u - \psi_h \| \leq c h^2 \| u \|_{H^2} \) and \( \| u - \psi_h \|_1 \leq c h \| u \|_{H^2} \) hold. One example of such a \( \psi_h \) is the projection of \( u \) in \( \mathbb{U}_h \) with respect to the scalar product \( \langle \nabla, \nabla \cdot \cdot \rangle \) see (3.89). Using the a priori estimate \( \| u \|_1 \leq \frac{c}{\varepsilon} \| f \| \) (a simple proof of this estimate is given, e.g., in [156]) we obtain
\[
\varepsilon \| e_h \|^2_{H^1} + \| e_h \|^2_{L^2} \leq c \frac{b^2}{\varepsilon} \left(1 + \frac{b^2}{\varepsilon}\right) \| f \|^2_{L^2}.
\]
Thus, it holds that
\[
\varepsilon \| e_h \|_1 \leq c \frac{b^2}{\varepsilon} \left(1 + \frac{b^2}{\varepsilon}\right) \| f \|_1. \tag{5.17}
\]
Now we use an argument based on the dual problem: Let \( w \in H^1_0(\Omega) \) be such that
\[
a(w, v) = (e_h, v) \quad \forall \ v \in H^1_0(\Omega).
\]
Since \( e_h \in L^2(\Omega) \), the \( H^2 \)-regularity assumption (see (3.86) and subsection 3.4.4) ensures that the function \( w \) belongs to the space \( H^2(\Omega) \) and satisfies the estimate \( \| w \|_{H^2} \leq \frac{c}{\varepsilon} \| e_h \|_1 \). We set \( w_h \) to be the projection of \( w \) in \( \mathbb{U}_h \). Then
\[
\| e_h \|^2_{L^2} = a(w, e_h) = a(w - w_h, e_h) \leq \varepsilon \| w - w_h \|_1 \| e_h \|_1 \\
+ \| w - w_h \| \| e_h \| \leq c \varepsilon b \| w \|_{H^2} \| e_h \|_1 + \varepsilon h^2 \| w \|_{H^2} \| e_h \| \\
\leq c \left(b \| e_h \|_1 + \frac{b^2}{\varepsilon} \| e_h \| \right) \| e_h \|.
\]
Using (5.15) and (5.17), we get for \( \frac{b^2}{\varepsilon} \leq 1, \)
\[
\| e_h \| \leq c \left(b \| e_h \|_1 + \frac{b^2}{\varepsilon} \| f \| \right) \\
\leq c b \frac{b^2}{\varepsilon} \left(1 + \frac{b^2}{\varepsilon}\right) \| f \|_1 + c \frac{b^2}{\varepsilon} \| f \| \leq c \frac{b^2}{\varepsilon} \| f \|. \tag{5.18}
\]
The inequality (5.15), together with (5.18) and \( \min \{1, \frac{b^2}{\varepsilon}\} \leq \left(\frac{c}{\varepsilon} + 1\right)^{-1} \), yields (5.14). □

The finite element formulation (5.12) leads to the system of algebraic equations
\[
Ax = b
\]
with a stiffness matrix \( A \) (see subsection 3.1.4).
The arguments similar to those for the Poisson problem in Section 3.4 provide us with the bounds for the spectral norm of the stiffness matrix:

\[ c_1 \left( \frac{\varepsilon}{h^2} + 1 \right) \leq ||A||_2 \leq c_2 \left( \frac{\varepsilon}{b^2} + 1 \right). \]  

(5.19)

Thanks to (5.19) and estimate (5.14) of Lemma 5.5, we have the approximation property (5.7) with a constant \( C_A \) independent of \( \varepsilon \) and \( h \). We will see that the simplest smoother, such as the relaxed Jacobi method, leads to robust multiplicative multigrid methods.

From (5.19) it follows that

\[ ||D^{-1}||_2 = \left( \min_i (A_{ii}) \right)^{-1} \leq c_1^{-1} \left( \frac{\varepsilon}{h^2} + 1 \right)^{-1} \leq \frac{C_2}{C_1} ||A||^{-1}_2. \]

Hence, for the diagonal part of \( A \) we have the estimate

\[ ||D^{-1}||_2 \leq \frac{c_1}{||A||_2}. \]

For the relaxed Jacobi smoother, we take \( W = \omega D \); then for appropriate \( \omega \) it holds that

\[ \text{sp}(W^{-1}A) \subset (0,1]. \]  

(5.20)

Now we shall check the following smoothing property,

\[ ||AS^\nu||_2 \leq c \frac{1}{\nu + 1} ||A||_2, \quad \nu = 1,2,\ldots, \]  

(5.21)

with the iteration matrix \( S = I - W^{-1}A \). Denote \( B := W^{-\frac{1}{2}}AW^{-\frac{1}{2}} \). Note that \( B \) is symmetric and \( \text{sp}(B) \subset (0,1]. \) Moreover,

\[ ||AS^\nu||_2 = ||W^{-\frac{1}{2}}B(I-B)^\nu W^{\frac{1}{2}}||_2 \leq ||W||_2 ||B(I-B)^\nu||_2. \]  

(5.22)

By Lemma 3.11, \( ||B(I-B)^\nu||_2 \leq (\nu + 1)^{-1} \). Since \( W \) is the scaled diagonal part of \( A \), then \( ||W||_2 \leq c ||A||_2 \) with the constant \( c \) independent of \( b \) and \( \varepsilon \). Hence the smoothing property (5.21) holds. The smoothing property for the V-cycle also follows directly from (5.21).

Lemma 5.5, together with (5.19), (5.20), and (5.21), implies that approximation and smoothing properties hold with some constants independent of \( \varepsilon \). This implies robustness of V- and W-cycles for (5.12). All missing arguments are similar to those we applied to the Poisson problem in Section 3.4.

### 5.1.6 ILU decomposition, ILU smoother

In the previous subsection, we showed that to build a robust multigrid preconditioner for the reaction-diffusion problem it is sufficient to use standard relaxed Jacobi smoothing iterations, canonical prolongation, and restriction. The only challenge was checking the approximation and smoothing properties, with particular attention paid to the dependence of the estimates on \( \varepsilon \). However, to approach efficiently other singular-perturbed problems from the list in subsection 5.1.1, we need more appropriate and advanced tools. One of these is the ILU smoothing iterations considered in this section.

An arbitrary nonsingular matrix can be decomposed (not uniquely!) as

\[ A = LU, \]  

(5.23)
5.1. Multigrid preconditioners for singular-perturbed problems

where \( L \) is a lower triangle matrix and \( U \) is an upper triangle matrix. Since

\[
\det(A) = \det(L) \det(U),
\]

matrices \( L \) and \( U \) are invertible. If the decomposition (5.23) is given for a matrix \( A \), then one can find the solution to \( Ax = b \) in two subsequent steps: In the first step, one finds an auxiliary vector \( \tilde{z} \) from the system \( L \tilde{z} = b \), and in the second step, one recovers the solution from \( Uz = \tilde{z} \). Both steps are straightforwardly performed by the Gauss elimination.

Computing factorization (5.23) is closely related to the Gauss elimination method applied to solve \( Ax = b \). To define an algorithm for finding decomposition (5.23), denote \( A^{(1)} = A \). The elements of the \( k \)th matrix \( A^{(k)} \) are denoted by \( a_{ij}^{(k)} \). Consider the first step. We find \( A^{(2)} = L_1 A^{(1)} \), where

\[
A^{(2)} = \begin{bmatrix}
 1 & 0 & 0 & \cdots & 0 \\
-\frac{a_{11}^{(1)}}{s_{11}} & 1 & 0 & \cdots & 0 \\
 0 & 1 & 0 & \cdots & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots \\
-\frac{a_{n1}^{(1)}}{s_{11}} & 0 & \cdots & 0 & 1
\end{bmatrix},
\]

\[
L_1 = \begin{bmatrix}
 1 & 0 & 0 & \cdots & 0 \\
\frac{a_{11}^{(1)}}{s_{11}} & 1 & 0 & \cdots & 0 \\
 0 & 1 & 0 & \cdots & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots \\
\frac{a_{n1}^{(1)}}{s_{11}} & 0 & \cdots & 0 & 1
\end{bmatrix},
\]

\[
a_{ij}^{(2)} = a_{ij}^{(1)} - a_{ij}^{(1)} \frac{a_{11}^{(1)}}{s_{11}} a_{1j}^{(1)}, \quad i, j = 2, \ldots, n.
\]

Repeating the procedure for the right bottom block of the matrix \( A^{(2)} \) and so on, we find (5.23), where

\[
L = L_1^{-1} L_2^{-1} \cdots L_{n-1}^{-1}, \quad U = A^{(n)}.
\]

Elements of \( A^{(n)} \) are computed recursively as

\[
a_{ij}^{(s+1)} = \begin{cases} 
  \frac{a_{ij}^{(s)}}{s_{ij}} & \text{for } i \leq s, \\
  0 & \text{for } i > s, \quad j \leq s, \\
  a_{ij}^{(s)} - a_{ij}^{(s)} \left[ \frac{a_{ij}^{(s)}}{s_{ij}} \right]^{-1} a_{ij}^{(s)} & \text{for } i > s, \quad j > s.
\end{cases}
\]

(5.24)

The matrix \( L \) is found to be

\[
L = \begin{bmatrix}
  1 & 0 & 0 & \cdots & 0 \\
\frac{a_{11}^{(1)}}{s_{11}} & 1 & 0 & \cdots & 0 \\
 0 & \ddots & \ddots & \ddots & \ddots \\
\frac{a_{11}^{(n-1)}}{s_{11}} & \frac{a_{12}^{(n-1)}}{s_{22}} & \cdots & \frac{a_{1n-1}^{(n-1)}}{s_{n-1,n-1}} & 1
\end{bmatrix}.
\]

If \( A \) is a symmetric matrix it is convenient to look for the symmetric factorization:

\[
A = (L + D)D^{-1}(L + D)^T,
\]

(5.25)

where \( D \) is a diagonal matrix with \( d_{ii} = a_{ii}^{(1)} \) and \( L \) is a low triangle matrix with the zero main diagonal.
Suppose, that the matrix $A$ is sparse. By “sparse matrix” we mean here that every row of the matrix has at most $O(1)$ nonzero elements independent on the matrix size. Unfortunately, matrices $L$ and $U$ from (5.24) do not necessarily inherit this property of $A$. In particular, this is true for matrices resulting from finite element or finite difference discretizations of elliptic partial differential equations in a space dimension greater than one. For most practical cases, the complexity of finding $LU$ decomposition is not optimal with respect to the number of matrix nonzero entries. The observation making the decomposition useful for a sparse matrix is that many elements of the matrices $L$ and $U$ can be “close” to zero. Setting these element equal to zero, we find matrices $\tilde{L}$ and $\tilde{U}$ such that the matrix $W = \tilde{L}\tilde{U}$ is “close” to $A$. Moreover, systems of equations with matrices $\tilde{L}$ and $\tilde{U}$ can be solved with the help of Gauss elimination evoking $O(N)$ arithmetic operations. Hence the computation of $W^{-1}v$ for any vector $v \in \mathbb{R}^N$ has optimal complexity. All of these are good reasons to use the matrix $W$ as the preconditioner and to expect that the iterations

$$z_{new}^{\nu} = z_{old}^{\nu} - W^{-1}(Az_{old}^{\nu} - b)$$

(5.26)

rapidly converge—or at least to expect that iteration (5.26) has good smoothing properties.

Let us decompose a matrix $A$ as

$$A = \hat{L}\hat{U} - N = W - N.$$  

(5.27)

The decomposition (5.27) is called the incomplete LU (ILU) decomposition, and in the context of multigrid methods, iterations (5.26) are called ILU smoothing iterations.

A common approach for finding $\hat{L}$ and $\hat{U}$ is defined by the choice of the pattern of nonzero elements $P$, where $P$ is a set of indexes $(i,j)$. A necessary assumption on $P$ is that it includes all diagonal elements, $(i,i) \in P$. Now, in the course of computing $\hat{L}$ and $\hat{U}$, an element $a_{ij}^{(i+1)}$ is updated via (5.24) iff $(i,j) \in P$; otherwise one sets $a_{ij}^{(i+1)} = 0$. If $card(P) = O(n)$, then the arithmetic complexity of computing $W = \hat{L}\hat{U}$, as well as of matrix-vector operation with $W^{-1}$, is $O(n)$. One common choice of $P$ is the set of those indexes for which elements of the original matrix $A$ are nonzero, $(i,j) \in P$ iff $a_{ij}^{(1)} \neq 0$. In the literature, such incomplete factorization is called “ILU with zero fill-in” and denoted by ILU(0).

The following lemma from [216] shows that if a decomposition (5.27) is reasonable (a norm of the defect matrix $N$ is small), then we get an efficient smoother. Moreover, the function $\eta$ from the smoothing property (see (3.71)) scales like $\|N\|_2$.

**Lemma 5.6.** Let $A = A^T > 0$ and $W = W^T > 0$, and consider decomposition (5.27) and the iteration matrix corresponding to this decomposition: $S = I - W^{-1}A = W^{-1}N$. Assume that for some $\alpha > 2$ and $\delta > 0$ it holds that

$$A + \alpha N \geq 0,$$

(5.28)

$$\|N\|_2 \leq \delta,$$

(5.29)

$$\|S\|_2 \leq C;$$

(5.30)

then for $\nu = 2,3,\ldots$, the following estimate is valid:

$$\|AS^\nu\|_2 \leq C_\nu \delta \max \left\{ \frac{1}{\nu-1}, \left(1 + \frac{1}{\alpha-1}\right)\left(\alpha - 1\right)^{2-\nu} \right\}.$$  

(5.31)
5.1.7 Anisotropic diffusion equation

**Proof.** Denote $X := W^{-\frac{i}{2}}NW^{-\frac{i}{2}}$ and consider two chains of equivalent relations:

$$A \geq 0 \iff W - N \geq 0 \iff I - W^{-\frac{i}{2}}NW^{-\frac{i}{2}} \geq 0 \iff I \geq X$$

and

$$A + \alpha N \geq 0 \iff W + (\alpha - 1)N \geq 0 \iff (\alpha - 1)W^{-\frac{i}{2}}NW^{-\frac{i}{2}} \geq -I \iff X \geq -(\alpha - 1)^{-1}I.$$  

Denote $\theta = (\alpha - 1)^{-1}$. From the above inequalities it follows that $sp(X) \in [-\theta, 1]$. Consider the following equalities:

$$AS^\nu = (W - N)(W^{-1}N)^\nu$$

$$= W^\frac{i}{2}(I - W^{-\frac{i}{2}}NW^{-\frac{i}{2}})W^\frac{i}{2}(W^{-\frac{i}{2}}NW^{-\frac{i}{2}})W^\frac{i}{2}$$

$$= W^\frac{i}{2}(I - X^\nu)W^\frac{i}{2}. $$

Therefore, we get

$$||AS^\nu||_2 = ||W^\frac{i}{2}X(I - X^\nu)X^{\nu - 2}XW^\frac{i}{2}||_2$$

$$\leq ||(I - X^\nu)X^{\nu - 2}||_2||W^\frac{i}{2}X^2W^\frac{i}{2}||_2 = ||(I - X)X^{\nu - 2}||_2||NW^{-1}N||_2$$

$$\leq ||(I - X)X^{\nu - 2}||_2||N||_2||S||_2 \leq C_1 \delta \max_{\nu \in [-\theta, 1]} |(1 - \theta)\theta^{\nu - 2}|$$

$$\leq C_1 \delta \max \left\{ \frac{1}{\nu - 1}, (1 + \theta)\theta^{\nu - 2} \right\}. \quad \Box$$

5.1.7 Anisotropic diffusion equation

We start with considering a finite element method for equation (5.3). Arguing similarly to the Poisson and the reaction-diffusion problems, we get the following finite element formulation: Find $u_h \in U_h$ satisfying the equality

$$\int_{\Omega} \frac{\partial u_h}{\partial x} \frac{\partial v_h}{\partial x} + \frac{\partial u_h}{\partial y} \frac{\partial v_h}{\partial y} \, dx = \int_{\Omega} f \, v_h \, dx \quad \forall v_h \in U_h. \quad (5.32)$$

We consider in detail the case when $\Omega$ is the unit square and $U_h$ is the space of finite element piecewise linear functions with respect to the uniform grid. Assume that the mesh size $b$ is equal to $\frac{1}{n}$. Let us enumerate the mesh nodes in the following order:

$$(b, b), (b, H), \ldots, (b, (n - 1)b), (H, b), \ldots, (H, (n - 1)b), \ldots.$$  

The matrix of the corresponding system of equations, i.e., the stiffness matrix of the finite element method (5.32), will be of the block-tridiagonal form

$$A = \frac{1}{b^2} \begin{bmatrix} A_1 & I_x & \cdots & 0 \\ I_x & A_2 & I_x & \cdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & I_x & A_{n-2} \\ 0 & \cdots & 0 & I_x & A_{n-1} \end{bmatrix}, \quad (5.33)$$
where the blocks $A_i$ and $I_\epsilon$ are the matrices of $(n-1) \times (n-1)$ size, and they are given by

$$A_i = \begin{bmatrix}
2(1+\epsilon) & -1 & \cdots & \cdots & -1 \\
-1 & 2(1+\epsilon) & -1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
0 & -1 & 2(1+\epsilon) & -1 & 0 \\
\end{bmatrix},$$

$$I_\epsilon = -\epsilon I.$$

Note that the standard five-point finite difference approximation of equation (5.3) leads to the same system. If a grid is uniform, a useful alternative to the matrix-vector form of writing out the system of algebraic equations is the stencil\footnote{To define the action of a discrete operator on a function defined in grid nodes, besides the matrix notion one often uses the stencil as a more illustrative form. The stencil defines the rule of computing the result of the operator action in a given node using the original function values in this and neighboring nodes. The stencil collects the factors which multiply the values of the function in these nodes. A value in the middle of the stencil corresponds to the given node, where the result of the operation is sought.} of the discrete (finite element) scheme. Here the stencil takes the form

$$\bar{A}(\epsilon) = \frac{1}{h^2} \begin{bmatrix}
0 & -1 & 0 \\
-\epsilon & 2(1+\epsilon) & -\epsilon \\
0 & -1 & 0 \\
\end{bmatrix}.$$

To distinguish between defining the discrete operator by its matrix or by its stencil, in the latter case we shall use square brackets and a bar over the operator notation.

**ILU smoothing iterations**

First, we look for a symmetric version of the ILU(0) decomposition,

$$A \approx (L+D)D^{-1}(D+L^T),$$

where $A_h$ is the stiffness matrix from (5.33), $L$ is the lower triangle part of $A(\epsilon)$ (without the main diagonal), and the elements of the diagonal matrix $D$ are computed according to the recursive relations

$$d(i,j) = \begin{cases}
2(1+\epsilon), & i = j = 1, \\
2(1+\epsilon) - \epsilon^2/d(i,j-1), & i = 1, j > 1, \\
2(1+\epsilon) - 1/d(i-1,j), & i > 1, j = 1, \\
2(1+\epsilon) - \epsilon^2/d(i,j-1) - 1/d(i-1,j), & i > 1, j > 1,
\end{cases} \quad (5.34)$$

where $d(i,j)$ is the diagonal element of $D$ corresponding to the node $(jh, ih)$.

Now we find the remainder $N = A - (L+D)D^{-1}(D+L^T)$. To compute the product of the matrices in the decomposition, it is convenient to use the stencil form of the grid operators. Thus the matrix-vector multiplication with

$$(L+D)D^{-1} = LD^{-1} + I$$

corresponds to applying the stencil

$$\bar{P}_1 = \begin{bmatrix}
0 & -\epsilon \\
\frac{-\epsilon}{d(i-1,j)} & 1 \\
0 & \frac{-1}{d(i-1,j)} \\
\end{bmatrix},$$
and $D + L^T$ corresponds to applying the stencil
\[
\tilde{P}_2 = \frac{1}{h^2} \begin{bmatrix}
0 & -1 & 0 \\
0 & d(i, j) & -\varepsilon \\
0 & 0 & 0
\end{bmatrix}.
\]

We compute the product $\tilde{P}_1 \tilde{P}_2$, decomposing $\tilde{P}_1$ into three terms. We get for $h^2 \tilde{P}_1 \tilde{P}_2$,
\[
\begin{bmatrix}
0 & -1 & 0 \\
0 & d(i, j) & -\varepsilon \\
0 & 0 & 0
\end{bmatrix} + \begin{bmatrix}
\varepsilon & 0 & 0 \\
0 & \frac{\varepsilon}{d(i, j)} & 0 \\
0 & 0 & 0
\end{bmatrix} + \begin{bmatrix}
0 & 0 & 0 \\
0 & \frac{1}{d(i, j)} & 0 \\
0 & 0 & \frac{\varepsilon}{d(i, j)}
\end{bmatrix}.
\]

For the remainder $N$ from (5.27) we find
\[
\tilde{N} = \hat{A}(\varepsilon) - \tilde{P}_1 \tilde{P}_2 = \frac{1}{h^2} \begin{bmatrix}
\varepsilon & 0 & 0 \\
0 & \frac{\varepsilon}{d(i, j)} & 0 \\
0 & 0 & \frac{\varepsilon}{d(i, j)}
\end{bmatrix}.
\]

Using an induction argument and the relations (5.34), one checks $d(i, j) \geq 1$. Therefore, the condition (5.29) is satisfied with $\bar{\varepsilon} = \frac{1}{\varepsilon}$. One can also verify (5.28), with $\alpha = \frac{2(1 + \varepsilon + \sqrt{\varepsilon})}{1 - \varepsilon}$. However, the proof is quite technical and thus we skip it (we refer the interested reader to [216]).

