PRECONDITIONED ITERATIVE METHODS FOR LINEAR SYSTEMS,

EIGENVALUE AND SINGULAR VALUE PROBLEMS

by

Eugene Vecharynski

M.S., Belarus State University, 2006

A thesis submitted to the

University of Colorado Denver

in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Applied Mathematics

2010

This thesis for the Doctor of Philosophy

degree by

Eugene Vecharynski

has been approved

by

———————————————————————

Andrew Knyazev

———————————————————————

Merico Argentati

———————————————————————

Michele Benzi

———————————————————————

Julien Langou

———————————————————————

Jan Mandel

———————————————————
Date

Vecharynski, Eugene (Ph.D., Applied Mathematics)

Preconditioned Iterative Methods for Linear Systems, Eigenvalue and Singular Value Problems

Thesis directed by Professor Andrew Knyazev

ABSTRACT

In the present dissertation we consider three crucial problems of numerical linear algebra: solution of a linear system, an eigenvalue, and a singular value problem. We focus on the solution methods which are iterative by their nature, matrix-free, preconditioned and require a fixed amount of computational work per iteration. In particular, this manuscript aims to contribute to the areas of research related to the convergence theory of the restarted Krylov subspace minimal residual methods, preconditioning for symmetric indefinite linear systems, approximation of interior eigenpairs of symmetric operators, and preconditioned singular value computations.

We first consider solving non-Hermitian linear systems with the restarted generalized minimal residual method (GMRES). We prove that the cycle-convergence of the method applied to a system of linear equations with a normal (preconditioned) coefficient matrix is sublinear. In the general case, however, it is shown that any admissible cycle-convergence behavior is possible for the restarted GMRES at a number of initial cycles, moreover the spectrum of the coefficient matrix alone does not determine this cycle-convergence.

Next we shift our attention to iterative methods for solving symmetric indefinite systems of linear equations with symmetric positive definite preconditioners. We describe a hierarchy of such methods, from a stationary iteration to the optimal Krylov subspace preconditioned minimal residual method, and suggest a preconditioning strategy based on an approximation of the inverse of the absolute value of the coefficient matrix (absolute value preconditioners). We present an example of a simple (geometric) multigrid absolute value preconditioner for the symmetric model problem of the discretized real Helmholtz (shifted Laplacian) equation in two spatial dimensions with a relatively low wavenumber.

We extend the ideas underlying the methods for solving symmetric indefinite linear systems to the problem of computing an interior eigenpair of a symmetric operator. We present a method that we call the Preconditioned Locally Minimal Residual method (PLMR), which represents a technique for finding an eigenpair corresponding to the smallest, in the absolute value, eigenvalue of a (generalized) symmetric matrix pencil. The method is based on the idea of the refined extraction procedure, performed in the preconditioner-based inner product over four-dimensional trial subspaces, and relies on the choice of the (symmetric positive definite) absolute value preconditioner.

Finally, we consider the problem of finding a singular triplet of a matrix. We suggest a preconditioned iterative method called PLMR-SVD for computing a singular triplet corresponding to the smallest singular value, and introduce preconditioning for the problem. At each iteration, the method extracts approximations for the right and left singular vectors from two separate four-dimensional trial subspaces by solving small quadratically constrained quadratic programs.

We illustrate the performance of the method on the example of the model problem of finding the singular triplet corresponding to the smallest singular value of a gradient operator discretized over a two-dimensional domain. We construct a simple multigrid preconditioner for this problem.

This abstract accurately represents the content of the candidate's thesis. I recommend its publication.

Signed _____

Andrew Knyazev

DEDICATION

To my family and friends.

ACKNOWLEDGMENT

CONTENTS

FIGURES

TABLES

1. Introduction

Complex numerical simulations and solutions of mathematical problems on large-scale data sets have become the routine tasks in cutting edge research and industry, resulting in a broad variety of computational algorithms. The nature of the algorithms can be very diverse, however, often their efficiency and robustness rely, ultimately, on the underlying techniques for solving basic problems of numerical linear algebra.

In this work, we consider numerical solution of linear systems, eigenvalue problems, and singular value problems; see, e.g., [42]. We assume that the problems are of an extremely large size, and possibly sparse, i.e., the involved coefficient matrices contain a significant number of zero entries. The *exact* solutions of such problems are rarely needed. Instead, it is desirable to construct computationally inexpensive *approximations* to the exact solutions. In this context, the use of iterative methods, see, e.g., [3, 33, 59, 56], may be the only option. The study of theoretical and practical aspects of several iterative methods, as well as the introduction of novel iterative techniques for solving the above mentioned problems constitutes the core of the present dissertation.