The assumption (5.30) is valid with $C_\varepsilon = 1$ due to
\[
||S||_2 = ||W^{-1}N||_2 \leq ||W^{-1}||_2 ||N||_2 \leq ||D||_2 ||(L + D)^{-1}||_2 ||N||_2 < 1. \tag{5.35}
\]

Since $||A|| = c h^{-2}$, for the function $\eta$ from the approximation property we get
\[
\eta(v, \varepsilon) = c \varepsilon \max \left\{ \frac{1}{\nu - 1}, \theta^{\nu - 2} \right\}. \tag{5.36}
\]

The smoothing property should be complemented with the approximation property: For the function $\zeta$ from (5.8) we will prove the estimate $\zeta(\varepsilon) \leq \frac{\varepsilon}{\varepsilon}$.

**Block Gauss–Seidel smoothing iterations**

For the matrix $A$ from (5.33), consider the decomposition on block-diagonal and block left and right triangle matrices $D$, $L$, and $L^T$:
\[
D = \frac{1}{h^2} \begin{bmatrix}
A & 0 & \cdots & 0 \\
0 & \ddots & \ddots & \vdots \\
0 & \cdots & A_{n-1}
\end{bmatrix}, \quad L = \frac{1}{h^2} \begin{bmatrix}
I_\varepsilon & 0 & \cdots & 0 \\
I_\varepsilon & \ddots & \ddots & \vdots \\
0 & \cdots & I_\varepsilon & 0
\end{bmatrix}.
\]

Consider a block variant of the symmetric Gauss–Seidel method. One iteration in the block method is defined as
\[
(L + D)x^{k+\frac{1}{2}} + L^T x^k = b,
\]
\[
(D + L^T)x^{k+1} + L x^{k+\frac{1}{2}} = b, \quad k = 0, 1, 2, \ldots.
\]

Since the blocks $A_i$ are the tridiagonal matrices, one iteration of the method (one smoothing step) has the optimal complexity with respect to the number of unknowns.
The symmetric block Gauss–Seidel method can be written as the basis iteration (1.11) with the preconditioner
\[ W = (L + D)D^{-1}(D + L^T). \]  
(5.37)

To check the smoothing property, we apply Lemma 5.6. The iteration matrix is \( S = I - W^{-1}A \). In the notation of Lemma 5.6, we have \( S = W^{-1}N, N = W - A \). To get an expression for \( N \), we note that
\[ W = (L + D)D^{-1}(D + L^T) = A + LD^{-1}L^T. \]
Hence
\[ N = LD^{-1}L^T. \]  
(5.38)

Now we check the assumptions (5.28)–(5.30) of Lemma 5.6. First, \( N > 0 \); therefore (5.28) is valid for any \( \alpha \geq 0 \). Second, we note that all the eigenvalues of the symmetric blocks \( A_i \) satisfy \( \lambda(A_i) > 2\epsilon \), and thus for the block-diagonal matrix we have
\[ \|D^{-1}\|_2 = \max_i \|A_i^{-1}\|_2 \leq \frac{1}{2} \theta^{-1} h^2. \]

Also we know that \( \|L\| = h^{-2} ||L|| = \theta h^{-2} \). We get
\[ \|N\|_2 \leq \|L\|_2 \|D^{-1}\|_2 \leq \frac{\epsilon}{H^2}. \]

Therefore, the constant \( \delta \) from (5.29) can be taken as \( \delta = \frac{\epsilon}{2} \theta h^{-2} \). Finally, similarly to (5.35) we check the assumption (5.30) with \( C_i = 1 \).

Therefore, Lemma 5.6 provides us with the smoothing property
\[ \|AS\|_2 \leq \frac{\epsilon}{H^2} \frac{1}{\nu-1}. \]

As in the case of ILU iterations, the smoothing property of the block Gauss–Seidel method should be complemented by the approximation property (5.7) such that the function \( \zeta \) from (5.8) should obey the estimate \( \zeta(\epsilon) \leq \epsilon^{-\tau} \), with a parameter-independent constant \( \epsilon \).

**Approximation property**

As we already know, the key to proving the approximation property is a suitable estimate of the \( L^2 \)-norm of the finite element error by the \( L^2 \)-norm of the right-hand side \( f \). To obtain the desired estimate for \( \zeta(\epsilon) \) from (5.8), it is sufficient to verify the error estimate
\[ \|u - u_h\| \leq c \frac{h^2}{\epsilon} \|f\|. \]  
(5.39)

We employ the arguments already familiar to the reader (see (5.16) for the example of the reaction-diffusion problem). The bilinear form corresponding to (5.3) is given by
\[ a(u, v) = \int_\Omega \{ \frac{\partial u}{\partial x} \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \frac{\partial v}{\partial y} \} dx. \]

For the error function \( e_h = u - u_h \) we have
\[ \epsilon \left( \frac{\partial e_h}{\partial x} \right)^2 + \left( \frac{\partial e_h}{\partial y} \right)^2 = a(e_h, e_h) = a(u - v_h, e_h) \]
\[ \leq \left( \frac{\epsilon}{\| \partial(u - v_h) \|} \right) + \left( \frac{\epsilon}{\| \partial(u - v_h) \|} \right) \left( \frac{\epsilon}{\| \partial u \|} + \frac{\epsilon}{\| \partial v_h \|} \right)^2. \]
Since $v_b$ is an arbitrary function from $U_h$, we get

$$
\varepsilon \left\| \frac{\partial e_b}{\partial x} \right\|^2 + \left\| \frac{\partial e_b}{\partial y} \right\|^2 \leq c h^2 \left( \varepsilon \left\| \frac{\partial^2 u}{\partial x^2} \right\|^2 + \left\| \frac{\partial^2 u}{\partial y^2} \right\|^2 + \left\| \frac{\partial^2 u}{\partial xy} \right\|^2 \right). \tag{5.40}
$$

Here we used the following approximation properties of the finite element spaces:

$$
\inf_{v_h \in U_h} \left\| \frac{\partial (u - v_h)}{\partial x} \right\| \leq c \left( \left\| \frac{\partial^2 u}{\partial x^2} \right\| + \left\| \frac{\partial^2 u}{\partial xy} \right\| \right), \tag{5.41}
$$

$$
\inf_{v_h \in U_h} \left\| \frac{\partial (u - v_h)}{\partial y} \right\| \leq c \left( \left\| \frac{\partial^2 u}{\partial y^2} \right\| + \left\| \frac{\partial^2 u}{\partial xy} \right\| \right). \tag{5.42}
$$

The proof of these simple estimates can be found, for example, in [40].

To show (5.39), we also need a priori estimates of the solution $u$ norms by the $L^2$-norm of the right-hand side $f$. We begin by squaring the equality (5.3) and integrating it over the domain $\Omega$. We get

$$
\|f\|^2 = \int_{\Omega} \left( \varepsilon \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right)^2 \, dx. \tag{5.43}
$$

Thanks to the homogeneous Dirichlet boundary conditions and the simple geometry of $\Omega$, by applying integration by parts we get

$$
\int_{\Omega} \left( \varepsilon \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right)^2 \, dx = \varepsilon^2 \left\| \frac{\partial^2 u}{\partial x^2} \right\|^2 + \left\| \frac{\partial^2 u}{\partial y^2} \right\|^2 + 2\varepsilon \int_{\Omega} \frac{\partial^2 u}{\partial x^2} \frac{\partial^2 u}{\partial y^2} \, dx
$$

$$
= \varepsilon^2 \left\| \frac{\partial^2 u}{\partial x^2} \right\|^2 + \left\| \frac{\partial^2 u}{\partial y^2} \right\|^2 + 2\varepsilon \int_{\Omega} \frac{\partial^2 u}{\partial x^2} \frac{\partial^2 u}{\partial y^2} \, dx.
$$

Using this relation and (5.43), we find

$$
\varepsilon^2 \left\| \frac{\partial^2 u}{\partial x^2} \right\|^2 + \left\| \frac{\partial^2 u}{\partial y^2} \right\|^2 + 2\varepsilon \left\| \frac{\partial^2 u}{\partial xy} \right\|^2 = \|f\|^2. \tag{5.44}
$$

Now (5.44) and (5.40) yield

$$
\varepsilon \left\| \frac{\partial e_b}{\partial x} \right\|^2 + \left\| \frac{\partial e_b}{\partial y} \right\|^2 \leq c \frac{h^2}{\varepsilon} \|f\|^2. \tag{5.45}
$$

Applying the one-dimensional Friedrichs inequality, we get in $\Omega = (0,1) \times (0,1)$ the estimate

$$
\|v\| \leq \left\| \frac{\partial v}{\partial y} \right\| \quad \forall v \in H^1(\Omega).
$$

Using this in (5.45), one proves the $O(\frac{1}{\varepsilon})$ order of convergence: $\|e_b\| \leq c \frac{1}{\sqrt{\varepsilon}} \|f\|$. However, to prove the robust convergence estimate for multigrid using smoothing and approximation properties, we need the $O(\frac{1}{\varepsilon})$ order of convergence for the finite element error. To improve the convergence order, we employ a duality argument similar to those used in
the proof for the reaction-diffusion problem: Given the error function \( e_h \in L^2(\Omega) \cap U_h \), consider the function \( w \in H^1(\Omega) \cap H^2(\Omega) \), which solves the problem

\[
-\varepsilon \frac{\partial^2 w}{\partial x^2} - \frac{\partial^2 w}{\partial y^2} = e_h \text{ in } \Omega, \quad w|_{\partial \Omega} = 0. \tag{5.46}
\]

Squaring the equality and integrating by parts, we get

\[
\varepsilon^2 \left( \left\| \frac{\partial^2 w}{\partial x^2} \right\|^2 + \left\| \frac{\partial^2 w}{\partial y^2} \right\|^2 + 2\varepsilon \left\| \frac{\partial^2 w}{\partial xy} \right\|^2 \right) = \|e_h\|^2. \tag{5.47}
\]

Furthermore, multiplying (5.46) with \( e_h \), integrating over \( \Omega \) and by parts, and using the orthogonality of the error to any function \( w_h \) from \( U_h \), and the Cauchy inequality, we obtain

\[
\|e_h\|^2 = a(w, e_h) = a(w - w_h, e_h)
\]

\[
\leq \left( \varepsilon \left( \left\| \frac{\partial e_h}{\partial x} \right\|^2 + \left\| \frac{\partial e_h}{\partial y} \right\|^2 \right) \right) \left( \left\| \frac{\partial (w - w_h)}{\partial x} \right\|^2 + \left\| \frac{\partial (w - w_h)}{\partial y} \right\|^2 \right)^{\frac{1}{2}}.
\]

Now (5.45) and the approximation property (5.41)-(5.42) imply

\[
\|e_h\|^2 \leq c \frac{h}{\sqrt{\varepsilon}} \|f\| \left( \varepsilon \left( \left\| \frac{\partial^2 w}{\partial x^2} \right\|^2 + \left\| \frac{\partial^2 w}{\partial y^2} \right\|^2 + \left\| \frac{\partial^2 w}{\partial xy} \right\|^2 \right)^{\frac{1}{2}} \right).
\]

Finally, (5.47) and the last estimate lead us to the desired result:

\[
\|e_h\| \leq c \frac{h^2}{\varepsilon} \|f\|. \quad \blacksquare
\]

Summarizing, the symmetric block Gauss–Seidel iteration as a smoother ensures the robustness of the multigrid. ILU(0) smoothings result in an almost robust multigrid method. We say "almost" since the parameter \( \theta \) from (5.36) still depends on \( \varepsilon \). A modification of ILU(0) from [216] gets rid of this dependence.

### 5.1.8 Upwind schemes for the convection-diffusion equations

Discrete convection-diffusion equations are known to be challenging to solve numerically. Some interesting and important properties of the equations become clear if one considers the simplest one-dimensional problem

\[
-\varepsilon u'' - u' = 0, \quad u(0) = 0, \quad u(1) = 1, \tag{5.48}
\]

with the known exact solution

\[
u(x) = \frac{1 - \exp(-\frac{x}{\varepsilon})}{1 - \exp(-1/\varepsilon)}. \tag{5.49}
\]

This solution is plotted in Figure 5.1(a). We start with this model example and briefly discuss stable discretization schemes for convection-diffusion equations.
5.1. Multigrid preconditioners for singular-perturbed problems

Figure 5.1. (a) The solution to the convection-diffusion equation (5.48) on \([0, 0.6]\) for \(\varepsilon = 10^{-2}\); (b) the central finite differences solution (5.50), \(h = \frac{1}{20}\); (c) the upwind finite differences solution (5.52).

A well-known stability issue is related to discretizations of the convection-diffusion problem. To understand it, consider a finite difference method. This differs from the rest of this book, where finite element methods are used. (A finite element method for the convection diffusion equations will be considered in the next section.) We start with the second order accurate finite difference scheme,

\[-\varepsilon \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} - \frac{u_{i+1} - u_{i-1}}{2h} = 0, \quad i = 1, \ldots, n - 1, \quad u_0 = 0, \quad u_n = 1,\]  

(5.50)

where \(u_i = u_b(i \frac{h}{b})\), \(h = 1/n\). The same discrete relations as in (5.50) are obtained from the finite element method for (5.48) with piecewise linear continuous \(u_b\). The exact solution for the difference scheme (5.50) can be readily checked to be

\[u_b(i \frac{h}{b}) = \frac{1 - r^i}{1 - r^n}, \quad \text{where} \quad r = \frac{2\varepsilon - b}{2\varepsilon + b},\]  

(5.51)

For \(h > 2\varepsilon\), \(r^i\) changes its sign from one node to the next, and the solution (5.51) has oscillations (see Figure 5.1(b)). There are no oscillations in the continuous solution (5.49), and therefore their appearance in a discrete one should be avoided.\(^{12}\) The magnitude of these oscillations grows when the relation \(\frac{h}{2\varepsilon}\) becomes larger.

The condition \(h \leq 2\varepsilon\) can be very restrictive (especially for two- and three-dimensional problems), and hence small values of \(\varepsilon\) lead to systems of algebraic equations with an unacceptably large number of unknowns. The simplest scheme that gives qualitatively reasonable results for \(h > 2\varepsilon\) is the first order scheme with upwind differences. For the problem (5.48) this scheme reads

\[-\varepsilon \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} - \frac{u_{i+1} - u_i}{h} = 0, \quad i = 1, \ldots, n - 1, \quad u_0 = 0, \quad u_n = 1,\]  

(5.52)

\(^{12}\)Indeed, for real-life problems the exact solution is usually not known. Hence, the occurrence of oscillations in numerical solutions can be interpreted as a correct physical phenomenon, for example, a resonance phenomenon. For this reason, numerical schemes able to produce nonphysical (numerical) oscillations are commonly considered as inappropriate (unstable). From the mathematical point of view the occurrence of numerical oscillations manifests the lack of suitable bounds for the \(C\)-norm or \(H^1\)-norm of the error of the discretization method. All these questions gained a lot of attention in the literature—the reader may consult the monographs [78, 172], among others.
with the exact solution (see Figure 5.1(c))

\[ u_h(i, j) = \frac{1 - r^i}{1 - r^a}, \quad \text{where } r = \frac{\epsilon}{\epsilon + h}. \]  

(5.53)

The scheme (5.52) can be rewritten in the form

\[-\left(\epsilon + \frac{h}{2}\right) \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} - \frac{u_{i+1} - u_{i-1}}{2h} = 0, \quad i = 1, \ldots, n - 1, \]

\[ u_0 = 0, \quad u_n = 1. \]

Here one can recognize the second order central difference scheme for an equation similar to (5.48), with \( \epsilon \) replaced by \( \epsilon + \frac{h}{2} \). Hence, the upwind difference scheme can be interpreted as a method adding artificial (numerical) diffusion to discrete equations.

Consider now a more general case of two-dimensional convection-diffusion equations,

\[-\epsilon \Delta u + a_1 \frac{\partial u}{\partial x} + a_2 \frac{\partial u}{\partial y} = f \quad \text{in } \Omega, \quad u|_{\partial \Omega} = 0. \]  

(5.54)

A common approach to discretizing (5.54) is a finite difference or finite element method which adds (sometimes implicitly) a certain amount of artificial diffusion if the relation \( \frac{h}{\epsilon}(\|a_1\| + \|a_2\|) \) is large. In this more general case, in any point \((x', y') \) of the domain, in particular in all grid nodes, the vector \((a_1(x', y'), a_2(x', y')) \) defines the direction of a “wind.” The simplest stable scheme uses upwind finite differences to approximate the first order terms \( a_1 \frac{\partial u}{\partial x} \) and \( a_2 \frac{\partial u}{\partial y} \).

### 5.1.9 A smoother for the convection-diffusion equations

For the problem (5.54) consider the case of constant coefficients \( a_1 > 0 \) and \( a_2 > 0 \). Let us approximate the problem using the first order upwind differences for the first order terms. The stencil of the discretization is

\[ \hat{L}_h(\epsilon) = \frac{1}{h^2} \begin{bmatrix} 0 & -\epsilon & 0 \\ -\epsilon - a_1 h & 4\epsilon + a_1 h + a_2 b & -\epsilon \\ 0 & -\epsilon - a_2 b & 0 \end{bmatrix}. \]  

(5.55)

For the resulting system of linear algebraic equations, the Gauss-Seidel method (with the lexicographical ordering of nodes) is an “exact solver” in the limit case of \( \epsilon \to 0 \) (explain why?). Thus, one desirable (but neither necessary nor sufficient) condition for building a robust multigrid method is satisfied. In the case of varying \( a_1 \) and \( a_2 \), one should enumerate grid nodes in such a way that the node indexes increase in the upstream direction. For example, if, for a node with the coordinates \((i h_x, j h_y)\), it holds that \( a_1(i h_x, j h_y) > 0 \), \( a_2(i h_x, j h_y) < 0 \), then unknowns corresponding to the nodes \(((i + 1) h_x, j h_y)\) and \((i h_x, (j - 1) h_y)\) should have larger indexes in the resulting algebraic system. In this case, the matrix of the finite difference system preserves a structure such that the upper-left block vanishes for \( \epsilon \to 0 \). Algorithms for performing the downstream enumeration can be found in the literature [98, 23]. They were proved to work well under certain assumptions on \( a_1 \) and \( a_2 \). The influence of different enumeration strategies on solver performance for convection-diffusion equations is studied in [28].

A canonical finite element method does not produce stencil (5.55). A finite element method may lead to discretization stencils similar to (5.55) only if one uses special weighted...
quadrature formulas to compute the elements of stiffness matrix $A_h$. The weights depend, in general, on $a$ and mesh properties; see, for example, [83]. In this case, the choice of a coarse grid matrix is not straightforward, since the construction of the coarse grid problem by $A_{H} = r A_h p$ does not always lead to a suitable coarse grid matrix $A_{H}$. For the canonical choice of $p$ and $r$ one loses the structure of the discrete problem, which admits efficient smoothing steps. To overcome this, one can define $p$ or $r$ depending on $a_1, a_2$, or one can discretize (5.54) directly on a coarse grid, using the upwind difference scheme on the coarse grid. In the latter case, the vector function $(a_1, a_2)$ should be restricted to the coarse grid. The construction of problem-dependent $p$ and $r$ is discussed in [228].

For certain vector functions $(a_1, a_2)$ this leads to robust multigrid methods.

The downside of the approach described in this section is the low (only first order) accuracy of the discretization scheme. Another disadvantage is the lack of feasible enumeration strategies for complex real-life flows. In the next section, we study in more detail a higher order "upwind" scheme. This scheme inherits the stability property of the first order upwind schemes and, in general, provides more accurate discrete solutions. Using schemes of high order, one often loses the property of the discrete problem to admit exact solvers in the limit case of $\epsilon = 0$. Nevertheless, multigrid methods remain efficient for the resulting systems of linear algebraic equations.

### 5.1.10. The streamline-diffusion finite element method

In this section, we consider an approach to building higher order schemes for the convection-diffusion equations, which are stable and resistant to numerical oscillations. The method is well known as the streamline-diffusion finite element method or streamline upwinding Petrov–Galerkin (SUPG) method. It was first proposed by Brooks and Hughes in [49] and further gained significant attention. The scheme is well suited for finite element discretization of problems involving transport (convection) terms.

Consider a triangulation $T_h$ of a computation domain. Assume that for any element $\tau \in T_h$ we are given parameter $\delta_{\tau}$, depending on $\epsilon$, $\tau \in T_h$, and $a(x) = (a_1(x), a_2(x))$. The finite element solution $u_h \in U_h$ of the SUPG method satisfies the following relation for any $v_h$ from $U_h$:

$$
\epsilon (\nabla u_h, \nabla v_h) + (a \nabla u_h, v_h) + \sum_{\tau \in T_h} \delta_{\tau} (-\epsilon \Delta u_h + a \nabla u_h - f, a \nabla v_h) = (f, v_h).
$$

(5.56)

For a more compact presentation, here and in what follows we use the common notation for the convection terms:

$$
a \nabla u = a_1 \frac{\partial u}{\partial x} + a_2 \frac{\partial u}{\partial y}.
$$

The following comments are in order.

1. The third term in (5.56) is evaluated elementwise. We use the notation

$$
(\phi, \psi)_\tau := \int_\tau \phi(x) \psi(x) \, dx.
$$

Since on every triangle $\tau \in T_h$ the function $u_h$ is smooth (a polynomial of degree $k$), the third term in (5.56) is well defined. Moreover, for linear or bilinear $u_h$ it holds that $\epsilon \Delta u_h = 0$. 


2. The first and second terms on the left-hand side of (5.56), as well as the right-hand side, emerge from the weak formulation of (5.54) and constitute the standard finite element method. The additional term (the role of this term will become clear) can be observed as the identity (5.54) multiplied by \( \delta \cdot a \nabla v_h \) elementwise. If we were to substitute the solution of the differential problem (5.54) instead of \( u_h \), then this additional term would vanish.

3. If \( \delta = 0 \) for all \( \tau \in T_h \), the scheme (5.56) becomes the standard Galerkin finite element approximation to problem (5.54). A “stabilizing” effect of the additional term has a clear interpretation for the uniform parameters \( \delta = \delta \). Assume linear or bilinear finite elements. In this case, \( \epsilon \Delta u_h = 0 \), and the only term added to the left-hand side of (5.56) is

\[
\delta(a \nabla u_h, a \nabla v_h) = \delta([a \otimes a] \nabla u_h, \nabla v_h).
\]

The latter term is the finite element discretization of the anisotropic diffusion term (for smooth \( u \)),

\[
\delta \text{ div}([a \otimes a] \nabla u). \tag{5.57}
\]

For \( \delta > 0 \), the additional numerical diffusion is added, and thus we may expect an enhancement of the stability of the method. The tensor \([a \otimes a]\) has its principal axis in the direction \( a \), so (5.57) adds additional diffusion only in the flow direction.

4. In addition, and unlike the first order upwind differences from the previous section, the SUPG method adds the additional term

\[
\sum_\tau \delta_\tau(f, a \nabla v_h)
\]

to the right-hand side in (5.56). With the help of this extra term, it is ensured that the finite element formulation (5.56) is consistent: For any \( v_h \in U_h \) the continuous solution to (5.54) satisfies (5.56) if it is inserted instead of \( u_h \). This leads to the orthogonality property similar to (5.13) and enables one to obtain a higher order accuracy method.

5. An “optimal” choice of stabilization parameters \( \delta \) is a delicate question. Roughly speaking, one wishes to add artificial diffusion only in those regions where the mesh Péclet number \( \text{Pe}_h \) is high. The mesh Péclet number, \( \text{Pe}_h \), is defined locally for each element \( \tau \in T_h \) as

\[
\text{Pe}_h = \frac{b_\tau \|a\|_\tau}{2\epsilon},
\]

where \( \|a\|_\tau \) is a local norm of \( a \) on an element \( \tau \) (we will use the \( L^\infty \)-norm), and \( b_\tau \) is the diameter of the element \( \tau \).

The general recommendation for defining \( \delta \) is to respect the “double asymptotic law” [172]:

\[
\delta \sim \frac{b_\tau^2}{\epsilon} \quad \text{for} \quad \text{Pe}_h \to 0, \tag{5.58}
\]

\[
\delta \sim \frac{b_\tau}{\|a\|_\tau} \quad \text{for} \quad \text{Pe}_h \to \infty. \tag{5.59}
\]

Sometimes it is alternatively recommended to set \( \delta = 0 \) for \( \text{Pe}_h \leq 1 \), recovering the standard Galerkin approximation.
5.1. Multigrid preconditioners for singular-perturbed problems

Some particular recipes for the choice of $\delta_\tau$ from the literature are given below:

$$\delta_\tau = \frac{1}{2} \left( \coth(\text{Pe}_h) - \frac{1}{\text{Pe}_h} \right),$$  \hspace{1cm} (5.60)

$$\delta_\tau = \frac{h_\tau}{\mu |a|_r} \left( 1 + \text{Pe}_h \right), \hspace{1cm} \hat{\delta} \in [0.2, 1],$$  \hspace{1cm} (5.61)

$$\delta_\tau = \frac{h_\tau}{2|a|_r} \left( 1 - \frac{1}{\text{Pe}_h} \right) \text{ if } \text{Pe}_h \geq 1, \hspace{1cm} \delta_\tau = 0 \text{ otherwise.}$$  \hspace{1cm} (5.62)

Theorem 5.7 below presents a typical higher order error estimate for the SUPG method. First we make some assumptions on $\mathbb{U}_h$.

Assume that $\mathbb{U}_h$ consists of polynomials of the degree $k \geq 1$ and the triangulation is such that the standard interpolation results hold: For any smooth function $v : v|_{\partial \Omega} = 0$ there exists an interpolant $\hat{v} \in \mathbb{U}_h$ such that for every $\tau \in \mathcal{T}_h$ it holds that

$$||v - \hat{v}||_{H^1(\tau)} \leq ch^{k-i+1} ||v||_{H^{i+1}(\omega(\tau))}, \hspace{1cm} i = 0, 1, k = 0, \ldots, l,$$  \hspace{1cm} (5.63)

where $\omega(\tau)$ is the neighborhood of $\tau$ consisting of all triangles from $\mathcal{T}_h$ sharing a vertex with $\tau$. Note that (5.63) is the local interpolation result. In particular, it implies the global (in the entire domain $\Omega$) interpolation result with the same norms and of the same orders with respect to $h$.

The natural norm for the analysis of the stability and convergence of the method is the following mesh-dependent norm on $H^1(\Omega)$:

$$|||u||| := \left( \varepsilon ||\nabla u||^2 + \sum_{\tau \in \mathcal{T}_h} \delta_\tau ||a \nabla u||^2 \right)^{\frac{1}{2}},$$

where $a$ is assumed to have a finite $L^\infty$-norm on every $\tau$. For further convenience, we denote

$$a_\tau := ||a||_{\infty, \tau} = \text{ess sup}_{x \in \tau} (|a_1(x)| + |a_2(x)|).$$

Finally, we will need the constant $\mu_h$ from the inverse inequality

$$||\Delta v_h||_\tau \leq \mu_h h^{-1}_\tau ||\nabla v_h||_\tau \hspace{1cm} \forall \tau \in \mathcal{T}_h, v_h \in \mathbb{U}_h.$$  \hspace{1cm} (5.64)

Note that $\mu_h = 0$ for linear finite elements.

**Theorem 5.7.** Assume the interpolation properties (5.63) hold and $a \in L^\infty(\Omega)^2$, $\text{div} a = 0$. Let $u(x, y)$ be a sufficiently smooth solution to (5.54). Consider (5.56), with a set of parameters $\delta_\tau$, satisfying the condition

$$0 \leq \delta_\tau \leq \frac{h^2_\tau}{\mu \varepsilon} \hspace{1cm} \forall \tau \in \mathcal{T}_h,$$  \hspace{1cm} (5.65)

where $\mu = \min\{1, \mu_h\}$. Then (5.56) possesses a unique finite element solution $u_h$ from $\mathbb{U}_h$ and

$$|||u - u_h|||^2 \leq \varepsilon \sum_{\tau \in \mathcal{T}_h} \left( \varepsilon + a_\tau^2 \delta_\tau + \min \left( \frac{a_\tau}{\varepsilon}, \frac{1}{\delta_\tau} \right) \frac{h^2_\tau}{\mu \varepsilon} \right) h^{2k} ||u||_{H^{i+1}(\tau)}^2$$  \hspace{1cm} (5.66)

for $k = 1, \ldots, l$. 

Chapter 5. Some Applications

2. The finite element solution to (5.56). The stability estimate (5.67), in particular, implies the existence and uniqueness of the solution. Introducing the bilinear form on $H_0^1(\Omega) \times H_0^1(\Omega)$,

$$a_h(u, v) := \varepsilon \left( \nabla u, \nabla v \right) + (a \nabla u, v) + \sum_{\tau \in T_h} \delta_\varepsilon \left( -\varepsilon \Delta u + a \nabla u, \nabla v \right)_\tau,$$

one gets for all $v_h \in \mathbb{U}_h$

$$a_h(v_h, v_h) \geq \frac{1}{2} \|v_h\|^2. \tag{5.67}$$

Indeed, the skew-symmetry of the convection term, $(a \nabla v, v) = 0$ for $\text{div} a = 0$, gives

$$a_h(v_h, v_h) = \varepsilon \| \nabla v_h \|^2 + \sum_{\tau \in T_h} \delta_\varepsilon \left( -\varepsilon \Delta v_h + a \nabla v_h \right)_\tau + \sum_{\tau \in T_h} \delta_\varepsilon |a \nabla v_h|_\tau^2 \geq \|v_h\|^2 - \sum_{\tau \in T_h} \delta_\varepsilon \varepsilon^2 |\Delta v_h|_\tau^2 - \sum_{\tau \in T_h} \delta_\varepsilon |a \nabla v_h|_\tau^2 \geq \|v_h\|^2 - \frac{1}{2} \sum_{\tau \in T_h} \delta_\varepsilon \varepsilon^2 |\Delta v_h|_\tau^2 - \frac{1}{2} \sum_{\tau \in T_h} \delta_\varepsilon |a \nabla v_h|_\tau^2 \geq \|v_h\|^2 - \frac{1}{2} \sum_{\tau \in T_h} \delta_\varepsilon \varepsilon^2 \mu h^{-2} |\Delta v_h|_\tau^2 - \frac{1}{2} \sum_{\tau \in T_h} \delta_\varepsilon |a \nabla v_h|_\tau^2 \geq \|v_h\|^2 - \frac{1}{2} \sum_{\tau \in T_h} \delta_\varepsilon |\nabla v_h|_\tau^2 - \frac{1}{2} \sum_{\tau \in T_h} \delta_\varepsilon |a \nabla v_h|_\tau^2 = \frac{1}{2} \|v_h\|^2.

The stability estimate (5.67), in particular, implies the existence and uniqueness of the finite element solution to (5.56).

2. The next step is to check the continuity estimate of the following form: For any $v \in H_0^1(\Omega) : \Delta v \in L^2(\Omega)$ and any $v_h \in \mathbb{U}_h$ it holds that

$$a_h(v, v_h) \leq \|v\| \|v_h\| \left( \delta_\varepsilon \varepsilon^2 |\Delta v|_\tau^2 + \delta_\varepsilon |a \nabla v|_\tau^2 \right) \leq \|v\| \|F(v)\| \tag{5.68}$$

This estimate holds due to the inequality

$$\|a \nabla v_h, v\| = \|a \nabla v_h, v\| \leq \sum_{\tau \in T_h} \min \left( \frac{\delta_\varepsilon \varepsilon}{\sqrt{\delta_\varepsilon}}, \frac{1}{\sqrt{\delta_\varepsilon}} \right) \left( \sqrt{\varepsilon} |\nabla v_h|_\tau + \sqrt{\delta_\varepsilon} |a \nabla v_h|_\tau \right) \|v\|_\tau$$

for the convection term, the Cauchy inequalities for the other terms, and the condition (5.65) on $\delta_\varepsilon$, which implies $\varepsilon \delta_\varepsilon \leq \varepsilon h^{-2}$.

3. Finally, one needs the orthogonality property of the SUPG method:

$$a_h(u - u_h, v_h) = 0 \quad \forall v_h \in \mathbb{U}_h.$$

Thanks to the stability and continuity properties, one gets for arbitrary $v_h \in \mathbb{U}_h$,

$$\frac{1}{2} \|u_h - v_h\|^2 \leq a_h(u_h - v_h, u_h - v_h) = a_h(u_h - v_h, u - v_h) \leq \|u_h - v_h\| \|F(u - v_h)\|.$$
This yields
\[ \| | u_h - v_h | | \leq 2F(u - v_h). \]
Therefore, the triangle inequality implies the estimate on the error
\[ \| | u - u_h | | \leq \| | u - v_h | | + \| | u_h - v_h | | \leq \| | u - v_h | | + 2F(u - v_h). \]
Since \( v_h \) is arbitrary, one can take \( v_h \) as an interpolant to \( u \), satisfying the assumptions (5.63). One proves the assertion of the theorem after elementary computations (check it!).

To design the stabilization parameter \( \delta \tau \), let us optimize the error estimate (5.66) by balancing the \( \delta \)-dependent terms on the right-hand side. We obtain
\[ \delta \tau a^2 b^2 \sim \min \left\{ \frac{a^2}{\varepsilon}, \frac{1}{\delta \tau} \right\} b^4. \]
This balance provides us with the double asymptotic law (5.58)–(5.59). Note that \( \delta \tau \) from the relations (5.60)–(5.61) satisfies this law as well as the assumption (5.65).

**Corollary 1.** Define the parameter \( \delta \tau \) by the formula (5.61). Substituting these values into (5.66), after elementary manipulations we get
\[ \| | u - u_h | | \leq c \sum_{\tau \in \mathcal{T}_h} \left( \varepsilon b^2 + a \frac{Pe_h}{Pe_h + 1} b^3 \right) + a \min \left\{ \frac{Pe_h}{Pe_h + 1} \right\} b^4 \leq c \sum_{\tau \in \mathcal{T}_h} \left( \varepsilon b^2 + a \frac{Pe_h}{Pe_h + 1} \right) b^{2(\varepsilon - 1)} \| | u | |^2_{H^{2(\varepsilon - 1)}(\tau)} \right\} \leq c \sum_{\tau \in \mathcal{T}_h} \left( \varepsilon b^2 + a \frac{Pe_h}{Pe_h + 1} \right) b^{2(\varepsilon - 1)} \| | u | |^2_{H^{2(\varepsilon - 1)}(\tau)} \right\}, \] (5.69)

We can scale the equations (5.54) to ensure \( a \leq 1 \). Then (5.69) yields the well-known \( \frac{3}{2} \) order of the convergence of the SUPG method for the linear finite element and sufficiently small \( \varepsilon (\varepsilon \lesssim h) \):
\[ \| | u - u_h | | \| | u_h | |_{H^2(\Omega)}, \quad b = \max_{\tau} b_{\tau}. \] (5.70)
Actually, for special grids the convergence order can be proved to be even higher; see [230].

It should be noted that for \( \varepsilon = 0 \) the norm \( \| | \cdot | | \) can be degenerated. The \( \varepsilon \)-dependent estimate follows only for the second term in the definition of the norm \( \| | u - u_h | | \):
\[ \sum_{\tau} \delta \tau \| | a \nabla (u - u_h) | |^2. \] (5.71)
This term controls the \( L^2 \)-norm of the error for a specific class of the vector functions \( a \); e.g., \( a \) is a homogeneous flow in some direction, but it may fail to define a norm in a general case. Nevertheless, numerical experiments with the SUPG method typically demonstrate the \( \frac{3}{2} \) (or even better) order of the convergence in the \( L^2 \)-norm. To see how the estimate (5.69) explains stability properties of the method, we note that the \( \varepsilon \)-independent estimate on the streamwise derivative of the error, i.e., the estimate on

\[ \sum_{\tau} \delta \tau \| | a \nabla (u - u_h) | |^2. \] (5.71)
∥a∇(u − uh)∥, shows that numerical oscillations in the streamwise directions are suppressed by the method even for small ϵ.

The system of linear algebraic equations (Az = b), corresponding to the finite element SUPG formulation (5.56), is known to be amenable to multigrid methods. Numerical experience shows that for a robust convergence, coarse grid matrices, Ah, should be constructed by the direct discretization of (5.4) on a coarse grid by the SUPG scheme (5.56), with stabilization parameters δk corresponding to the coarse grid. Then, at least for some relatively simple flows, standard multigrid methods provide robust and convergent iterations.

5.1.11 Analysis of a multigrid preconditioner for convection-diffusion equations

We discuss analysis of a multigrid preconditioner for a model convection-diffusion problem. The discussion follows [157], where omitted details can be found. Currently, only a simple problem is amenable to rigid theoretical analysis of multigrid preconditioners.

Considering the following model convection-diffusion problem: Find a function u such that

\[-\varepsilon \Delta u + u_x = f \quad \text{in} \quad \Omega := (0,1)^2,\]
\[\frac{\partial u}{\partial x} = 0 \quad \text{on} \quad \Gamma_E := \{(x,y) \in \partial \Omega | x = 1\},\]
\[u = 0 \quad \text{on} \quad \partial \Omega \setminus \Gamma_E,\]

with \(\varepsilon \in [0,1]\). Consider streamline-diffusion finite element method (5.56) based on \(P_1\) triangular elements. Let \(U_k\) be the finite element space of continuous piecewise linear functions on a uniform triangulation of the unit square, with mesh size \(h_k := 2^{-k}\). All functions in \(U_k\) are assumed to be zero at \(\partial \Omega \setminus \Gamma_E\). The streamline-diffusion method leads to a hierarchy of discrete problems: Find \(u_k \in U_k\) such that

\[(\varepsilon + \delta_k h_k)(u_k)_x + \varepsilon ((u_k)_y, v_y) + ((u_k)_x, v_x) = (f, v + \delta_k h_k v_x)\]

for all \(v \in U_k\), where we set

\[\delta_k = \begin{cases} \tilde{\delta} & \text{if } \frac{h_k}{2^k} \geq 1, \\ 0 & \text{otherwise.} \end{cases}\]

The finite element method results in a stable discretization on every grid level because of the \(k\)-dependence of the stabilization parameter \(\delta_k\). This property is important for the performance of the multigrid method.

In \(U_k\) we use the standard nodal basis functions. The resulting stiffness matrix is denoted by \(A_k\). In each interior grid point this matrix has a stencil representation

\[\tilde{A}_k = \frac{1}{h_k^2} \begin{bmatrix} 0 & -\varepsilon_k & 0 & 0 \\ -\varepsilon_k & 2(\varepsilon_k + \varepsilon) & -\varepsilon_k & 0 \\ 0 & -\varepsilon & 0 & 0 \\ 0 & 0 & -\varepsilon & 0 \end{bmatrix} + \frac{1}{h_k} \begin{bmatrix} 0 & -\frac{1}{6} & \frac{1}{6} & 0 \\ -\frac{1}{6} & 0 & \frac{1}{6} & 0 \\ -\frac{1}{6} & \frac{1}{6} & 0 & 0 \end{bmatrix},\]

\[\varepsilon_k := \varepsilon + \delta_k h_k.\]

For the construction of a V- or W-cycle multigrid method for solving a system with matrix \(A_k\), we need prolongation and restriction operators and a smoother.
Approximation property

We use a smoother of the form
\[ x^{\text{new}} = x^{\text{old}} - \omega W_k^{-1}(A_k x^{\text{old}} - b) \]
with corresponding iteration matrix \( S_k := I - \omega W_k^{-1}A_k \). For \( W_k \) we take the bidiagonal matrix which corresponds to the stencil
\[ W_k = \frac{1}{b_k^2} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 4 \epsilon & 0 \\ 0 & 0 & 0 \end{bmatrix} + \frac{1}{b_k} \begin{bmatrix} 0 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \]

This choice of \( W_k \) is based on the following considerations: We use a pointwise relaxation term for the diffusion part and simple upwind differences for the convective term. An alternative (which, however, causes technical complications in the analysis) would be the use of Gauss-Seidel type smoothers with a numbering of unknowns starting from the inflow boundary and going downwind \([98, 23]\).

For the prolongation, denoted by \( p \), we choose the canonical operator induced by the embedding \( \mathcal{V}_{k-1} \subset \mathcal{V}_k \). For the restriction operator \( r \) we take the adjoint of the prolongation.

In order to prove convergence of a W-cycle multigrid method based on these prolongation, restriction, and smoothing operators, we first analyze a two-grid method with iteration matrix \( T_k = S_k^1(I - pA_{k-1}^{-1}rA_k)S_k^0 \). The parameters \( \mu, \nu \in \mathbb{N} \) correspond to the number of smoothing iterations before and after the coarse grid correction. The standard technique (see subsections 3.3.5 and 3.3.6) to bound the two-grid contraction number (with respect to a suitable norm) is to show that the following smoothing, approximation, and stability properties hold:

\[ ||A_k S_k^0||_2 \leq C_{st}(b_k, \epsilon) \eta(\nu) \quad \text{with } \eta(\nu) \to 0 \text{ for } \nu \to \infty, \]
\[ ||A_k^{-1} - pA_{k-1}^{-1}r||_2 \leq C_a(b_k, \epsilon), \]
\[ ||S_k^0||_2 \leq C_{st} \quad \forall \nu \geq 0. \]

In many cases an inequality as in (5.79) holds with a constant \( C_{st} \) independent of any parameters (often even \( C_{st} = 1 \)). The “constants” \( C_s(b, \epsilon) \) and \( C_a(b_k, \epsilon) \) in (5.77) and (5.78) should be such that their product is bounded from above independently of \( b_k \) and \( \epsilon \). If these conditions are satisfied, then for \( \nu \) and \( \mu \) large enough an inequality \( ||T_k||_2 \leq c < 1 \) holds with \( c \) independent of \( b_k \) and \( \epsilon \). However, it turns out that for the discrete convection-diffusion problem described above it is very unlikely that (5.77) and (5.78) hold with “constants” \( C_s(b, \epsilon) \) and \( C_a(b_k, \epsilon) \) such that their product in bounded independently of \( b_k \) and \( \epsilon \). Therefore, in \([157]\) another splitting was suggested,
\[ T_k = (S_k W_k^{-1}) \cdot (W_k (A_k^{-1} - p_k A_{k-1}^{-1}r_k)) \cdot (A_k S_k^0), \]
which induces the corresponding nonstandard smoothing and approximation properties. In this splitting, the pre conditioner \( W_k \) from the smoother influences the approximation property.