The methods that we consider in this work share a number of common characteristics. First, their mathematical formulations are based on *short-term recurrent* relations, which allows constructing solvers with a fixed amount of computational work and storage per iteration. Second, the methods are *preconditioned*, i.e., they can use auxiliary operators, called preconditioners, which, if

properly defined, significantly improve the convergence, and, ideally, only modestly affect the cost of each iteration.

In the current manuscript, we address a set of computationally challenging problems, such as numerical solution of symmetric indefinite and nonsymmetric linear systems, computation of interior eigenpairs of symmetric matrix pencils, and finding the smallest singular triplets of general matrices. Our main results concern the convergence theory of the restarted Krylov subspace minimal residual methods, novel preconditioning strategies for symmetric indefinite linear systems and eigenvalue problems, as well as the extension of the concept of preconditioning to singular value problems.

In Chapter 2, we consider the restarted generalized minimal residual method (GMRES) for non-Hermitian linear systems. We prove that the cycle-convergence of the method applied to a system of linear equations with a normal (preconditioned) coefficient matrix is sublinear. In the general case, however, it is shown that any admissible cycle-convergence behavior is possible for the restarted GMRES at a number of initial cycles, moreover the spectrum of the coefficient matrix alone does not determine this cycle-convergence. The results of this chapter are mostly published in [77, 76].

In Chapters 3, 4, and 5, we consider iterative solution of symmetric indefinite systems, symmetric eigenvalue, and singular value problems, respectively. The material is presented in such a way that we can emphasize the interconnections between the problems, which allows us to treat their numerical solution within a unified approach. The obtained results, presented in the chapters, appear here for the first time. We note that the choice of the real vector spaces has been

motivated merely by the desire to simplify the presentation. The extension to the complex case is straightforward.

In Chapter 3, first, we describe a hierarchy of methods for solving symmetric indefinite linear systems with symmetric positive definite (SPD) preconditioners. These methods are, mainly, based on the known idea of the minimization of the residual in the preconditioner-based norm. The careful study of such methods is motivated by a search of appropriate iterative schemes, which can be extended to the problems of finding interior eigenpairs of symmetric operators, as well as computing the smallest singular triplets of general matrices. For example, we describe a method, which can be viewed as a natural analogue of the preconditioned steepest descent algorithm for solving SPD systems, and is the simplest proved to be convergent residual-minimizing method for solving symmetric indefinite systems with an SPD preconditioner. We use the locally optimal accelerations of this method to construct the base scheme, which is further extended to eigenvalue and singular value problems.

Second, in Chapter 3, we suggest a novel preconditioning strategy, which is based on the idea of approximating the inverse of the absolute value of the coefficient matrix. We call preconditioners, which are obtained using this strategy, the absolute value preconditioners. We show, for a model problem of the discretized real Helmholtz (shifted Laplacian) equation in two spatial dimensions with a relatively low wavenumber, that such preconditioners can be efficiently constructed, e.g., in the multigrid framework. It is significant that the same preconditioners can be used for finding interior eigenpairs of symmetric matrix pencils, if applied within the scheme described in Chapter 4. The absolute value

preconditioners for symmetric indefinite systems also motivate the definition of preconditioners for the singular value problems in Chapter 5.

Using the results of Chapter 3, in Chapter 4, we present a new method, that we call the Preconditioned Locally Minimal Residual method (PLMR), which represents a technique for finding an eigenpair corresponding to the smallest, in the absolute value, eigenvalue of a (generalized) symmetric matrix pencil. The method is based on the idea of the refined extraction procedure, also called the refined projection procedure, performed in the preconditioner-based inner product over four-dimensional trial subspaces, and relies on the choice of the (SPD) absolute value preconditioner. We applied the described technique to the model problem of finding an eigenpair of the two-dimensional discretized Laplace operator, which corresponds to the eigenvalue, closest to a given shift. The method demonstrated a satisfactory convergence behavior, with the convergence rate comparable, at a number of initial steps, to that of the optimal preconditioned minimal residual method, applied to the problem of finding the corresponding null space (eigenspace) of the shifted Laplacian.

Finally, in Chapter 5, we consider the problem of finding a singular triplet of a matrix. We suggest a new preconditioned iterative method, that we refer to as PLMR-SVD, for computing a singular triplet corresponding to the smallest singular value. The method has several important features. First, at every step, it uses a pair of separate four-dimensional trial subspaces for extracting the right and left singular vectors, respectively. Second, it admits *two* SPD preconditioners, designed specifically for a singular value problem. We show that even the proper choice of only one of the two preconditioners can result in

4

a significantly improved convergence behavior. As a model problem we consider computing a singular triplet, corresponding to the smallest singular value, of a discrete two-dimensional gradient operator. We present a simple construction of the multigrid preconditioner for this problem.