Approximation property

One can show that there does not exist a constant \( c \) independent of \( \epsilon \) and \( k \) such that
\[ ||W_k (A_k^{-1} - p_k A_{k-1}^{-1}r_k)||_2 \leq c. \]
To motivate the modified approximation property, let us analyze the term on the left-hand side of this inequality. The matrix $W_k$ is a sum of a diagonal matrix (scaled identity) $\frac{4\nu}{hk}I_k$ and an upwind differencing matrix, denoted by $D_x$, that corresponds to the second term in (5.76). Hence, we have

$$\|W_k(A_k^{-1} - pA_{k-1}^{-1} r)\|_2 \leq \|D_xA_k^{-1}\|_2 + \|D_x pA_{k-1}^{-1} r\|_2 + \frac{4\nu}{hk}\|A_k^{-1} - pA_{k-1}^{-1} r\|_2.$$ 

The first two terms on the right-hand side of this inequality satisfy the uniform estimates $\|D_xA_k^{-1}\|_2 \leq c$ and $\|D_x pA_{k-1}^{-1} r\|_2 \leq c$, which is a consequence of the stability of the discretization (recall from subsection 5.1.10 that the streamline-diffusion method suppresses streamwise oscillations in a discrete solution!). To bound the third term, assume that $x$ is a vector of nodal values which vanish at the nodes near the inflow boundary. Then one can prove that $\|(A_k^{-1} - pA_{k-1}^{-1} r)x\|_2 \leq c \frac{b_k}{\epsilon} \|x\|_2$ holds. It turns out that in this estimate the assumption on $x$ cannot be relaxed. Let $J_k$ be the diagonal matrix in which the diagonal entries corresponding to the nodes near the inflow boundary are zero and the other diagonal entries are equal to 1. Using this projection matrix, the final modified approximation property is as follows:

$$\|W_k(A_k^{-1} - pA_{k-1}^{-1} r)_k\|_2 \leq c$$

with $c$ independent of $\epsilon$ and $b_k$. In terms of the multigrid algorithm this result implies the condition that, before the coarse grid correction step, a residual should vanish or be sufficiently small near the inflow boundary. It turns out (cf. (5.83) below) that indeed the smoothing iterations make the residual small near the inflow boundary.

**Smoothing property**

We already pointed out that the preconditioner $W_k$ appears in the new approximation property (compare (5.78) and (5.81)). Due to this, the additional term $W_k^{-1}$ appears in the modified smoothing property (compare to (5.77)),

$$\|A_k S_k W_k^{-1}\| \leq \frac{c}{\sqrt{\nu}} \tag{5.82}$$

with a constant $c$ independent of $\epsilon$ and $b_k$.

In subsection 5.1.3 we discussed the concept of robust smoother. Such a smoother should become a direct solver if the singular perturbation parameter tends to zero. This would mean that $A_k - W_k = O(\epsilon)$. The robust smoothers are commonly based on a block (Jacobi or Gauss-Seidel) method or an ILU factorization. If the convection-diffusion problem is discretized using standard finite differences, then, as we saw in subsection 5.1.9, a Gauss-Seidel iteration with an appropriate numbering of unknowns yields a robust smoother. However, in the finite element setting such methods do not yield a robust smoother. This is clear from the stencil (5.74): For $\epsilon \rightarrow 0$ the diffusion part yields an $x$-line difference operator which can be represented exactly by an $x$-line smoother, but in the convection stencil the $[0 - \frac{1}{b} \frac{1}{\nu}]$ and $[-\frac{1}{b} \frac{1}{\nu} 0]$ parts of the difference operator are not captured by such a smoother. It is not clear how a robust smoother can be constructed for such a finite element discretization.

\[14\] The precise result requires that components of $x$ corresponding to grid nodes $(x, y)$ with $x < (3k + 1)b_k$ be zero.
5.1. Multigrid preconditioners for singular-perturbed problems

Smother is a solver near the inflow

As noted above, close to the inflow boundary the residual after smoothing should be small. The following estimate shows that the smoother near the inflow has geometric convergence with a factor that is independent of $\varepsilon$ and depends only logarithmically on $h_k$:

$$\|(I-J_k)\tilde{S}_k\|_2 \leq \left(1 - \frac{c}{k^4}\right)^{\nu}, \quad \text{with } \tilde{S}_k = A_k S_k A_k^{-1}. \quad (5.83)$$

Note that in the diagonal matrix $I-J_k$ all diagonal entries corresponding to nodes outside a “thin” strip adjacent to the inflow boundary are zero. The bound (5.83) can be compared with a global (i.e., for the whole domain) bound of the form $\|\tilde{S}_k\|_2 \leq (1 - c h^2)^{\nu}$.

A convergence result

Based on the new approximation and smoothing properties and the special property (5.83) for the smoother near the inflow boundary, one can prove a contraction number bound for the two-grid method,

$$\|T_k\|_{A^*A} \leq \frac{c_1}{\sqrt{v}} + c_2 \left(\frac{1 - c}{k^4}\right)^{\mu},$$

where $\|\cdot\|_{A^*A}$ is the norm induced by the scalar product $\langle A^*_k \cdot, A^*_k \cdot \rangle$. From this estimate one easily deduces the final multigrid convergence result.

Theorem 5.8. Assume that the number of smoothing steps on every grid level is sufficiently large: $\nu_k \geq c_{p_{0}}, \mu_k \geq c_{p_{r}} k^4$, with suitable constants $c_{p_{0}}, c_{p_{r}}$. Then for the contraction number of the multigrid $W$-cycle, the inequality $\|M_k^{\text{mgm}}\|_{A^*A} \leq \xi^*$ holds, with a constant $\xi^* < 1$ independent of $k$ and $\varepsilon$.

The dependence of $\mu_k$ on $k (k \sim \ln h_k^{-1})$ results in a suboptimal estimate (up to logarithmic factor) of the multigrid complexity: $O(n(\ln n)^4)$, where $n$ is a number of unknowns on the finest grid level. In numerical experiments, however, one does not see a need to increase the number of smoothing steps, thus preserving the optimal complexity.

5.1.12 Linear elasticity

Consider the linear elasticity problem, written in displacement variables,

$$-2\mu \text{div}D(u) - \frac{\nu}{1-2\nu} \nabla(\text{div}u) = f \quad \text{in } \Omega,$$
$$u = 0 \quad \text{on } \partial \Omega,$$

(5.84)

where $u$ is the displacement, $f$ is the body force, $D(u) = \frac{1}{2}(\nabla u + (\nabla u)^T)$ is the strain (deformation) tensor, $\nu > 0.5$ denotes Poisson’s ratio and $\mu$ is the shear modulus given by

$$\mu = \frac{E}{2(1+\nu)},$$

where $E$ is Young’s modulus. Noting that $2\text{div}D(u) = \Delta u + \nabla \text{div}u$ and performing rescaling, we arrive at the parameter-dependent problem (5.5) with $\varepsilon = \frac{\mu}{\mu + (1-2\nu)^{-1}}$. A particularly interesting case of the nearly incompressible media, when the Poisson ratio $\nu \approx 0.5$, leads to singular-perturbed elliptic problem with a small parameter $\varepsilon$. 
As we saw from the example of the convection-diffusion problem, singular-perturbed problems may require certain care in applying the finite element method. Only stable discretizations are suitable and corresponding linear algebraic systems are worth building robust iterative solvers. This is exactly the situation for the linear elasticity equation. Standard finite element methods, such as piecewise linear conforming elements, often fail or provide suboptimal accuracy; cf. [45, 48]. To realize this, consider a finite element space $\mathbb{U}_h \subset H^2_0(\Omega)$, let $\varepsilon \to 0$, and take $f = \nabla g$; then the corresponding discrete problem turns to finding $u_h \in \mathbb{U}_h$ such that

$$P_h(g - \text{div} u_h) = 0,$$

(5.85)

where $P_h$ is the $L^2$-orthogonal projector on $\text{div}(\mathbb{U}_h)$. A solution to (5.85) is not necessarily unique, and stability means that there exists $u_h \in \mathbb{U}_h$ solving (5.85), with a uniformly bounded $H^1$-norm: $||u_h||_1 \leq c \|g\|_{L^2}$, where $c$ is independent of $h$ and $g \in L^2(\Omega)$. Failure to satisfy this indicates that the space $\text{div}(\mathbb{U}_h)$ is not "large" enough in a certain sense and leads to what is known as locking phenomena.

There are two ways to avoid locking phenomena. One can try to enhance the finite element space in a way that $\text{div}(\mathbb{U}_h)$ is large enough. Examples are classical Scott–Vogelius elements [229, 155] or the recently developed “pointwise mass conservative” elements [75, 95]. Alternatively, one may ask for the equality $g - \text{div} u_h = 0$ to hold not in $\text{div}(\mathbb{U}_h)$ but in a smaller space $\mathbb{P}_h$. In the latter case, the corresponding finite element formulation reads as follows: Find $u \in \mathbb{U}_h$, satisfying

$$\varepsilon (\nabla u, \nabla v) + (P_h \text{div} u, \text{div} v_h) = (f, v_h) \quad \forall \; v_h \in \mathbb{U}_h,$$

(5.86)

where $P_h$ is the $L^2$-orthogonal projector on $\mathbb{P}_h$. The space $\mathbb{P}_h$ should be chosen carefully enough such that the solution to (5.85) is stable and at the same time $\mathbb{P}_h$ has necessary approximation properties.

It is easy to see that the solution $u_h$ of (5.86) and $p_h = \varepsilon^{-1} P_h \text{div} u_h$ also solve a finite element formulation of (5.9), where the space $\mathbb{P}_h$ is used to approximate the auxiliary variable $p = \varepsilon^{-1} \text{div} u$. In Section 5.2, we shall study preconditioners and multigrid methods for algebraic systems resulting from discretizations of (5.9) and similar problems. Here we made several comments on a multigrid method developed in [176] for solving (5.86) directly.

**Block smoothers handle kernel vectors**

For small values of parameter $\varepsilon$, standard pointwise smoothers such as damped Jacobi or Gauss–Seidel fail to eliminate the oscillatory components of the error which belong to the kernel of $P_h \text{div}$. An effective smoother should be able to handle this. For certain choices of finite elements and $P_h$, building such smoothers is feasible.

Let $\mathbb{U}_h$ be a space of piecewise quadratic conforming finite element functions with respect to a regular triangulation $\mathcal{T}_h$. Define $P_h$ to be an $L^2$-projector into the space of piecewise constant functions with respect to the same triangulation $\mathcal{T}_h$. Such a pair of spaces is commonly denoted as $P_2 - P_0$. One finds that $u_h \in \ker(P_h \text{div})$ iff for any $\tau \in \mathcal{T}_h$ the total flux for $u_h$ through $\partial \tau$ equals zero. Thus the kernel of $P_h \text{div}$ can be described explicitly by presenting its basis [48]. These basis functions are illustrated in Figure 5.2 (these are actually three of nine basis functions “captured” inside the oval on the right). Note that all basis functions have local support. Therefore an efficient smoother can be built as a block relaxation procedure, where degrees of freedom supporting each basis function from $\ker(P_h \text{div})$ are organized in one block. The blocks can be overlapping. This can be described more formally using the notions of space decomposition iteration from subsection 4.1.1.
5.1. Multigrid preconditioners for singular-perturbed problems

We recall that a space decomposition preconditioner is based on a decomposition $\mathcal{U}_h = \sum_{i=0}^{l} \mathcal{U}_i$, $\mathcal{U}_i \subset \mathcal{U}_h$, where the sum is not necessarily a direct sum. To define a suitable decomposition, let $\mathcal{T}_i$, $i = 1, \ldots, l$, be the set of all interior vertices of $\mathcal{G}_h$. For each $i$ define $\tau^i$ as a union of all triangles $\tau \in \mathcal{T}_h$ sharing $\mathcal{T}_i$. Then

$$\mathcal{U}_i := \{ v_h \in \mathcal{U}_h : v_h = 0 \quad \text{in} \quad \Omega \setminus \tau^i \}.$$ 

One smoothing step is then a space decomposition iteration as defined in (4.5).

In [176] it was shown that these smoothing iterations are suitable, since the smoothing property holds in the parameter-dependent norm

$$\|v\|_\epsilon = (\epsilon (\nabla v, \nabla v) + (P_h \operatorname{div} v, \operatorname{div} v))^{1/2}.$$ 

This norm can be also seen as the problem specific “energy” norm on $H^1_0$. This smoothing property has to be complemented with an appropriate approximation property. The key ingredient for such an approximation result is discussed below.

**Prolongation should respect kernel vectors**

Consider the standard prolongation operator $p : \mathcal{U}_H \rightarrow \mathcal{U}_h$ induced by the embedding $\mathcal{U}_H \subset \mathcal{U}_h$. This is a common choice in a multigrid method for solving elliptic problems. However, it appears not to work well for (5.5) in the case of small $\epsilon$. The reason is that $v_H \in \ker (P_H \operatorname{div})$ does not yield in general $p v_H \in \ker (P_h \operatorname{div})$. Therefore the standard prolongation is not uniformly stable in the energy norm; i.e., the constant $C$ in the estimate

$$\|p v_H\|_\epsilon \leq C \|v_H\|_\epsilon \; \forall \; v_H \in \mathcal{U}_H$$

(5.87)

could grow unboundedly for $\epsilon \rightarrow 0$. In [176] it is suggested that to enforce the stability of the prolongation one can make a correction of $v_H$ on a fine grid in a way involving solution of local problems.

Assume the fine grid triangulation $\mathcal{T}_h$ is obtained from the coarse grid triangulation $\mathcal{T}_H$ by one step of regular refining, and define a subspace of $\mathcal{U}_h$ of all functions that vanish in the coarse grid nodes:

$$\mathcal{U}_h = \{ v_h \in \mathcal{U}_h : v_h|_{\partial \tau} = 0 \quad \forall \; \tau \in \mathcal{T}_h \}.$$ 

Define the bilinear form on $\mathcal{U}_h \times \mathcal{U}_h$,

$$a_h(u_h, v_h) := \epsilon (\nabla u_h, \nabla v_h) + (P_h \operatorname{div} u_h, \operatorname{div} v_h).$$
For some given coarse grid finite element function \( u_H \in \mathbb{U}_H \), let \( \tilde{u}_h \) be a solution of the problem

\[
\mathcal{A}_h(\tilde{u}_h, \tilde{v}_h) = \mathcal{A}_h(u_H, \tilde{v}_h) \quad \forall \tilde{v}_h \in \mathbb{U}_h.
\] (5.88)

Define the prolongation \( p : \mathbb{U}_H \to \mathbb{U}_h \) by

\[
p u_H = u_H - \tilde{u}_h.
\] (5.89)

Prolongation \( p \) is stable, i.e., (5.87) holds with a constant \( C \) independent of \( h \) and \( \varepsilon \); see [19].

Obviously the subspace \( \tilde{\mathbb{U}}_h \) is the direct sum of small subspaces corresponding to each coarse grid triangle. Hence, solving (5.88) requires the solution of \( N_H \) independent problems of small dimension (for regular subdivisions the dimension is 6 in two dimensions and 9 in three dimensions), where \( N_H \) is the number of elements in \( \mathcal{G}_H \).

Results in [176, 19] show that geometric multigrid V- and W-cycles, with standard restriction and coarse grid operators, block smoothing iterations, and modified prolongation as explained above are robust with respect to \( \varepsilon \). A deeper analysis of this and other multigrid methods for the linear elasticity problem relies on results for indefinite (saddle point) problem, which are studied in the next section.

We note that if \( P_h \) is the projection on \( \text{div}(\mathbb{U}_h) \), then one may set \( P_h = I \) in (5.86) and the definition of the energy norm. In this case, the canonical prolongation based on the embedding \( \mathbb{U}_H \subset \mathbb{U}_h \) is obviously stable in the sense of (5.87). However, letting \( P_h \) to be the projection on \( \text{div}(\mathbb{U}_h) \) is suitable for conforming finite elements, which are polynomials of degree \( k \geq 2d \) on a triangulation of \( \Omega \subset \mathbb{R}^d \), \( d = 2, 3 \), but not for lower order elements. If the triangulation is obtained by a barycenter refinement of a coarser mesh, then the order of polynomials can be reduced to \( k \geq d \). This finite element is known as the Scott–Vogelius element. The stability of canonical prolongation and robust smoothing properties of the vertex oriented block iterations, similar to those described above, is exploited in [129] to prove the robustness of a simple multigrid cycle for the linear elasticity problem discretized with the Scott–Vogelius elements.

**Exercises**

5.1.1. (See [9].) Find the ILU(0) decomposition from (5.27) for the matrix

\[
A = \begin{bmatrix}
4 & -1 & -1 & 0 \\
1 & -2 & 0 & -1 \\
-1 & 0 & -2 & -1 \\
0 & -1 & -1 & 4
\end{bmatrix}.
\]

Check that the remainder \( N \) from (5.27) is

\[
N = \begin{bmatrix}
0 & 0 & 0 & 0 \\
0 & 0 & \frac{1}{4} & 0 \\
0 & \frac{1}{4} & 0 & 0 \\
0 & 0 & 0 & 0
\end{bmatrix}.
\]
5.2. Preconditioners for certain problems of fluid mechanics

In this section, we discuss preconditioners for systems of algebraic equations resulting from discretizations of equations of fluid mechanics. As a prototypical problem, we consider the system of the Navier–Stokes equations, which models a flow of incompressible viscous fluid:

\[
\frac{\partial \mathbf{u}}{\partial t} - \nu \Delta \mathbf{u} + (\mathbf{u} \cdot \nabla) \mathbf{u} + \nabla p = \mathbf{f} \quad \text{in} \quad \Omega \times (0, T),
\]

\[
-\nabla \cdot \mathbf{u} = 0 \quad \text{in} \quad \Omega \times [0, T). \tag{5.90, 5.91}
\]

Here \( \mathbf{u}(t, x) = (u_1(t, x), u_2(t, x), u_3(t, x)) \) is a vector function that has the physical meaning of fluid velocity in a point of space \( x = (x, y, z) \in \Omega \subset \mathbb{R}^3 \) (coordinates are given with...
The Stokes problem

respect to an inertial frame of reference) and time \( t \in [0, T] \). Other notations are

\[
\Delta u = \begin{cases} 
\Delta u_1, \\
\Delta u_2, \\
\Delta u_3,
\end{cases} \quad (u \cdot \nabla)u = \begin{cases} 
\nu \frac{\partial u_1}{\partial x} + u_2 \frac{\partial u_1}{\partial y} + u_3 \frac{\partial u_1}{\partial z}, \\
\nu \frac{\partial u_2}{\partial x} + u_2 \frac{\partial u_2}{\partial y} + u_3 \frac{\partial u_2}{\partial z}, \\
\nu \frac{\partial u_3}{\partial x} + u_2 \frac{\partial u_3}{\partial y} + u_3 \frac{\partial u_3}{\partial z};
\end{cases}
\]

\( p(t, x) \) is the scalar function that has the physical meaning of the kinematic pressure of fluid; and \( \nu > 0 \) is the given viscosity coefficient. Of course, (5.90)–(5.91) should be supplemented with initial and boundary conditions. For the purpose of analysis and for building algebraic solvers, it is common to impose homogeneous boundary conditions for velocity. If no boundary values are given for the pressure, the function \( p \) is defined up to a constant. For the uniqueness, it is sufficient to assume that pressure has zero integral mean. Thus, we supplement (5.90)–(5.91) with

\[
u\|u\|_{L^2}^2 + \int_{\Omega} |p| \, dx = 0. \tag{5.92}
\]

The derivation of equations and a thorough discussion of underlying physics can be found in many texts on fluid mechanics; see, for example, [124] and [78].

5.2.1 The Stokes problem

The problem (5.90)–(5.91) is the system of differential equations. We shall study algebraic solvers for such systems starting with the greatly simplified problem of the system of stationary Stokes equations:

\[
\begin{align*}
-\Delta u + \nabla p &= f \quad \text{in } \Omega, \\
-\text{div}\, u &= g \quad \text{in } \Omega, \\
|u|_{\partial \Omega} &= 0.
\end{align*}
\tag{5.93}
\]

Moreover, for the sake of clarity, we consider only the two-dimensional case, \( \Omega \subset \mathbb{R}^2 \), \( u = (u_1, u_2) \).

A solution to (5.93)–(5.95) is sought in \( H^1_0(\Omega) := (H^1_0(\Omega))^2 \) for \( u \) and \( L^2(\Omega) \) for \( p \). It is convenient to use the following space for the pressure:

\[
L^2_0 := \left\{ p \in L^2(\Omega) : \int_{\Omega} p \, dx = 0 \right\}.
\]

It is widely accepted that the solution of the Stokes problem is a good approximation to the Navier–Stokes solution in the case of large values of viscosity \( \nu \) and for low velocities \( u \), so-called creeping flows. Besides this, the Stokes problem often serves as an auxiliary one in numerical algorithms for the full system of the Navier–Stokes equations. In this case, the function \( g \) in the right-hand side of (5.94) can be nonzero. Integrating (5.94) over \( \Omega \) and applying the Stokes formula yields the necessary condition on \( g \) for the consistency of the system: \( \int_{\Omega} g \, dx = 0 \). We note that for \( g = 0 \), the Stokes problem is the limit case of \( \varepsilon \to 0 \) for the alternative form (5.9) of the linear elasticity equations. Much of the discussion below also carries over to the system (5.9) with \( \varepsilon > 0 \). We shall make corresponding remarks when they are appropriate. One can consider \( \nu = 1 \) in (5.93) without any loss of generality, since it is always possible to multiply the equations (5.93) with \( \nu^{-1} \) and further rescale the pressure and the right-hand side, \( p_{\text{new}} := \nu^{-1} p, f_{\text{new}} := \nu^{-1} f \).
5.2. Preconditioners for certain problems of fluid mechanics

To deduce a weak formulation of the problem, we multiply two equations in \(5.93\) by arbitrary \(v_1 \in H_0^1(\Omega)\) and \(v_2 \in H_0^1(\Omega)\), integrate the resulting equalities over \(\Omega\), apply integration by parts, and sum up these two equalities. We also multiply the equation \((5.94)\) by \(q \in L_0^2(\Omega)\) and integrate over \(\Omega\). This leads to the following weak formulation of the Stokes problem: Find a pair of functions \(u \in H_0^1(\Omega)\) and \(p \in L_0^2(\Omega)\) satisfying

\[
(\nabla u, \nabla v) - (p, \text{div} v) = (f, v) \quad \forall v \in H_0^1(\Omega),
\]

\[
-(\text{div} u, q) = (g, q) \quad \forall q \in L_0^2(\Omega),
\]

where

\[
(\nabla u, \nabla v) = \int_\Omega \left( \sum_{i=1}^2 \frac{\partial u_1}{\partial x_i} \frac{\partial v_1}{\partial x_i} + \frac{\partial u_2}{\partial x_i} \frac{\partial v_2}{\partial x_i} \right) dx.
\]

Sometimes it is more convenient to use the equivalent definition of the weak solution as a pair \(\{u, p\} \in H_0^1(\Omega) \times L_0^2(\Omega)\) satisfying

\[a(u, p; v, q) = (f, v) - (g, q) \quad \forall \{v, q\} \in H_0^1(\Omega) \times L_0^2(\Omega),\]

with nonnegative definite, nonsymmetric bilinear form

\[a(u, p; v, q) := (\nabla u, \nabla v) - (p, \text{div} v) + (\text{div} u, q).
\]

We continue to use the notation

\[||u||_1 := (\nabla u, \nabla u)^{1/2}\]

for a norm on \(H_0^1(\Omega)\) that is equivalent to the \(H^1\)-norm.

One can also formulate the Stokes problem as finding a saddle point of the Lagrangian

\[\mathcal{L}(u, p) = ||u||_1^2 + 2(\text{div} u - g, p) - 2(f, u).
\]

Indeed, the solution to \((5.96)-(5.97)\) satisfies

\[\mathcal{L}(u, q) \leq \mathcal{L}(u, p) \leq \mathcal{L}(v, p) \quad \forall v \in H_0^1(\Omega), q \in L_0^2.
\]

Thanks to this optimization property, the linear algebraic systems resulting from the constrained equations \((5.93)-(5.95)\) are often referred to as saddle point problems.

In the particular case of \(g = 0\), an a priori estimate for \(u\) readily follows from \((5.96)\) by setting \(v = u\) and using \((5.97)\) with \(q = p\):

\[||u||_1 = (f, u) \leq ||f|| ||u|| \leq C_f ||f|| ||u||_1.
\]

Therefore, \(||u||_1 \leq C_f ||f||\).

An estimate for pressure is possible thanks to the following inequality known as the Necas or LBB (Ladyzhenskaya–Babushka–Brezzi) inequality \([36]\):

\[c_0 ||q|| \leq ||\nabla q||_{L^2} \quad \forall q \in L_0^2,
\]

with a positive constant \(c_0\). Using the definition of \(\nabla q\) as a functional on \(H_0^1(\Omega)^2\), the Necas inequality can be written as

\[
\sup_{v \in H_0^1(\Omega)^2} \frac{(q, \text{div} v)}{||v||_1} \geq c_0 ||q|| \quad \forall q \in L_0^2.
\]
Now the pressure estimate follows from (5.96):

$$c_0 \|p\| \leq \sup_v \left( \frac{(p, \nabla v)}{\|v\|_1} + \frac{\|\nabla u - (f, v)\|}{\|v\|_1} \right) \leq \|u\|_1 + C_\Omega \|f\| \leq 2C_\Omega \|f\|.$$  

If $g \neq 0$, an estimate for the solution of the Stokes problem can be found by similar arguments and takes the form:

$$\|u\|_1 + \|p\| \leq c (\|f\| + \|g\|).$$

For the existence and uniqueness of the solution to (5.96)–(5.97), it is sufficient to consider $f \in L^2(\Omega)^2$, $g \in L^2(\Omega)$.

Let finite element spaces $U_h$ and $P_h$ be given to approximate $u = (u_1, u_2)$ and $p$, respectively. The space $U_h$ is the direct sum of two finite element spaces (three in three dimensions); each of these two approximates $H^1(\Omega)$. The finite element method is based on the weak formulation (5.96)–(5.97). Finite element solution $\{u_h, p_h\}$ is sought to satisfy

$$\begin{align*}
(\nabla u_h, \nabla v_h) - (p_h, \nabla v_h) &= (f, v_h) \quad \forall v_h \in U_h, \\
-(\nabla u_h, q_h) &= (g, q_h) \quad \forall q_h \in P_h.
\end{align*}$$  

To ensure the existence and the uniqueness of the solution $\{u_h, p_h\}$, the spaces $U_h$ and $P_h$ need to satisfy a certain compatibility condition, which is known as the LBB condition: There exists a constant $\mu(\Omega) > 0$ independent of $h$ such that

$$\mu(\Omega) \leq \inf_{q_h \in P_h} \sup_{v_h \in U_h} \frac{(q_h, \nabla v_h)}{\|q_h\|_{\|v_h\|_1}}.$$  

This condition can be interpreted as the discrete counterpart of the Næset inequality.

The LBB condition does not necessarily hold for any pair of finite element spaces $U_h$ and $P_h$. Failure of (5.101) may cause instability. Thus, if the right-hand side of (5.101) equals zero, then it may happen that there is no solution $u_h$ and $p_h$ to (5.99)–(5.100); see Exercise 5.2.1. If the infsup term in the right-hand side of (5.101) is positive but tends to zero for $h \to 0$, then a unique solution to (5.99)–(5.100) exists but may fail to converge to the smooth solution of the differential problem (5.93)–(5.95).

Inequality (5.101) may fail to be true for some simple and naive choices of $U_h$ and $P_h$. For example, the condition (5.101) is not satisfied if $U_h$ consists of piecewise linear continuous functions and $P_h$ consists of piecewise linear continuous or piecewise constant functions. The corresponding finite element pairs are denoted by $P_1 - P_1$ and $P_1 - P_0$. Two examples of finite element pairs satisfying (5.101) are given by piecewise linear (or constant) pressure $p_h$ and piecewise linear continuous velocity $u_h$ on a refined triangulation, as shown in Figure 5.3. The standard notations for these elements are $isoP_2 - P_1$ and $isoP_2 - P_0$ finite element pairs. Another example is the family of Taylor–Hood elements $P_{k+1} - P_k$, $k \geq 1$ (both velocity and pressure are continuous, and on each triangle $u_h$ and $p_h$ are polynomials of degree $k + 1$ and $k$, respectively). One more example of a pair satisfying (5.101) is the nonconforming $\bar{P}_1 - P_0$ Crouzeix–Raviart element (piecewise constant pressure $p_h$ and piecewise linear velocity $u_h$ on the same triangulation; here $u_h$ is not globally continuous). Since for the nonconforming elements the discrete velocity $u_h$ can be discontinuous across the edges of the triangulation, the bilinear forms in (5.99)–(5.100) should be defined elementwise. The monograph [48] discusses in detail why the LBB condition is important and how it can be verified for a given elements, and considers many other stable pairs $\{U_h, P_h\}$. 

5.2. Preconditioners for certain problems of fluid mechanics

Similarly to the analysis of multigrids for elliptic problems, the multigrid analysis for the Stokes problem requires error estimates for finite element solutions. With the help of Galerkin orthogonality and the LBB condition, one shows the following error estimates:

\[
\|u - u_h\|_1 + \|p - p_h\| \leq C \left( \inf_{v_h \in \mathcal{U}_h} \|u - v_h\|_1 + \inf_{q_h \in \mathcal{P}_h} \|p - q_h\| \right).
\] (5.102)

Using interpolation properties of \( \mathcal{U}_h \) and \( \mathcal{P}_h \) we get

\[
\|u - u_h\|_1 + \|p - p_h\| \leq c h (\|u\|_{H^1} + \|p\|_{H^1}).
\] (5.103)

With additional assumptions on \( \Omega \) and \( g \), the following estimates are valid:

\[
\|u - u_h\| \leq c h \|u - u_h\|_1,
\] (5.104)

\[
\|u\|_{H^1} + \|p\|_{H^1} \leq c (\|f\| + \|g\|_{H^1}).
\] (5.105)

Estimates (5.102)–(5.105) are fairly standard. Proofs of (5.102)–(5.105) are given in many references. One of the best texts on finite element method for the Navier–Stokes problem is the monograph [83]. The estimate (5.105) for the case of convex domains with piecewise smooth boundary were proved in [62]. We note that to ensure the inequality (5.105) is valid in a convex polyhedral domain, in addition to the assumption \( g \in H^1 \) one needs assumptions on the behavior of \( g \) in the neighborhoods of angle points (and edges for three-dimensional domains).

Inequalities (5.103)–(5.105) serve as a basis for checking an approximation property in the following \( h \)-dependent norm:

\[
\|\|u_h, p_h\|\| := (\|u_h\|^2 + h^2 \|p\|_{L^2}^2)^{1/2}.
\]

Indeed, (5.103)–(5.105) imply, for the difference between fine and coarse grid solutions, the estimate

\[
\|\|u_h - u_{hH}, p_h - p_{hH}\|\| \leq c h^2 (\|f\| + \|g\|_1).
\] (5.106)

In the next subsection, we consider a system of linear algebraic equations resulting from (5.99)–(5.100). We will use (5.106) to write an approximation property in the matrix form.

5.2.2 The algebraic Stokes system

To write the finite element problem (5.99)–(5.100) in the form of an algebraic system, we enumerate all basis functions from \( \mathcal{U}_h \): \( \psi_i \), \( i = 1, \ldots, n \). It is usually convenient to enumerate first basis functions for the first velocity component and then for the second.
Finally, we enumerate all pressure basis functions $\mathbb{P}_h$: $\phi_i$, $i = 1, \ldots, m$. Equalities (5.99)–(5.100) lead to the algebraic system of the form

$$
\begin{bmatrix}
A & B^T \\
B & 0
\end{bmatrix}
\begin{bmatrix}
u \\
p
\end{bmatrix}
= \begin{bmatrix}
f \\
g
\end{bmatrix}
\iff A\mathbf{x} = \mathbf{b},
$$

(5.107)

where $u, p$ are the vectors of coefficients of functions from $U_h$ and $\mathbb{P}_h$. The coefficients of matrices $A$ and $B$ are defined through the identities

$$
\langle \hat{\mathbf{e}}_i, \hat{\mathbf{e}}_j \rangle = (\nabla \psi_i, \nabla \psi_j), \quad i, j = 1, \ldots, n,
$$

$$
\langle B\hat{\mathbf{e}}_i, \hat{\mathbf{e}}_j \rangle = (\text{div} \psi_i, \phi_j), \quad i = 1, \ldots, n, \quad j = 1, \ldots, m,
$$

where $\hat{\mathbf{e}}_i$ is the unit vector, with the $i$th coordinate equal to one and all other coordinates equal to zero. Matrix $A$ is symmetric and positive definite. If the matrix $A\mathbf{x}$ in (5.107) is nonsingular, then it follows that $m \leq n$ and $B$ has full rank. The system (5.107) inherits the optimization property (5.98) to the Stokes problem, and its solution is a saddle point of the Lagrangian

$$
\mathcal{L}(u, p) = \langle Au, u \rangle + 2\langle Bu - g, p \rangle - 2\langle \hat{f}, u \rangle.
$$

Linear algebraic systems as in (5.107) are often called algebraic saddle point problems. Since such systems are ubiquitous in applications, they enjoyed extensive treatment in the literature; see the review [17].

By $M$ we denote the mass matrix for the pressure space $\mathbb{P}_h$,

$$
\langle M\hat{\mathbf{e}}_i, \hat{\mathbf{e}}_j \rangle = (\phi_i, \phi_j), \quad i, j = 1, \ldots, m.
$$

(5.108)

Since $A > 0$ and $A^{-1}$ exists, one can perform the block Gauss elimination procedure and arrive at the following equations:

$$
u + A^{-1}B^T p = A^{-1}f,
$$

(5.109)

$$
BA^{-1}B^T p = BA^{-1}f - g.
$$

(5.110)

In (5.110) the velocity vector $\mathbf{u}$ is eliminated from the system. The matrix $S = BA^{-1}B^T$ on the left-hand side is called the (pressure) Schur complement for the system (5.107). This matrix plays an important role for the analysis of saddle point problems. Note that, first, the matrix $S$, in contrast to the matrices $A$ and $B$, is not a sparse matrix—most entries of $S$ are nonzero, since most entries of $A^{-1}$ are nonzero. Second, computing all entries of $S$ is pretty expensive. In practice, it is very rare that the matrix $S$ is found explicitly. In the next lemma, we prove some useful properties of $S$ that can be verified, avoiding the computation of the entries of $S$.

**Lemma 5.9.** Assume that the LBB condition (5.101) holds for a pair of conforming finite element spaces $U_h$ and $\mathbb{P}_h$ with a constant $\mu(\Omega)$. Then the Schur complement $S = BA^{-1}B^T$ is the symmetric positive definite matrix such that

$$
\mu^2(\Omega) M \leq S \leq M,
$$

(5.111)

where $\mu(\Omega)$ is a constant from the LBB condition (5.101).
5.2. Preconditioners for certain problems of fluid mechanics

**Proof.** For arbitrary vectors $q \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$, consider the corresponding finite element functions $q_h = \sum_{i=1}^m q_i \phi_i$ and $v_h = \sum_{i=1}^n v_i \psi_i$. Then it holds that

$$||q_h||^2 = \langle M q, q \rangle, \quad ||v_h||_1 = \langle A v, v \rangle, \quad (q_h, \text{div} \ v_h) = \langle q, B v \rangle, \quad (5.112)$$

$$\sup_{v_h \in U_h} \frac{(q_h, \text{div} v_h)}{||v_h||_1} = \sup_{v \in \mathbb{R}^n} \frac{\langle q, B v \rangle}{\langle A v, v \rangle}. \quad (5.113)$$

As before, if $x$ appears in a denominator, then $\sup_x$ is computed over $x \neq 0$. One easily verifies that the following equalities are valid:

$$\sup_{v \in \mathbb{R}^n} \frac{(q, B v)^2}{\langle A v, v \rangle} = \sup_{v \in \mathbb{R}^n} \frac{(B^T q, v)^2}{\langle A v, v \rangle} = \sup_{v \in \mathbb{R}^n} \frac{\langle A^{-1} B^T q, A v \rangle^2}{\langle A v, v \rangle} = \sup_{v \in \mathbb{R}^n} \frac{\langle A^{-1} B^T q, v \rangle_A^2}{||v||_A^2} = ||A^{-1} B^T q||_A^2 = \langle A^{-1} B^T q, A^{-1} B^T q \rangle_A$$

$$= \langle B^T q, A^{-1} B^T q \rangle = \langle S q, q \rangle.$$

Here we used that for a scalar product and the corresponding norm it holds that $||b|| = \sup_{v \in \mathbb{R}^n} \frac{\langle v, b \rangle}{||v||}$ for any $b \in \mathbb{R}^n$.

The equality $||\nabla v||^2 = ||\text{div} v||^2 + ||\text{rot} v||^2$ holds for any $v \in H^1_0(\Omega)$ (Exercise 5.2.2). This implies $||\text{div} v|| \leq ||\nabla v|| = ||v||_1$. Thus for arbitrary function $q_h \in P_h$ it holds that

$$\sup_{v_h \in U_h} \frac{(q_h, \text{div} v_h)}{||v_h||_1} \leq \sup_{v \in \mathbb{R}^n} \frac{||q_h||}{||\text{div} v_h||} \leq ||q_h||.$$

(5.114)

Now the bounds (5.101) and (5.114) and relations (5.112) imply

$$\rho^2(\Omega)(M q, q) \leq \langle S q, q \rangle \leq (M q, q),$$

which is (5.111) by the definition.

Finally, since $A$ is symmetric, $A^{-1}$ is symmetric, and so is the matrix $S$. $\square$

The block Gauss elimination procedure is equivalent to the block LU decomposition:

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} = \begin{bmatrix} I & O \\ B A^{-1} & I \end{bmatrix} \begin{bmatrix} A & B^T \\ O & S \end{bmatrix}. \quad (5.115)$$

There is a variety of preconditioners for discrete Stokes problem built upon this or similar block factorizations. Most of them resort to building preconditioners for each block, $A$ and $S$. We discuss this approach later in this section. First, we consider some direct (also known as “coupled”) multigrid methods for the discrete Stokes problem.

### 5.2.3 Some useful operator norms

Consider the fine grid matrix

$$\mathcal{A}_h = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix}. \quad (5.116)$$

Similarly, one defines the coarse grid matrix $\mathcal{A}_h$.

We will need one simple inequality. Assume we are given two norms $|| \cdot ||_a$ and $|| \cdot ||_b$ on $\mathbb{R}^n$. Then for arbitrary matrices $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{n \times n}$, it holds that

$$||\mathcal{A} \mathcal{B}||_a \leq ||\mathcal{A}||_{\beta \to a} ||\mathcal{B}||_{a \to \beta}, \quad (5.117)$$
where

$$
\|A\|_{\beta \rightarrow \alpha} = \sup_{x \in \mathbb{R}^n} \frac{\|Ax\|_{\alpha}}{\|x\|_{\beta}}, \quad \|B\|_{\alpha \rightarrow \beta} = \sup_{x \in \mathbb{R}^n} \frac{\|Bx\|_{\beta}}{\|x\|_{\alpha}}.
$$

Let $\mathcal{S}_h$ be the iteration matrix of a smoothing iteration, and consider the two-grid iteration matrix (see subsection 3.2.5)

$$
\mathcal{M} = (I - p \cdot A^{-1} \cdot \mathcal{S}_h) \mathcal{S}_h^\nu.
$$

To prove the convergence of the two-grid method in a norm $\|\cdot\|_{\alpha}$, it is sufficient, due to (5.117), to show that

$$
\|\mathcal{M}\|_{\alpha} \leq \|A^{-1} - p \cdot A^{-1} \cdot \mathcal{S}_h \cdot \mathcal{S}_h^\nu\|_{\alpha \rightarrow \beta} < 1,
$$

with any norm $\|\cdot\|_{\beta}$. Therefore, it is sometimes convenient to look for the approximation property in the norm $\|\cdot\|_{\beta \rightarrow \alpha}$ and for the smoothing property in the norm $\|\cdot\|_{\alpha \rightarrow \beta}$. We recall that the fulfillment of this properties is sufficient for the proof of the convergence of a multigrid W-cycle. In this and previous chapters, the norms $\|\cdot\|_{\alpha}$ and $\|\cdot\|_{\beta}$ were assumed to be the same, but they can be chosen to be different.

For the analysis of multigrids for the system (5.107), we need the following norms:

$$
\|u, p\|_{\alpha} = (\|u\|^2 + b^2 \|p\|^2)^{\frac{1}{2}}, \quad \|u, p\|_{\beta} = (\|u\|^2 + b^{-2} \|p\|^2)^{\frac{1}{2}}.
$$

We continue with presenting several smoothing strategies.

Further in the section we discuss smoothing iterations. Smoothing iterations for the Stokes problem found in the literature can be divided into two classes: block iterations and distributive iterations. We consider the most well-known methods from both classes. We warn the reader that the theory of iterative methods for the systems of the type in (5.116) is more complicated than in the symmetric positive definite case. In particular, the smoothing property has not been proven for all methods successfully used in practice. Yet, on the theoretical side, we shall look for the smoothing property of the form

$$
\|A^{-1} - p \cdot A^{-1} \cdot \mathcal{S}_h \cdot \mathcal{S}_h^\nu\|_{\alpha \rightarrow \beta} \leq \eta(\nu)b^{-2}, \quad \eta(\nu) \to 0 \quad \text{for} \ \nu \to \infty. \quad (5.118)
$$

### 5.2.4 The Vanka smoother for the Stokes system

We start with one of the most popular smoothing strategies. The method is related to block iterations and was proposed at the late 1980s in [209]. It is known in the literature as the Vanka or SCGS (symmetric coupled Gauss–Seidel) method.

The Gauss–Seidel method cannot be directly applied to (5.107) because of the zero right bottom block. Thus, one way of applying the Gauss–Seidel method consists of regrouping unknowns in such a way that a block method can be applied with respect to these blocks. A natural way of enumeration and regrouping of unknowns is based on a division of the triangulation $\mathcal{T}_h$ on macroelements.

We shall consider the SCGS iterations for the example of an $isoP_2 - P_0$ finite element pair. We start with enumerating all pressure elements of the triangulation (large triangles): $\tau_i, i = 1, \ldots, I$. Denote by $u_i, p_i$ the sets of coefficients (nodal values) of pressure and velocity defined on the element $\tau_i$. For this particular example, every set contains 13 coefficients: 6 values come from each velocity component of $u$ and 1 value comes from $p$. Note that an entry of vector $u$ can belong to several sets $u_i$ at the same time.
Let $A_i, B_i$ be matrices of $(12 \times 12)$ and $(1 \times 12)$ dimensions. These matrices are composed from elements of $A$ and $B$, which correspond to the $i$th set of coefficients. The matrices $A_i$ and $B_i$ contain entries standing on the cross of rows and columns, corresponding to the indexes of unknowns from the $i$th set. In the same fashion as $u_i, p_i$, we define sets $r_i, q_i$ for the residual vector $r = \begin{bmatrix} r & q \end{bmatrix}$. The SCGS smoothing iterations are defined as follows. For $i = 1, \ldots, I$ do

$$
\begin{bmatrix}
  u_i \\
  p_i
\end{bmatrix}^{\text{new}} = \begin{bmatrix}
  u_i \\
  p_i
\end{bmatrix}^{\text{old}} - \omega \sum_{i \in I} \begin{bmatrix}
  A_i & B_i^T \\
  B_i & 0
\end{bmatrix}^{-1} \begin{bmatrix}
  r_i \\
  q_i
\end{bmatrix},
$$

(5.119)

Calculations in (5.119) are performed sequentially for different values of $i$. Here new values $u_i, q_i$ are used to compute the new residual $r_{i+1}, q_{i+1}$ and so on, and $\omega$ is the relaxation parameter. One can put $\tilde{A}_i = A_i$; another popular choice is $\tilde{A}_i = \text{diag}(A_i)$. In the latter case, the inverted matrix in (5.119) has the form shown in Figure 5.4. To solve systems with matrices $\tilde{A}_i$, the Gauss elimination is a good choice because of the small size of these matrices.