Let us summarize the main results obtained within the scope of the present dissertation: we have proved two theoretical results which concern the convergence theory of the restarted GMRES algorithm, introduced a new preconditioning strategy for symmetric indefinite linear systems, suggested a novel preconditioned method for computing interior eigenpairs of symmetric matrix pencils, and described a preconditioned method for finding the smallest singular triplets.

2. Convergence of the restarted GMRES

The *generalized minimal residual method* (GMRES) was originally introduced by Saad and Schultz [61] in 1986, and has become a popular method for solving non-Hermitian systems of linear equations,

$$Ax = b, \quad A \in \mathbb{C}^{n \times n}, \quad b \in \mathbb{C}^n. \tag{2.1}$$

Without loss of generality, to simplify the presentation below, we assume that system (2.1) is already preconditioned.

GMRES is classified as a Krylov subspace (projection) iterative method. At every new iteration i, GMRES constructs an approximation $x^{(i)} \in x^{(0)} + \mathcal{K}_i\left(A, r^{(0)}\right)$ to the exact solution of (2.1) such that the 2-norm of the corresponding residual vector $r^{(i)} = b - Ax^{(i)}$ is minimized over the affine space $r^{(0)} + A\mathcal{K}_i\left(A, r^{(0)}\right)$, i.e.,

$$r^{(i)} = \min_{u \in A\mathcal{K}_i\left(A, r^{(0)}\right)} \|r^{(0)} - u\|, \tag{2.2}$$

where $\mathcal{K}_i\left(A, r^{(0)}\right)$ is the i-dimensional Krylov subspace

$$\mathcal{K}_i\left(A, r^{(0)}\right) = \mathrm{span}\{r^{(0)}, Ar^{(0)}, \dots, A^{i-1}r^{(0)}\}$$

induced by the matrix A and the initial residual vector $r^{(0)} = b - Ax^{(0)}$ with $x^{(0)}$ being an initial approximate solution of (2.1),

$$A\mathcal{K}_i\left(A, r^{(0)}\right) = \mathrm{span}\{Ar^{(0)}, A^2r^{(0)}, \dots, A^ir^{(0)}\}.$$

As usual, in a linear setting, a notion of minimality is associated with some orthogonality condition. In our case, minimization (2.2) is equivalent to forcing the new residual vector $r^{(i)}$ to be orthogonal to the subspace $A\mathcal{K}_i\left(A, r^{(0)}\right)$ (also known as the Krylov residual subspace). In practice, for a large problem size, the latter orthogonality condition results in a costly procedure of orthogonalization against the expanding Krylov residual subspace. Orthogonalization together with storage requirement makes the GMRES method complexity and storage prohibitive for practical application. A straightforward treatment for this complication is the so-called restarted GMRES [61].

The *restarted GMRES*, or GMRES(m), is based on restarting GMRES after every m iterations. At each restart, we use the latest approximate solution as the initial approximation for the next GMRES run. Within this framework a single run of m GMRES iterations is called a GMRES(m) cycle, and m is called the restart parameter. Consequently, restarted GMRES can be regarded as a sequence of GMRES(m) cycles. When the convergence happens without any restart occurring, the algorithm is known as the *full GMRES*.

In the context of restarted GMRES, our interest will shift towards the residual vectors $r^{(k)}$ at the end of every kth GMRES(m) cycle (as opposed to the residual vectors $r^{(i)}$ (2.2) obtained at each iteration of the algorithm).

Definition 2.1 (cycle-convergence) *We define the cycle-convergence of the restarted GMRES(m) to be the convergence of the residual norms $\|r^{(k)}\|$, where, for each k, $r^{(k)}$ is the residual at the end of the kth GMRES(m) cycle.*

GMRES(m) constructs approximations $x^{(k)} \in x^{(k-1)} + \mathcal{K}_m\left(A, r^{(k-1)}\right)$ to the exact solution of (2.1) such that each residual vector $r^{(k)} = b - Ax^{(k)}$ satisfies

the local minimality condition

$$r^{(k)} = \min_{u \in A\mathcal{K}_m\left(A, r^{(k-1)}\right)} \|r^{(k-1)} - u\|, \qquad (2.3)$$

where $\mathcal{K}_m\left(A, r^{(k-1)}\right)$ is the m-dimensional Krylov subspace produced at the kth GMRES(m) cycle,

$$\mathcal{K}_m\left(A, r^{(k-1)}\right) = \mathrm{span}\{r^{(k-1)}, Ar^{(k-1)}, \ldots, A^{m-1}r^{(k-1)}\}, \qquad (2.4)$$

$A\mathcal{K}_m\left(A, r^{(k-1)}\right) = \mathrm{span}\{Ar^{(k-1)}, A^2 r^{(k-1)}, \ldots, A^m r^{(k-1)}\}$ is the corresponding Krylov residual subspace.