Nowadays, a multigrid method with the SCGS smoother is generally believed to be one of the best ways solve systems of Stokes type. Besides partial results in [139], a proof of its mesh-independent convergence is still largely missing even for the simplest case of the Stokes problem. A disadvantage of the method is a significant degradation of its convergence on grids with a strong anisotropy. To fix this shortcoming, further modification of the method is required; see [191].

Of course, the SCGS iteration can be defined in terms of a space decomposition method if the spaces $U_h$ and $P_h$ are decomposed into subspaces $U_i$ and $P_i$, where the latter subspaces span over all basis functions with nonzero support over a macroelement. However, the analysis of section 4.1 does not apply, since the matrix of the system is not sign definite.

We considered a sequential method to define the Vanka smoother. Naturally, one may also try a parallel method. This corresponds to Jacobi type block iterations (sequential correction Jacobi or SCJ), where a smoothing procedure coincides with (5.119); with the only difference being that the iterations (5.119) are performed now for all $i = 1, \ldots, I$ (i.e., iterations are performed for all sets of unknowns) independently:
A smoothing property of the SCJ iterations for the Stokes problem was proved in [177] under the assumption that $\sum_{\epsilon \ell} P_{\ell} P_{\ell}^T = 1$. However, the SCJ method is less competitive than the SCGS method, at least on computers with a sequential architecture, and may not converge unless the number of smoothing steps is large enough. Numerical experiments with the Vanka multigrid method can be found in many references; see, for example, [198].

### 5.2.5 Distributive iterations as smoothers

Let us write the system (5.107) in the abstract form $A x = b$. A general idea of the method of distributive iterations is to find a matrix $B$ such that the matrix $A B$ admits the construction of efficient smoothing iterations of the form

$$y^{m+1} = y^m - \mathcal{S}^{-1}(A B y^m - b), \quad x = B y.$$  \hspace{1cm} (5.120)

One of the simplest ways to define $B$ is to set $B = A^T$ (for the Stokes problem it holds that $A = A^T$). For this choice, $A B = A A^T$ is a symmetric positive definite matrix. Consider another simple choice: Introduce the diagonal matrix $D$ of the form

$$D = \begin{bmatrix} I & 0 \\ 0 & b^2 I \end{bmatrix}$$

and set $B = D^{-1} A D^T$, $C = D$. For such a choice of $B$ and $C$, the Richardson iteration was considered as the smoothing method in [181] and [211]. In [211] the desired smoothing property was proved (although formulated in a form different from (5.121)):

$$\| A (I - w_h^2 D^{-1} A D^{-1} A) y \|_{2 \rightarrow \beta} \leq \frac{c b^{-2}}{\sqrt{2\nu + 1}} \text{ with } w_h = O(h^2). \hspace{1cm} (5.121)$$

To verify (5.121), we rewrite the estimate in the spectral matrix norm by substituting $x = D^{-1} y$ and $A = D^{-1} A D^{-T}$:

$$\| A (I - w_h^2 D^{-1} A D^{-T} A) y \|_{2 \rightarrow \beta} \leq \frac{c b^{-2}}{\sqrt{2\nu + 1}} \hspace{1cm} (5.122)$$

The matrix $A$ is symmetric. Using the result of Exercise 5.2.3 and $|\rho(A)| \leq \|A\|$, one easily finds that $\rho(A) = c b^{-2}$. Let $w_h^{-1} := \rho(A)$; then $\sp(w_h A) \in [-1, 1]$, and the estimate (5.122) follows from

$$\| A (I - w_h^2 A) y \|_{2 \rightarrow \beta} \leq w_h^{-1} \max_{x \in [-1, 1]} |x(1 - x^2)| \leq \frac{c b^{-2}}{\sqrt{2\nu + 1}}.$$  \hspace{1cm} (5.123)

Another example is the distributive Gauss-Seidel smoothing method of Brandt and Dinar formulated in the form (5.120) in [96]. In this method, the matrix $B$ is given by

$$B = \begin{bmatrix} I & B^T \\ 0 & BB^T \end{bmatrix}.$$  \hspace{1cm}

This yields

$$A B = \begin{bmatrix} A & A B^T + B^T (BB^T) \\ B & BB^T \end{bmatrix}. \hspace{1cm} (5.123)$$
Now the matrix $\mathcal{A}/\mathcal{B}$ is sign-definite and the inconvenient zero (2,2)-block disappears. The matrix $BB^T$ mimics a discretization of the Poisson problem for pressure. Assume that $\hat{A}$ and $R$ are preconditioners for $A$ and $BB^T$, respectively. These preconditioners can be built with the help of ILU decomposition. The block preconditioner $C$ for system (5.123) in the method of Brandt and Dinar is defined as

$$C = \begin{bmatrix} \hat{A} & 0 \\ B & R \end{bmatrix}.$$  

In the next example we consider one of the most popular algorithms in engineering software, the SIMPLE (semi-implicit method for pressure-linked equations) method, first introduced in [163]: For a given $u^{old}$ and $p^{old}$ and a relaxation parameter $\alpha > 0$ compute the next iterate $u^{new}$ and $p^{new}$ in several steps.

1. Find an auxiliary function $\tilde{u}$ by performing several iterations of an iterative method to solve the system

$$A\tilde{u} = f - B^T p^{old}. $$

2. Find an auxiliary function $\tilde{p}$ from the system (the system mimics the mixed approximation of the Poisson problem):

$$BD^{-1}B^T \tilde{p} = \alpha B\tilde{u}, \quad (5.124)$$

where $D \approx A$, e.g., $D = \text{diag}(A).$\footnote{For the choice of $D = A$, solving the system (5.124) is as hard as solving the original system (5.107) (explain why?).} System (5.124) can also be solved approximately.

3. Correction:

$$u^{new} = \tilde{u} - \alpha^{-1}D^{-1}B^T \tilde{p}, \quad p^{new} = p^{old} + \tilde{p}. $$

Different modifications of the method are found in the literature under different names (SIMPLER and other). These modifications differ in how the system with matrix $A$ is approximately solved in the first step and in the choice of the matrix $D$ in (5.124).

However, numerical experiments show that the use of the SIMPLE method as a smoothing iteration does not necessarily lead to typical multigrid efficiency in many cases. Further analysis of the smoothing properties of distributive iterations can be found in [217].

### 5.2.6 Smoother of Braess and Sarazin

This method was introduced by Braess and Sarazin in [33]. Let $D \approx A$, e.g., $D = \text{diag}(A)$; then the smoothing iterations in the method of Braess and Sarazin have the form

$$\begin{bmatrix} u^{m+1} \\ p^{m+1} \end{bmatrix} = \begin{bmatrix} u^m \\ p^m \end{bmatrix} - \begin{bmatrix} \alpha D & B^T \\ B & 0 \end{bmatrix}^{-1} \begin{bmatrix} \mathcal{A}/\mathcal{B} \begin{bmatrix} u^m \\ p^m \end{bmatrix} - \begin{bmatrix} f \\ g \end{bmatrix} \end{bmatrix}. \quad (5.125)$$

Every iteration of (5.125) needs the solution of the auxiliary problem

$$\begin{bmatrix} \alpha D & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} v \\ q \end{bmatrix} = \begin{bmatrix} r^m \\ B u^m - g \end{bmatrix}. \quad (5.126)$$
From (5.126) we see that \( B v = B u^m - g \). This implies \( B u^{m+1} = B(u^m - v) = g \) for all \( i \geq 0 \). Therefore the iterations (5.125) can be considered as smoothing iterations defined on the affine subspace of vectors satisfying inequality \( B v = g \).

Similarly to the Stokes system (see (5.109) and (5.110)), the vector \( v \) can be eliminated from (5.126). The system is reduced to

\[
\alpha^{-1} B D^{-1} B^T q = \alpha^{-1} B D^{-1} r^m - B u^m + g. \tag{5.127}
\]

Consider the simplest case \( D = I \). Then each iteration of (5.125) requires solving the following problem for the pressure update: \( \alpha^{-1} B B^T q = b \). Since \( B \) approximates divergence and \( B^T \) approximates gradient, this problem mimics the discrete Poisson problem. From (5.125) we can deduce the recursive relation for the error in velocity

\[
u - u^{m+1} = (I - B^T(B B^T)^{-1}B)(I - \alpha^{-1}A)(u - u^m). \tag{5.128}\]

This relation clearly shows that the velocity error does not depend on the pressure error on the previous iteration.

Operator \( P := I - B^T(B B^T)^{-1}B \) is the orthogonal projector on \( \text{Ker}(B) \); thus \( ||P|| = 1 \), and for \( \alpha \geq \lambda_{\max}(A) \) it immediately follows from (5.128) that

\[
||u - u^{m+1}||_2 \leq ||u - u^m||_2. \tag{5.129}\]

The following lemma is the basis for checking the smoothing property.

**Lemma 5.10.** Let \( \alpha \geq \lambda_{\max}(A) \); then there exists a vector \( q \) such that

\[
||A(u - u^\nu) + B^T q||_2 \leq \frac{\alpha}{\nu}||u - u^0||_2, \quad \nu = 1, 2, \ldots.
\]

**Proof.** The proof follows arguments from [33]. The recursive relation for the velocity error \( e^\nu = u - u^\nu \) can be written as

\[
e^m = P(I - \alpha^{-1}A)e^{m-1}. \tag{5.130}\]

We introduce the notation \( M = P(I - \alpha^{-1}A)P \); then, employing \( P e^1 = e^1 \), we get from (5.130) that

\[
e^\nu = M^{\nu-1} e^1. \tag{5.131}\]

At the same time, noting the identity \( P e^m = e^m \), we get

\[
\alpha(I - M)e^\nu = \alpha e^\nu - P(\alpha I - A)e^\nu = P Ae^\nu = Ae^\nu + B^T q,
\]

where \( q = -(B B^T)^{-1}B Ae^\nu \). From (5.131) and (5.132) it follows that

\[
||A e^\nu + B^T q||_2 = ||\alpha(I - M)M^{\nu-1} e^1||_2 \leq \alpha ||(I - M)M^{\nu-1}||_2 ||e^1||_2. \tag{5.133}\]

The matrix \( M \) is symmetric, since both matrices \( A \) and \( P \) are symmetric, and \( \text{sp}(M) \in [0, 1] \). Therefore, the estimate \( ||(I - M)M^{\nu-1}||_2 \leq \nu^{-1} \) follows from Lemma 3.11. Thanks to (5.130) we have \( ||e^1||_2 \leq ||e^0||_2 \). This completes the proof of the lemma. \( \square \)

Note that an unknown vector \( q \in P_h \) appears in the lemma assertion. Therefore, to have the complete proof of the smoothing property, and hence the convergence analysis of the multigrid method, one needs another auxiliary step in the algorithm. This step

\[\text{\textsuperscript{16}} e^\nu \] may fail to satisfy this condition if, for example, \( u^\nu \) is an approximation to \( u \) after a coarse grid correction step.
follows immediately after smoothing iterations and before the coarse grid correction. The step consists of finding a vector $\tilde{p}^\nu$ minimizing the norm

$$||Au^\nu + B^T \tilde{p}^\nu - f||_2 = ||A(u - u^\nu) + B^T (p - \tilde{p}^\nu)||_2 = ||A(u - u^\nu) + B^T q||_2.$$  \hspace{1cm} (5.134)

Similarly to one smoothing iteration, this step requires solving the problem of the Poisson type for $\tilde{p}^\nu$:

$$BB^T \tilde{p}^\nu = B(f - Au^\nu).$$  \hspace{1cm} (5.135)

The resulting pair $\{u^\nu, \tilde{p}^\nu\}$ is the output of the Braess–Sarazin smoothing iterations: it is found after executing $\nu$ iterations (5.125) and recovering $\tilde{p}^\nu$ from (5.135). Noting that $B(u - u^\nu) = 0$ for $\nu \geq 1$ and $\alpha = O(h^{-2})$ and employing Lemma 5.10 and (5.134), we get for the error vectors $e^\nu = u - u^\nu$, $q^\nu = p - \tilde{p}^\nu$ the estimate

$$||\alpha\{e^\nu, q^\nu\}||_B = \left(||Ae^\nu + B^T q^\nu||_2^2 + h^{-2}||Be^\nu||_2^2\right)^{\frac{1}{2}}$$

$$= ||Ae^\nu + B^T q^\nu||_2 \leq \frac{c h^{-2}}{\nu} ||e^\nu||_2$$

$$\leq \frac{c h^{-2}}{\nu} (||e^0||_2^2 + h^{-2}||q^0||_2^2)^{\frac{1}{2}} = \frac{c h^{-2}}{\nu} ||(e^0, q^0)||_2.$$  \hspace{1cm} (5.135)

This estimate is the necessary smoothing property (5.118).

5.2.7 - Inexact Uzawa smoother

The method of Braess and Sarazin requires an exact solution of the problem (5.127), which can be interpreted as a discrete pressure Poisson problem. Note that the distributive smoothers avoid this or require only an approximate solution of the similar problem in the form of preconditioner $R$. Below we also consider a smoother closely related to (5.125), which avoids the exact solution of (5.127).

Consider the block iterative method from [14], which can be seen as a variant of the inexact Uzawa method, with a relaxation parameter $\alpha > 0$. Let $G$ be a preconditioner for $\hat{S} := BD^{-1}B$ such that

$$G < \alpha^{-1} \hat{S} \leq (1 + \beta)G, \quad \beta > 0.$$  \hspace{1cm} (5.136)

The matrix $\hat{S}$ can be seen as a Schur complement of an “approximate” discrete Stokes problem, where the submatrix $A$ is replaced by some nonsingular $D$. One step of the method can be divided in the following three substeps:

$$\alpha D(u^{\text{aux}} - u^{\text{old}}) = f - Au^{\text{old}} - B^T p^{\text{old}},$$  \hspace{1cm} (5.137)

$$G(p^{\text{new}} - p^{\text{old}}) = Bu^{\text{aux}} - g,$$  \hspace{1cm} (5.138)

$$\alpha D(u^{\text{new}} - u^{\text{aux}}) = -B^T (p^{\text{new}} - p^{\text{old}}).$$  \hspace{1cm} (5.139)

The iteration method (5.137)–(5.139) is written as the simple linear iteration (1.11) with

$$\Psi = \begin{bmatrix} \alpha D & B^T \\ B & \alpha^{-1} \hat{S} - G \end{bmatrix}.$$  \hspace{1cm} (5.140)

Thus the Braess-Sarazin smoother (5.125) can be interpreted as a special case of (5.137)–(5.139) with an exact preconditioner for $\hat{S}_k$ (for the sake of analysis we need a strict lower
bound in (5.136), however. The condition on the relaxation parameter $\alpha$ from Lemma 5.10 now reads

$$\alpha > \lambda_{\text{max}}(D^{-1}A).$$

(5.140)

For uniform mesh and $D = \text{diag}(A)$, one can take

$$\alpha = O(1).$$

(5.141)

The smoothing property of (5.137)–(5.139) is based on the following lemma from [232].

**Lemma 5.11.** Assume (5.140) and (5.136). Denote

$$\tilde{\mathcal{G}}_s = \begin{bmatrix} \alpha D - A & 0 \\ 0 & \alpha^{-1} S - G \end{bmatrix};$$

then the matrix $L = \tilde{\mathcal{G}}_s^{-1}(I - W^{-1} \mathcal{A})\tilde{\mathcal{G}}_s^{-1}$ is symmetric and

$$\text{sp}(L) \subset \left[ -\beta - \sqrt{\beta^2 + \beta}, 1 \right].$$

Moreover, the identity $\mathcal{A}(I - W^{-1} \mathcal{A}) = \tilde{\mathcal{G}}_s^2 (I - \tilde{L}) \tilde{L}^{-1} \tilde{\mathcal{G}}_s^2$ holds.

Thanks to Lemma 5.11 we prove the smoothing property for inexact Uzawa iterations (5.137)–(5.139).

**Theorem 5.12.** Let $D = \text{diag}(A)$, and let $L = I - W^{-1} \mathcal{A}$ be the iteration matrix of (5.137)–(5.139). Assume (5.140) and (5.136) with $\beta < \frac{1}{4}$; then

$$\|\mathcal{A} L\|_{z\to\beta} \lesssim b^{-2} \frac{1}{\nu - 1}, \quad \nu > 1.$$  

(5.142)

**Proof.** Define the auxiliary matrix

$$\tilde{\mathcal{G}}_b = \begin{bmatrix} I & 0 \\ 0 & b^2 \end{bmatrix};$$

then $\|\cdot\|_b = \langle \tilde{\mathcal{G}}_b \cdot, \cdot \rangle^{\frac{1}{2}}$. Thanks to $D \simeq b^{-2} I$, (5.140), and (5.136) we obtain $\|\tilde{\mathcal{G}}^{-1}_b \tilde{\mathcal{G}}_s\| \lesssim b^{-2}$. Therefore Lemma 5.11 and assumption $\beta < \frac{1}{4}$ yield

$$\|\mathcal{A} L\|_{z\to\beta} = \|\tilde{\mathcal{G}}_b^{-\frac{1}{2}} \mathcal{A} L \tilde{\mathcal{G}}_b^{-\frac{1}{2}}\|_2 = \|\tilde{\mathcal{G}}_b^{-\frac{1}{2}} \tilde{\mathcal{G}}_s^2 (I - \tilde{L}) \tilde{L}^{-1} \tilde{\mathcal{G}}_s^{-\frac{1}{2}} \tilde{\mathcal{G}}_b^{-\frac{1}{2}}\|_2 \leq \|\tilde{\mathcal{G}}_b^{-1} \tilde{\mathcal{G}}_s\|_2 \|(I - \tilde{L}) \tilde{L}^{-1}\|_2 \lesssim b^{-2} \max_{x \in [-\beta - \sqrt{\beta^2 + \beta}, 1]} |(1 - x)x^{-1}| \leq \frac{b^{-2}}{\nu - 1}. \quad \Box$$

To complete the analysis of multigrid preconditioners for the Stokes problem, the smoothing property should be complemented by an approximation property.

### 5.2.8 Approximation property

Assume fine grid and coarse grid spaces $U_h, P_h$ and $U_H, P_H$ such that $U_H \subset U_h \subset H^1_0(\Omega)$, $P_H \subset P_h \subset L^2(\Omega)$ and that both fine grid and coarse grid pairs satisfy the LBB condition (5.101). In this section, we discuss an approximation property that complements
the smoothing property (5.118). Assume for a moment that the space \( \mathbb{P}_h \) is such that the finite element pressure \( p_h \) is continuous on \( \Omega \), and for functions from \( \mathbb{P}_h \) the inverse inequality (3.98) holds. Then from (5.106) one might conclude that

\[
\| u_h - u_H, p_h - p_H \| \leq c h^2 (\| f \|^2 + h^{-2} \| g \|^2)^{1/2}
\]  

(5.143)

if \( f \in U_h \) and \( g \in \mathbb{P}_h \). Further, the standard arguments (see subsection 3.4.1) and (5.143) would imply the approximation property

\[
\| \mathcal{A}_h^{-1} - \mathcal{A}_H^{-1} \alpha \|_{\beta \rightarrow \gamma} \leq c h^2.
\]  

(5.144)

However, the following difficulty arises: one necessary estimate, the inequality (5.106), is valid in the case of a convex polygonal domain subject to the additional condition \( g(x,y) = 0 \) in the vertices of the polygon. This condition is not necessarily satisfied by functions from \( \mathbb{P}_h \). The estimate (5.143) for \( f \in U_h \) and \( g \in \mathbb{P}_h \) was proved in [211] with the help of the following lemma (the proof of the lemma is elementary but rather technical and will be skipped).

**Lemma 5.13.** For finite element spaces \( U_h, \mathbb{P}_h, U_H, \mathbb{P}_H \), we assume the LBB condition (5.101) and standard interpolation properties. Let \( f_h \in U_h, g_h \in \mathbb{P}_h \), and

\[
(f_h, v_H) + (g_h, q_H) = 0 \quad \forall \ v_H \in U_H, \ q_H \in \mathbb{P}_H.
\]  

(5.145)

Then the solution of the finite element problem (5.99)–(5.100) satisfies the estimate

\[
\| u_h, p_h \| \leq c h^2 (\| f_h \|^2 + h^{-2} \| g_h \|^2)^{1/2}.
\]  

(5.146)

Now we show that the lemma yields the desired approximation property. We need to show (5.143). Using the Galerkin orthogonality of the defect \( u_h - u_H, p_h - p_H \) to any pair of vectors from \( \{U_H, \mathbb{P}_H\} \) and employing the LBB inequality (5.101), we get, similarly to (5.102),

\[
\| u_h - u_H, p_h - p_H \| \leq C \inf_{v_H \in U_H, q_H \in \mathbb{P}_H} \inf_{v_h \in U_h, q_h \in \mathbb{P}_h} \| u_h - v_H, p_h - q_H \|.
\]

We choose \( v_H \) and \( q_H \) to be the orthogonal projections of \( u_h \) and \( p_h \) on \( U_H \) and \( \mathbb{P}_H \). With this (as well as with any other) choice of \( v_H \) and \( q_H \), we have

\[
\| u_h - u_H, p_h - p_H \| \leq C \| u_h - v_H, p_h - q_H \|.
\]  

(5.147)

Now we estimate the right-hand side of (5.147). To do this, consider \( w_h \) and \( r_h \) solving the following finite element Stokes problem:

\[
(\nabla w_h, \nabla v_h) - (r_h, \text{div} v_h) = (u_h - v_H, v_h) \quad \forall \ v_h \in U_h,
\]

\[
-(\text{div} u_h, q_h) = b^2 (p_h - q_H, q_h) \quad \forall \ q_h \in \mathbb{P}_h.
\]

The right-hand side of the system satisfies the assumptions of Lemma 5.13, and hence it holds that

\[
\| w_h, r_h \| \leq c h^2 \| u_h - v_H, p_h - q_H \|.
\]  

(5.148)

Now, using the orthogonality of \( u_h - v_H \) to all functions from \( U_H \) and \( p_h - q_H \) to all functions from \( \mathbb{P}_H \), recalling the definition of \( u_h, p_h \) and \( w_h, r_h \), we write

\[
\| u_h - v_H, p_h - q_H \|^2 = (u_h, u_h - v_H) + b^2 (p_h, p_h - q_H)
\]

\[
= (f, w_h) + (g, r_h) \leq (\| f \|^2 + b^{-2} \| g \|^2)^{1/2} (\| w_h \|^2 + b^2 \| r_h \|^2)^{1/2}
\]

\[
\leq c h^2 (\| f \|^2 + b^{-2} \| g \|^2)^{1/2} \| u_h - v_H, p_h - q_H \|.
\]
Together with (5.147) this proves (5.143). Therefore, we obtain the desired approximation property (5.144).

5.2.9 Inner-outer iterations

There is another way to build iterative methods for the Stokes and other equations having similar block structure. The idea is to exploit fast solvers or efficient preconditioners available for subblocks of matrix $\mathcal{A}$ or for matrices from block factorizations of $\mathcal{A}$, as in (5.115). Using these inner solvers or preconditioners one can try to design an outer solver or preconditioner. First, we illustrate this idea of inner-outer solvers by the classical Uzawa method (the method was first considered in [7] to solve some problems of linear programming).

Consider the pressure equation (5.110) with the Schur complement matrix $S$:

$$S p := BA^{-1}B^T p = BA^{-1}f - g.$$  \hfill (5.149)

We know from Lemma 5.9 that for LBB stable discretizations the matrix $S = BA^{-1}B^T$ is symmetric positive definite and spectrally equivalent to the mass matrix $M$ with the equivalence constants independent of $h$. For piecewise constant elements, the pressure space mass matrix $M$ is diagonal. For higher order elements, it is well known (see [214]) that the diagonal matrix $\hat{M} = \text{diag}(M)$ is spectrally equivalent to $M$: $\text{cond}($$\hat{M}^{-1}M$$) \leq c$ with a constant $c$, independent of $h$. Therefore, $\text{cond}($$\hat{M}^{-1}S$$) \leq c \mu(\Omega)^{-2}$.\hfill (5.149)

Assume that we can accurately solve a linear algebraic system with the matrix $A$. Since the matrix $A$ results from the discretization of two Poisson problems (one for each velocity component), a conjugate gradient (CG) method with a multigrid preconditioner is a good candidate for accomplishing this task. If we can solve the system $Av = r$, then it is possible to calculate the matrix-vector product of $S$ with any vector from $\mathbb{R}^m$. As a result, the pressure equation (5.110) can be approximately solved by a preconditioned iterative method, with $M^{-1}$ or $\hat{M}^{-1}$ standing for the preconditioner. The preconditioned CG method is appropriate as the outer iterator. The Richardson method for solving (5.149) is known as the Uzawa algorithm. It is convergent if the parameter $\bar{w}$ is sufficiently small. Once the pressure part of the vector of unknowns is found, the velocity part can be recovered.

Unfortunately, it can be very expensive to solve the system $Av = r$ with high accuracy on every iterative step of an outer method. There are several strategies for replacing the calculation of $A^{-1}r$ by applying a preconditioner $\hat{A}^{-1}$. Virtually all of them exploit the block structure of the discrete Stokes system matrix $\mathcal{A}$.

5.2.10 Block-diagonal preconditioner

Based on (5.115) it is natural to consider the following block-diagonal preconditioner for $\mathcal{A}$ [73]:

$$\widehat{\mathcal{A}} = \begin{bmatrix} \hat{A} & 0 \\ 0 & \hat{S} \end{bmatrix},$$  \hfill (5.150)

where $\hat{A}$ and $\hat{S}$ are preconditioners for the block $A$ and the Schur complement matrix $S$. It is straightforward to check that for $\hat{A} = A$ and $\hat{S} = S$ the annihilating polynomial for the preconditioned matrix $\widehat{\mathcal{A}}^{-1}$ has degree 3. Thus a preconditioned minimal residual method would converge in three iterations. Of course, this choice of $\hat{A}$ and $\hat{S}$ is not very
practical. In a more general case, the efficiency of the block-diagonal preconditioner can be judged based on eigenvalues estimates from the following lemma.

**Lemma 5.14.** Assume $A, \hat{A}, \hat{S}$ are symmetric positive definite matrices and $S$ is the pressure Schur complement of (5.107). Assume also that

$$v_1 \hat{A} \leq A \leq v_2 \hat{A}, \quad \mu_1 \hat{S} \leq S \leq \mu_2 \hat{S},$$

with positive constants $v_1, v_2, \mu_1, \mu_2$. Then all eigenvalues of the problem

$$Au + B^T p = \lambda \hat{A} u, \quad Bu = \lambda \hat{S} p$$

belong to the union of the intervals,

$$[v_1, v_2] \cup \left[\frac{v_1 + \sqrt{v_1^2 + 4v_1 \mu_1}}{2}, \frac{v_2 + \sqrt{v_2^2 + 4v_2 \mu_2}}{2}\right]$$

and

$$\left[\frac{v_2 - \sqrt{v_2^2 + 4v_2 \mu_2}}{2}, \frac{v_1 - \sqrt{v_1^2 + 4v_1 \mu_1}}{2}\right].$$

**Proof.** We slightly modify the proof given in [106]. Assume for a moment that $\lambda > v_2$; then the operator $\lambda \hat{A} - A$ is nonsingular. Thus we can apply $(\lambda \hat{A} - A)^{-1}$ to the first equation in (5.153) to obtain $(\lambda \hat{A} - A)^{-1} B^T p = u$. Applying $B$ to this identity yields

$$B(\lambda \hat{A} - A)^{-1} B^T p = \lambda \hat{S} p.$$

Computing the scalar product with $p$ gives

$$\langle (\lambda \hat{A} - A)^{-1} B^T p, B^T p \rangle = \langle \lambda \hat{S} p, p \rangle.$$

Condition (5.152) implies

$$\frac{\lambda}{\mu_2} \langle S p, p \rangle \leq \langle (\lambda \hat{A} - A)^{-1} B^T p, B^T p \rangle \leq \frac{\lambda}{\mu_1} \langle S p, p \rangle.$$

Denoting $v = B^T p$ and recalling $S = BA^{-1} B^T$, we get

$$\frac{\lambda}{\mu_2} \leq \frac{\langle (\lambda \hat{A} - A)^{-1} v, v \rangle}{\langle A^{-1} v, v \rangle} \leq \frac{\lambda}{\mu_1}.$$

At the same time, it holds that

$$\sigma_{\min} \leq \frac{\langle (\lambda \hat{A} - A)^{-1} v, v \rangle}{\langle A^{-1} v, v \rangle} \leq \sigma_{\max},$$

where $\sigma_{\min}$ and $\sigma_{\max}$ are the lower and upper bounds for the eigenvalue problem

$$(\lambda \hat{A} - A)^{-1} v = \sigma A^{-1} v.$$
Using the fact that $A_1 \preceq A_2 \iff A_2^{-1} \preceq A_1^{-1}$ for two symmetric positive definite matrices and (5.151), one readily concludes that (5.156) holds with

$$\sigma_{\min} = \frac{\nu_1}{\lambda - \nu_1}, \quad \sigma_{\max} = \frac{\nu_2}{\lambda - \nu_2}.$$ 

Thus, the estimates (5.155) and (5.156) imply that any eigenvalue $\lambda > \nu_2$ satisfies

$$\frac{\lambda}{\mu_2} \leq \frac{\nu_2}{\lambda - \nu_2} \quad \text{and} \quad \frac{\nu_1}{\lambda - \nu_1} \leq \frac{\lambda}{\mu_1}. \quad (5.157)$$

Resolving (5.157), we get for $\lambda > \nu_2$ that

$$\lambda \in \left[ \frac{\nu_1 + \sqrt{\nu_1^2 + 4\nu_1\mu_1}}{2}, \frac{\nu_2 + \sqrt{\nu_2^2 + 4\nu_2\mu_2}}{2} \right].$$

We have the second interval in (5.154). The third interval can be obtained by similar considerations for the case $\lambda < \nu_1$. Finally, the first interval covers the case $\nu_1 \leq \lambda \leq \nu_2$, excluded from our previous considerations. □

Thanks to the lemma and the convergence estimate (1.108), which is valid for the (nongeneralized) minimal residual method, we can make some conclusions about the convergence of the minimal residual method with the block-diagonal preconditioner $\hat{A}$. Applying results of Lebedev [126] on optimal polynomials for the case of several intervals, we get the residual reduction in $k$ steps,

$$\frac{\|r^k\|_2}{\|r^0\|_2} \leq \left( \frac{1 - \sqrt{b_{-1}a_1/b_{-1}a_{-1}}}{1 + \sqrt{b_{-1}a_1/b_{-1}a_{-1}}} \right)^{k/2}, \quad (5.158)$$

where

$$a_{-1} = \frac{\nu_2 - \sqrt{\nu_2^2 + 4\nu_2\mu_2}}{2}, \quad b_{-1} = \frac{\nu_1 - \sqrt{\nu_1^2 + 4\nu_1\mu_1}}{2}.$$ 

In particular, if one applies few multigrid cycles to define $\hat{A}$ and a proper preconditioner for $S$, e.g., the pressure mass matrix for LBB stable finite elements, the constants $a_{-1}$, $b_{-1}$, $a_1$, and $b_1$ are independent of the mesh size $h$, since this is true for $\nu_1$, $\nu_2$, $\mu_1$, and $\mu_2$. The convergence of a preconditioned method is guaranteed to be bounded independent of the mesh size. The total arithmetic complexity for solving the system up to a fixed accuracy depends linearly on the number of unknowns; such complexity is optimal.

From among the many papers addressing preconditioned iterative strategies for the Stokes problem, we point to [37], where the problem is reformulated and preconditioned in a way that a positive definite symmetric matrix emerges. This makes possible the application of the CG method. As in most methods, the necessary ingredients are preconditioners for the $(1,1)$-block and the Schur complement matrices, i.e., $\hat{A}$ and $\hat{S}$.

### 5.2.11 Augmented-Lagrangian preconditioners I

A block-diagonal preconditioner needs a submatrix $\hat{S}^{-1}$ which is an approximate inverse of the Schur complement $S = BA^{-1}B^T$. For a stable discretizations of the Stokes problem, such $\hat{S}^{-1} = \hat{M}_p^{-1}$ is readily available. Here we discuss an approach that allows us to build
### 5.2. Preconditioners for certain problems of fluid mechanics

suitable block preconditioners when $\hat{S}^{-1}$ is not available or $\text{cond}(\hat{S}^{-1}S)$ is still too large.

A particular example of such a system is the nonsymmetric Oseen problem considered later in this section. However, we begin with the symmetric case.

Assume $W \in \mathbb{R}^{m \times m}$ is symmetric positive definite and $\gamma > 0$ is a parameter. The system (5.128) is equivalent to the system

$$
\begin{bmatrix}
A + \gamma B^T W^{-1}B & B^T \\
B & 0
\end{bmatrix}
\begin{bmatrix}
u \\
p
\end{bmatrix}
=
\begin{bmatrix}
\hat{f} \\
g
\end{bmatrix},
$$

(5.159)

with $\hat{f} := f + \gamma B^T W^{-1}g$. The solution to (5.159) and, because of the equivalence, the solution to (5.128) is the saddle point of the Lagrangian (check this!)

$$
\mathcal{L}_\gamma(u, p) = \langle Au, u \rangle + 2\langle Bu - g, p \rangle - 2\langle \hat{f}, u \rangle + \gamma \|W^{-1/2}(Bu - g)\|_2^2.
$$

(5.160)

This can be seen as an augmented-Lagrangian for the original system (5.128). This point of view on (5.159) was taken in [79] and gave the name augmented-Lagrangian preconditioning to the resulting approach.

In order to maintain sparsity in $A + \gamma B^T W^{-1}B$, the matrix $W$ should be chosen such that $W^{-1}$ is sparse. For the Stokes problem, the good choice is $W = \hat{M}_p$. The choice of $\gamma$ is important and will be discussed below.

Denote the (1,1)-block of the augmented matrix by $A_{\gamma} := A + \gamma B^T W^{-1}B$. Matrix $S_{\gamma} = BA_{\gamma}^{-1}B^T$ is the Schur complement of (5.159). The Schur complements of (5.128) and of the augmented systems satisfy the identity

$$
S^{-1}_{\gamma} = S^{-1} + \gamma W^{-1}.
$$

(5.161)

The proof of (5.161) for a slightly more general case will be given in Lemma 5.15.

Consider two generalized eigenvalue problems

$$
S q = \mu W q \quad \text{and} \quad S_{\gamma} q = \lambda W q.
$$

From (5.161), we have

$$
\lambda^{-1} = \mu^{-1} + \gamma.
$$

(5.162)

If we choose $W$ as a preconditioner for $S_{\gamma}$, we immediately get from (5.162) for spectral condition numbers (check this!)

$$
\text{cond}(W^{-1}S_{\gamma}) = \frac{\mu^{-1}_{\min} + \gamma}{\mu^{-1}_{\max} + \gamma}.
$$

Thus,

$$
\text{cond}(W^{-1}S_{\gamma}) < \text{cond}(W^{-1}S) \quad \text{for any } \gamma > 0,
$$

$$
\text{cond}(W^{-1}S_{\gamma}) \to 1 \quad \text{for } \gamma \to \infty.
$$

(5.163)

Properties (5.162) and (5.163) show the positive effect of the augmentation on the (preconditioned) Schur complement conditioned numbers. Therefore, instead of preconditioner (5.150) for (5.90), one considers the block-diagonal preconditioner for the equivalent system (5.159):

$$
\hat{A} = \begin{bmatrix}
\hat{A}_{\gamma} & 0 \\
0 & \delta W
\end{bmatrix}.
$$

(5.164)
The parameter $\delta > 0$ introduces the scaling necessary to balance the contributions of (1,1)- and (2,2)-blocks in (5.164). The need of such a balance is seen from the appearance of constants from (5.151) in the convergence estimate (5.158) for the MINRES method.

The possibility of improving the conditioning of the block system Schur complement comes at a price. The matrix $\gamma B^T W^{-1} B$, which is added to the (1,1)-block of the system, may have large kernel, and, for large $\gamma$, building a good preconditioner for $A_\gamma$ cannot be easier. A closer look at $A_\gamma$ reveals that this matrix is the stiffness (discretization) matrix of the finite element method (5.86) for the linear elasticity problem, where $P_h$ is chosen to be an orthogonal $L^2$-projector into the pressure space $P_h$. Therefore, the methods discussed in subsection 5.1.12 can be used here.

In practice, one rarely needs augmentation for symmetric Stokes type problems, since good preconditioners are usually readily available for the Schur complement of the original nonaugmented system; see, e.g., [51, 93, 142, 154, 158, 233] for the treatment of Schur complement preconditioners for more complicated parameter-dependent symmetric saddle point problems. Nevertheless, the augmentation can be important for a symmetric saddle point problem if the (1,1)-block is singular and the direct block factorization is not possible. This is the case, for example, for certain discretizations of Maxwell equations; see [85]. Moreover, augmentation can be vital for building efficient block preconditioners for saddle point algebraic problems with a (highly) nonsymmetric (1,1)-block, as discussed in the next subsection.

### 5.2.12 Augmented-Lagrangian preconditioners II

The augmented-Lagrangian approach can be especially useful if nonsymmetric saddle point type problems are treated. One prominent example of such problems is the Oseen equations,

\[
\sigma u - \nu \Delta u + (\nu \cdot \nabla) u + \nabla p = f \\
div u = 0 \quad \text{in } \Omega.
\] (5.165)

The problem arises from implicit time discretization or linearization of the Navier–Stokes system (5.90) by a fixed point iteration. Then $v$ is a known velocity field from a previous iteration or time step and $\sigma$ is proportional to the reciprocal of the time step parameter ($\sigma = 0$ for a steady problem). When $v = 0$ we have a (generalized) Stokes problem. Equations as in (5.165) are also a part of the Leray-alpha fluid model.

Similarly to the Stokes problem, spatial discretization of (5.165) using finite differences or finite elements results in large, sparse $2 \times 2$-block systems of the form

\[
\begin{bmatrix}
A & B^T \\
B & -C
\end{bmatrix}
\begin{bmatrix}
u \\
p
\end{bmatrix} =
\begin{bmatrix}
f \\
g
\end{bmatrix}.
\] (5.166)

The (1,1)-block matrix $A$ is the discretization of the diffusion, convection terms and the $\sigma$-term. Hence, $A$ is not symmetric unless $v = 0$. We would like to consider an even more general problem and so assume that the (2,2)-block of the matrix is not necessarily zero but a symmetric semidefinite matrix $C$ such that $C \leq 0$. If the discretization satisfies the LBB stability condition (5.101), then it is common to have $C = 0$. If the LBB condition is not satisfied, the matrix $C$ corresponds to additional stabilization terms added to a discrete problem; see, e.g., [73].

Further presentation in this section follows [20]. Due to the presence of a nonzero (2,2)-block, the augmentation of system (5.166) must be done differently than in the case of stable finite elements. As before, let $\gamma > 0$ and let $W$ be a symmetric positive definite
5.2. Preconditioners for certain problems of fluid mechanics

matrix. Then from \( Bu - C p = 0 \) it follows that

\[
\gamma B^T W^{-1} Bu - \gamma B^T W^{-1} C p = 0.
\]

Adding the above equation to \( Au + B^T p = f \) gives

\[
(A + \gamma B^T W^{-1} B) u + (B^T - \gamma B^T W^{-1} C) p = f.
\]

Therefore, the (first) augmented linear system is

\[
\begin{bmatrix}
  A_{\gamma} & B^T_{\gamma} \\
  B & -C
\end{bmatrix}
\begin{bmatrix}
  u \\
  p
\end{bmatrix}
= \begin{bmatrix}
  f \\
  0
\end{bmatrix}
\]

or \( \mathcal{A} x = b \), (5.167)

where \( A_{\gamma} = A + \gamma B^T W^{-1} B \) and \( B^T_{\gamma} = B^T - \gamma B^T W^{-1} C \).

Notice that in (5.167), the (1,2)-block \( B^T_{\gamma} \), is not equal to the transpose of the (2,1)-block \( B \). To get a more “symmetric” augmented linear system, we can obtain from \( Bu - C p = 0 \) the equation

\[
-\gamma C W^{-1} Bu + \gamma C W^{-1} C p = 0.
\]

Then, combining this equation with \( Bu - C p = 0 \), we have

\[
(B - \gamma C W^{-1} B) u - (C - \gamma C W^{-1} C) p = 0.
\]

Letting \( C_{\gamma} = C - \gamma C W^{-1} C \), we obtain the second augmented system

\[
\begin{bmatrix}
  A_{\gamma} & B^T_{\gamma} \\
  B_{\gamma} & -C_{\gamma}
\end{bmatrix}
\begin{bmatrix}
  u \\
  p
\end{bmatrix}
= \begin{bmatrix}
  f \\
  0
\end{bmatrix}.
\]

(5.168)

Numerical experience shows very a similar performance of iterative methods based on (5.167) and (5.168).

For nonsymmetric block systems it makes sense to consider block-triangle rather than block-diagonal preconditioners, since preserving a symmetry is not a concern anymore. When used in the (generalized) minimal residual method, the block-triangle version generally shows better results, similarly to the Gauss–Seidel preconditioner, which in most circumstances performs better than the Jacobi one. Thus, consider the block-triangular preconditioner in the form

\[
\mathcal{P} = \begin{bmatrix}
  A_{\gamma} & 0 \\
  B & -\hat{\mathcal{S}}
\end{bmatrix}.
\]

(5.169)

For the “symmetrized” system (5.168), \( B \) should be replaced by \( B_{\gamma} \). Instead of the lower triangular preconditioner (5.169) one may consider the upper triangular preconditioner:

\[
\mathcal{P} = \begin{bmatrix}
  A_{\gamma} & B^T_{\gamma} \\
  0 & -\hat{\mathcal{S}}
\end{bmatrix}.
\]

For the symmetric problem we set \( \hat{\mathcal{S}} = \gamma^{-1} W = \gamma^{-1} \hat{M}_p \). As we will see, for the case of \( C \neq 0 \) the choice of \( \hat{\mathcal{S}} \) and \( W \) is more delicate. Matrix \( \hat{\mathcal{S}} \) intends to approximate the pressure Schur complement of the augmented system, i.e., the matrices \( \hat{S}_\gamma := B_{\gamma} A_{\gamma}^{-1} B_{\gamma}^T + C_{\gamma} \)
for (5.168) and \( \tilde{S}_\gamma := BA_\gamma^{-1}B_\gamma^T + C \) for (5.167). The following result, which extends the representation in (5.161) to the case of \( C \neq 0 \), will help us to set \( W \), build the preconditioner \( \hat{S} \), and analyze the spectrum of the preconditioned system.

**Lemma 5.15.** Assuming all the relevant matrices are invertible, it holds that

\[
\begin{align*}
S_\gamma^{-1} &= S^{-1} + \gamma(W - \gamma C)^{-1}, \quad (5.170) \\
\tilde{S}_\gamma^{-1} &= S^{-1}(I - \gamma CW^{-1}) + \gamma W^{-1}. \quad (5.171)
\end{align*}
\]

**Proof.** The matrix \( X := -S_\gamma^{-1} \) is the (2,2)-block of the inverse of the coefficient matrix in (5.168). Denoting by \( Y \) the (1,2)-block of this inverse matrix, we get the following system of matrix equations:

\[
\begin{align*}
(A + \gamma B^T W^{-1} B)Y + B^T (I - \gamma W^{-1} C)X &= 0, \quad (5.172) \\
(I - \gamma CW^{-1})BY - (I - \gamma CW^{-1})CX &= I. \quad (5.173)
\end{align*}
\]

From (5.173) we get \( BY = (I - \gamma CW^{-1})^{-1} + CX \). Substituting this into (5.172) and applying \( A^{-1} \) leads to

\[
Y = -\gamma A^{-1} B^T W^{-1} (I - \gamma CW^{-1})^{-1} - A^{-1} B^T X.
\]

Now substituting \( Y \) into (5.173) gives, after simple manipulations,

\[
-(BA^{-1} B^T + C)X(I - \gamma CW^{-1}) = I + \gamma BA^{-1} B^T W^{-1}.
\]

By straightforward computations one verifies that the last equation is solved by matrix

\[
X = -(BA^{-1} B^T + C)^{-1} - \gamma(W - \gamma C)^{-1}.
\]

Thus (5.170) is proved. The result in (5.171) follows from the obvious identity \( (I - \gamma CW^{-1}) \tilde{S}_\gamma = S_\gamma \).

The expressions (5.170) and (5.171) suggest that the auxiliary matrix \( W \) should be such that \( W - \gamma C \) is positive definite. The following two choices of \( W \) satisfy this constraint:

\[
\begin{align*}
W_1 &= \tilde{M}_p + \gamma C, \\
W_2 &= \tilde{M}_p \quad \text{with} \quad 0 < \gamma \leq (2\|\tilde{M}_p^{-1} C\|_2)^{-1}.
\end{align*}
\]

Setting \( W = W_1 \) leads to a simple choice of preconditioner \( \hat{S} \) such that the preconditioned system enjoys eigenvalue bounds independent on matrix \( C \). At the same time, \( W = W_1 \) involves matrix \( C \), which often mimics the discrete Laplacian matrix. Hence the inverse \( W^{-1} \) may become an (almost) full matrix, resulting in that \( A_\gamma = A + B^T W^{-1} B \) is an (almost) full matrix, consequently making the solution of linear systems with \( A_\gamma \) much more difficult. This happens, for example, with \( Q_1 \cdot Q_1 \) elements. For \( Q_1 \cdot P_0 \) elements, however, the matrix \( C \) has a special block-diagonal structure, which leads to a relatively cheap solve with \( A_\gamma \).

The choice \( W = W_2 \) preserves the sparsity of \( A_\gamma \). However, the restriction on \( \gamma \) yields the decrease of \( \gamma \) when \( \nu \) is small and \( b \) tends to zero, since for small \( \nu \) it holds that \( \|\tilde{M}_p^{-1} C\|_2 = O(b^{-1}) \); see the discussion in [20]. Thus less augmentation is introduced and the performance of the solver becomes more sensitive to the variation in \( \nu \) and \( b \).

Below we present the eigenvalue analysis and show the corresponding choices of \( \hat{S} \) for both cases \( W = W_1 \) and \( W = W_2 \). We shall also discuss a third (intermediate) alternative...
of setting the augmentation and preconditioning, which is not covered by the analysis but shows stable and almost $\nu$- and $h$-independent convergence behavior while keeping the matrix $A_\gamma$ sparse.

Consider the following generalized eigenvalue problem:

$$
\begin{bmatrix}
A_\gamma & B^T \\
B & -C
\end{bmatrix}
\begin{bmatrix}
u \\
p
\end{bmatrix} = \lambda
\begin{bmatrix}
A_\gamma & 0 \\
B & -S
\end{bmatrix}
\begin{bmatrix}
u \\
p
\end{bmatrix}.
$$

(5.174)

For the symmetrized system, matrix $B$ in the (2,1)-block is replaced by $B_\gamma$, and matrix $C$ in the (2,2)-block is replaced by $C_\gamma$. As in the case of stable finite elements, we consider the eigenvalue problem

$$
S q = \mu M_p q,
$$

(5.175)

where $S = BA^{-1}B^T + C$, and obtain bounds on $\lambda$ in terms of $\mu$.

For $W = W_1$ one immediately gets from (5.170) and (5.171) that

$$
S^{-1}_\gamma = S^{-1} + \gamma M_p^{-1},
$$

$$
\tilde{S}^{-1}_\gamma = S^{-1}M_p(M_p + \gamma C)^{-1} + \gamma(M_p + \gamma C)^{-1}.
$$

Therefore, setting

$$
\tilde{S} := \gamma^{-1}M_p + C \text{ for (5.167)} \quad \text{or} \quad \tilde{S} := \gamma^{-1}M_p \text{ for (5.168),}
$$

(5.176)

we obtain that all nonunit eigenvalues of (5.174) satisfy $S^{-1} p + \gamma M_p^{-1} p = \lambda^{-1} \gamma M_p^{-1} p$, where $p \neq 0$, and thus

$$
\lambda = \frac{\gamma \mu}{1 + \gamma \mu}.
$$

This representation is identical to the one in (5.162). Therefore, we obtain the following theorem.

**Theorem 5.16.** Assume that $W = W_1$, and that $\tilde{S}$ is defined as in (5.176). The preconditioned matrix $\tilde{S}^{-1} \tilde{\mathcal{G}}$ has the eigenvalue 1 of multiplicity at least $n$. All other (nonunit) eigenvalues satisfy the following bounds:

$$
0 < \min_{\mu \neq 0} \frac{\gamma a_\mu}{1 + \gamma a_\mu} \leq a_\lambda \leq 1, \quad |b_\lambda| \leq \max_{\mu \neq 0} \left( \gamma \frac{|b_\mu|}{\gamma |b_\mu|} \right) \leq 1,
$$

where $\lambda = a_\lambda + \beta b_\lambda$ and $\mu = a_\mu + \beta b_\mu$.

We noted already that the choice $W = W_1$ is not always practical. The next theorem shows eigenvalue bounds for the case $W = W_2$ in terms of the bounds given by Bendixson’s theorem [188] for the generalized eigenvalue problem (5.175),

$$
\alpha_\mu := \min_{p \neq 0} \frac{|(p, Dp)|}{(p, M_p p)} \leq a_\mu, \quad |b_\mu| \leq \beta_\mu := \max_{p \neq 0} \left( \frac{|(p, Rp)|}{(p, M_p p)} \right),
$$

(5.177)

where $\mu = a_\mu + \beta b_\mu$, $D = B \frac{A^{-1} - B^T}{2} B^T + C$ is the symmetric part of $S$, and $R = B \frac{A^{-1} - B^T}{2} B^T$ is its skew-symmetric part.
Theorem 5.17. Assume $W = W_2$, $0 < \gamma \leq (2\|M_p^{-1}C\|_2)^{-1}$, and $\hat{S} = \gamma^{-1}M_p$. The preconditioned matrix $\mathcal{P}^{-1}\hat{S}$ has the eigenvalue 1 of multiplicity at least $n$. All other (nonunit) eigenvalues satisfy the following bounds for the nonsymmetric augmentation (5.167),

$$0 < \frac{\gamma a_\mu}{1 + \gamma a_\mu} \leq a_\lambda \leq 1, \quad |b_\lambda| \leq \max_{\mu} \{\gamma \beta_\mu, 1\} \leq 1,$$

(5.178)

where $\lambda = a_\lambda + \beta b_\lambda$.

Proof. From (5.174), we immediately get that $\lambda = 1$ is eigenvalue of multiplicity (at least) $n$ and any vector $[u;0]$ with $u \neq 0$ is a corresponding eigenvector. The remaining eigenvalues $\lambda$ satisfy

$$\tilde{S}_p = \lambda \tilde{S} p.$$

For $W = M_p$, $\tilde{S} = \gamma^{-1}M_p$, using representation (5.171), we obtain

$$S_p = \frac{\lambda}{1 - \lambda} \left(1 - M_p - C\right) p.$$

(5.179)

For brevity, we let $\eta = \frac{1}{\lambda}$ and $Q = \frac{1}{\gamma}M_p - C$. It follows from Bendixson’s theorem that

$$\min_{\eta \neq 0} \frac{\langle p, Dp \rangle}{\langle p, Qp \rangle} \leq a_\eta \leq \max_{\eta \neq 0} \frac{\langle p, Dp \rangle}{\langle p, Qp \rangle}, \quad |b_\eta| \leq \max_{\eta \neq 0} \frac{|\langle p, Rp \rangle|}{\langle p, Qp \rangle},$$

where $\eta = a_\eta + \beta b_\eta$. Using (5.179) we shall obtain bounds for $\lambda$ in terms of $\mu$ from (5.175). Since $\gamma$ satisfies $0 < \gamma \leq (2\|M_p^{-1}C\|_2)^{-1}$, it holds that

$$\frac{1}{2\gamma}M_p \leq \frac{1}{\gamma}M_p - C \leq \frac{1}{\gamma}M_p.$$

(5.180)

Therefore, we have

$$\gamma \min_{\eta \neq 0} \frac{\langle p, Dp \rangle}{\langle p, M_p p \rangle} \leq \min_{\eta \neq 0} \frac{\langle p, Dp \rangle}{\langle p, Qp \rangle} \leq a_\eta \leq \max_{\eta \neq 0} \frac{\langle p, Dp \rangle}{\langle p, Qp \rangle} \leq 2\gamma \max_{\eta \neq 0} \frac{\langle p, Dp \rangle}{\langle p, M_p p \rangle},$$

$$|b_\eta| \leq \max_{\eta \neq 0} \frac{|\langle p, Rp \rangle|}{\langle p, Qp \rangle} \leq 2\gamma \max_{\eta \neq 0} \frac{|\langle p, Rp \rangle|}{\langle p, M_p p \rangle}.$$

Applying (5.177) yields

$$a_\eta \geq \gamma a_\mu > 0, \quad |b_\eta| \leq 2\gamma \beta_\mu.$$

(5.181)

Solving $\eta = \frac{1}{\lambda}$ for $a_\lambda$ and $b_\lambda$, we have

$$a_\lambda = \frac{a_\eta (1 + a_\eta) + b_\eta^2}{(1 + a_\eta)^2 + b_\eta^2}, \quad b_\lambda = \frac{b_\eta}{(1 + a_\eta)^2 + b_\eta^2}.$$

From this and (5.181) the result in (5.178) follows. □

The bounds for $\lambda$ in Theorems 5.16 and 5.17 are written in terms of bounds for the eigenvalues $\mu$ from (5.175). Following the same argument as in [72], one can prove that $a_\mu$
and $\beta_\rho$ from (5.177), and hence the smallest real and the largest imaginary parts of $\mu$ are independent of $b$ but depend on $\nu$. The resulting eigenvalue bounds for $\lambda$ are very similar to those for the LBB stable case ($C = 0$) from [19, 20]. This suggests that the choice of sufficiently large $\gamma$ leads to a method which is essentially insensitive to variations of parameters $\nu$ and $b$. However, for the practical choice of $W = \hat{M}_p$ we have the restriction on $\gamma$. Numerical experiments show that with the setting of $W$ and $\hat{S}$ from Theorem 5.17 the restriction is indeed important and prohibits the choice of $\gamma = O(1)$ for all values $\nu$ and $b$ of interest. The situation looks better from the numerical viewpoint if one sets (for the nonsymmetrized case (5.167))

$$W = \hat{M}_p, \quad \hat{S} := \gamma^{-1}M_p + C.$$  \hspace{1cm} (5.182)

This combination of the augmentation and preconditioning, which is intermediate between those in Theorems 5.16 and 5.17, is not covered by the eigenvalue analysis above. However, in practice it leads to performance which is nearly robust with respect to variation of problem parameters. Further analysis of the augmented-Lagrangian approach for nonsymmetric saddle point problems can be found in [18, 21, 103, 153]. In particular, based on numerical range bounds of a preconditioner system, the paper [18] gives rigorous convergence estimates for the (generalized) minimal residual method.

### Exercises

5.2.1. Assume there is a pair of finite element spaces $U_h$ and $P_{bb} h$, and $U_h \subset H^1_0(\Omega)$. Prove the following statements:

a. Infimum in the right-hand side of (5.101) is attained for some $\hat{q}_b \in P_{bb} h$, $\hat{q}_b \neq 0$.

b. If $\mu(\Omega) = 0$ in (5.101), then there exists a nontrivial solution $\{\hat{u}_h, \hat{p}_h\}$ of the system (5.99)–(5.100) with $f = 0$ and $g = 0$.

c. Under the assumptions of item b there exists $f \in L^2(\Omega)^2$, $g \in L^2_0(\Omega)$ such that the problem (5.99)–(5.100) does not have a solution.

5.2.2. Using the identity $\Delta = \nabla \text{div} - \text{rot} \text{rot}$, check that for arbitrary $v \in H^1_0(\Omega)$ it holds that

$$||\nabla v||^2 = ||\text{div} v||^2 + ||\text{rot} v||^2,$$

where $\text{rot} v = -\frac{\partial v_y}{\partial x} + \frac{\partial v_x}{\partial y}$ in $\mathbb{R}^2$ and $\text{rot} v = \nabla \times v$ in $\mathbb{R}^3$.

5.2.3. For symmetric $A \in \mathbb{R}^{n \times n}$, $C \in \mathbb{R}^{m \times m}$, and $B \in \mathbb{R}^{m \times n}$, let

$$A' = \begin{bmatrix} A & B^T \\ B & C \end{bmatrix}.$$  

Show for the spectral matrix norm

$$||A'||_2 \leq ||A||_2 + ||B||_2 + ||C||_2.$$

5.2.4. Show that the SIMPLE method can be written in the form of distributive iterations (5.120).
5.2.5. (See [33].) For the following example (the one-dimensional Stokes problem),
\[ \gamma u - u'' + p' = f \quad \text{on} \quad (0,2\pi), \quad \gamma \geq 0, \]
\[ -u' = 0 \quad \text{on} \quad (0,2\pi), \]
\[ u(0) = u(2\pi) = u_r, \quad \int_0^{2\pi} p \, dx = 0, \]
demonstrate that the SIMPLE method with \( D = \text{diag}(A) \) is not a good smoother for \( \gamma \ll h^{-2} \).

5.2.6. Set \( \hat{A} = A \) and \( \hat{S} = S \), and let \( \Delta \) be a block-diagonal preconditioner from (5.150). Show that the annihilating polynomial for the preconditioned matrix \( \Delta^{-1} \Delta_b \) has degree 3.

5.2.7. Consider a \( 2 \times 2 \)-block matrix (5.107) with \( A = A^T > 0 \) and \( S \) the Schur complement. Show that \( \text{cond}(S) \leq \text{cond}(A) \) for spectral condition numbers.

5.2.8. Show that the solution \( \{u, p\} \) to the system (5.159) is the saddle point of the Lagrangian \( \mathcal{L}_\gamma(v, q) \) from (5.160) for any \( \gamma \geq 0 \), i.e., it holds that
\[ \mathcal{L}_\gamma(u, p) = \min_{v \in \mathbb{R}^n} \max_{q \in \mathbb{R}^m} \mathcal{L}_\gamma(v, q) = \max_{q \in \mathbb{R}^m} \min_{v \in \mathbb{R}^n} \mathcal{L}_\gamma(v, q). \]
Bibliography


Anderson mixing, 18, 33
breakdown, 32
Cauchy integral, 39
Cesaro sums, 82, 97
Chebyshev parameters, 12
Chebyshev polynomials, 11, 43, 49
cluster
general, 80
outlier function, 80, 84
proper, 77
conjugate gradient method, 27, 48
alternative convergence estimates, 60
classic convergence estimates, 49
inexact, 57
preconditioned, 59
stagnation periods, 49
superlinear convergence, 53
convergence factor, 4, 12
decomposition
block ILU, 211
Cholesky, 6, 29
ILU, 182, 186
LU, 3
QR, 3, 26
SVD, 3, 41
delay of convergence, 34
domain decomposition method, 158
alternating Schwarz, 158
Neumann–Neumann preconditioner, 165, 172
nonoverlapping domain decomposition, 161
Poincare–Steklov operator, 163, 164
Schur complement, 164
ellipticity condition, 36
equation
anisotropic diffusion, 176, 185
convection-diffusion, 176, 190
diffusion, 100
Navier–Stokes, 205
of linear elasticity, 176, 201
of Love, 94
Oseen, 224
Poisson, 103, 124, 158
reaction-diffusion, 175, 179
Stokes, 206
Symm’s, 96
error functional, 13, 20
error vector, 4
estimate
Kaniel–Page, 52
fast Fourier transform, 71
field of values, see numerical range
finite element method, 101
iso$P_2–P_1$ element, 208
Crouzeix–Raviart element, 138, 142, 208
inverse inequality, 130
locking phenomena, 202
nonconforming, 138
orthogonality property, 180
Scott–Vogelius element, 202, 204
stiffness matrix, 104, 106, 129
SUPG method, 193
Taylor–Hood element, 208
Wilson element, 139, 142
fixed point problem, 15
formal scalar product, 30
generalized minimal residual method, see minimal residual method
geometric convergence, 4
GMRES, see minimal residual method
Gram–Schmidt procedure, 26
$H^2$-regularity, 128, 135
index of an algorithm, 22
inequality
Friedrichs, 100
Hölder, 19
Harding, 171
Necas, 207
strengthened Cauchy–Schwarz, 148
inner-outer iterations, 220
Jordan normal form, 4, 5
K-condition number, 60, 74
Krylov subspaces, xv, 21, 23
Krylov vectors, xv
LBB condition, 208, 210
leading submatrix, 50
lemma
Lax–Milgram, 101
Reusken’s, 133
logarithmic potentials, 44
loss of orthogonality, 35
matrix
circulant, 69
Fourier image of, 74
cyclic downward shift, 69
diagonal dominant, 20
elliptic, 36
Fourier, 70
Hermitian, 2
Index

Hessenberg, 3
isodiagonal, 67
iteration, 5
Jacobian, 16, 17
\(M\)-matrix, 9
normal, 43
positive definite, 6
sparse, 107, 184
stiffness, see finite element method, stiffness matrix
symmetric, 2
Toeplitz, 67
algebra, 68
generating function, 75
generators, 67
multilevel, 69, 92
symbol, 75
method
Anderson, see Anderson mixing
Arnoldi, 50
biconjugate gradient, 30
CG, see conjugate gradient method
consistent, 5
fixed point, 15
Galerkin, 106
Gauss–Seidel, 8, 148
GMRES, see minimal residual method
inexact Krylov subspace, 55
inexact Richardson, 55
inexact Uzawa, 217
Jacobi, 8, 108, 131, 147
Krylov subspace, 21
Lanczos, 50
linear iterative, 5
minimal residual, see minimal residual method
Newton, 17
PCG, see conjugate gradient method, preconditioned
quasi-minimal residual, 32
relaxation, 8
Richardson, 8
SIMPLE, 215, 229
SOR, 9, 134
steepest descent, 14
symmetric Gauss-Seidel, 8, 131
Uzawa, 220

minimal residual method, 22, 25, 36
algebraic approach, 25
asymptotic convergence factor, 40, 44, 48
Chebyshev polynomials, 43
convergence for diagonalizable matrices, 39
convergence for normal matrices, 43
convergence for positive definite matrices, 37
equivalence to Anderson's mixing, 33
for nonlinear problems, 33
generalized, 22, 25
geometrical approach, 25
loss of orthogonality, 35
nongeneralized, 27
numerical range, 42
pseudospectra, 43
quasi-minimization, 32
with restarts, 37
MINRES, see minimal residual method, nongeneralized
multigrid method, 116
additive, 152
approximation property, 123, 124, 188, 219
as preconditioner, 117
coarse grid operator, 110, 111
complexity, 118
coarse grid correction, 109
F-cycle, 117
iteration matrix, 119
modified approximation property, 200
modified smoothing property, 200
preconditioner, 223
projected system, 21
quadratic functionals, 13
quasi-minimal residual method, 32
Rayleigh quotient, 13, 19
residual functional, 13, 20, 21
residual vector, 6
resolvent, 39, 41
Ritz value, 51
Ritz vector, 51
Robin constant, 44
robust iterative method, 177
robust smoother, 178
roundoff errors, 34
saddle point problem, 207, 210
Schur complement, 210, 220
preconditioner, 223
Index

singular perturbed problem, 175
singular value decomposition, 3
singular values, 3
smoothing iterations
  block Gauss–Seidel, 187
  distributive, 214
  ILU, 182
  of Braess and Sarazin, 215
spectral radius, 2
spectrum, 2
stencil, 192
subspace equation, 146
superlinear convergence, 53, 78
theorem
  Bendixon’s, 227
  interlacing of eigenvalues, 77
  Toeplitz–Hausdorff, 41
  van der Sluis–van der Vorst, 54
triangulation, 104
  minimal angle condition, 104, 106
upwind differences, 191
weak formulation of a PDE, 104,
  128, 139, 180, 207
weak solution of a PDE, 101, 207
Wiener class, 76