The price paid for the reduction of the computational work in GMRES(m) is the loss of global optimality in (2.2). Although (2.3) implies a monotonic decrease of the norms of the residual vectors $r^{(k)}$, GMRES(m) can stagnate [61, 80]. This is in contrast with the full GMRES which is guaranteed to converge to the exact solution of (2.1) in n steps (assuming exact arithmetic and nonsingular A). In practice, however, if n is too large, proper choices of a preconditioner and a restart parameter, e.g., [25, 26, 46], can significantly accelerate the convergence of GMRES(m), thus making the method attractive for applications.

While a great deal of effort has been devoted to the characterization of the convergence of the full GMRES, e.g., [74, 21, 34, 35, 43, 70, 72], our understanding of the behavior of GMRES(m) is far from complete, leaving us with more questions than answers, e.g., [25].

For a few classes of matrices, convergence estimates are available for the *restarted* GMRES and/or the *full* GMRES. For example, for real positive definite matrices (that is, for matrices A for which $H = (A + A^T)/2$ is symmetric positive

definite, or, equivalently, for matrices A for which $x^T A x > 0$ for any nonzero $x \in \mathbb{R}^n$), the Elman's bound [22, 23, 33, 61] can be stated as follows

$$\|r^{(k)}\|^2 \le (1 - \rho)^k \|r^{(0)}\|^2 \quad \text{where } 0 < \rho \equiv (\lambda_{min}(H)/\|A\|)^2 \le 1.$$

The latter guarantees the linear cycle-convergence of GMRES(m) for a positive definite matrix. Improvements and generalizations of this bound can be found in [8, 63, 82].

For normal matrices the convergence of the *full* GMRES is well studied. In particular, the convergence is known to be governed solely by the spectrum of A [62, 74]. In Section 2.1 of this manuscript, we prove that the cycle-convergence of *restarted* GMRES for normal matrices is sublinear. This statement means that, for normal matrices, the reduction in the norm of the residual vector at the current GMRES(m) cycle cannot be better than the reduction at the previous cycle. We would like to mention the simultaneous but independent work [5], where the authors present an alternative proof of the sublinear convergence of the restarted GMRES for normal matrices.

Assuming that the coefficient matrix A is diagonalizable, some characterizations of the convergence of the *full* GMRES rely on the condition number of the eigenbasis [74]. Other characterizations of the convergence of the *full* GMRES rely on pseudospectra [52]. More commonly, the field of values is used [8, 22, 23, 33, 61, 63, 82]. A discussion on how descriptive some of these bounds are is given by Embree [24].

In the general case, for the *full* GMRES, the following theorem shows that we cannot prove convergence results based only on the spectrum of the coefficient matrix alone.

Theorem 2.2 (Greenbaum, Pták, and Strakoš, 1996, [34]) *Given a non-increasing positive sequence $f(0) \geq f(1) \geq \cdots \geq f(n-1) > 0$, there exists an n-by-n matrix A and a vector $r^{(0)}$ with $\|r^{(0)}\| = f(0)$ such that $f(i) = \|r^{(i)}\|$, $i = 1, \ldots, n-1$, where $r^{(i)}$ is the residual at step i of the GMRES algorithm applied to the linear system $Ax = b$, with initial residual $r^{(0)} = b - Ax^{(0)}$. Moreover, the matrix A can be chosen to have any desired (nonzero) eigenvalues.*

This result states that, in general, eigenvalues alone do not determine the convergence of the *full* GMRES. A complete description of the set of all pairs $\{A, b\}$ for which the *full* GMRES applied to (2.1) generates the prescribed convergence curve while the matrix A has any (nonzero) eigenvalues, is given in [2].

In Section 2.2, we show that any admissible cycle-convergence behavior is possible for *restarted* GMRES at a number of initial cycles, moreover the spectrum of the coefficient matrix alone does not determine this cycle-convergence. The latter can be viewed as an extension of the result of Greenbaum, Pták, and Strakoš, given by Theorem 2.2, for the case of restarted GMRES.

2.1 The sublinear cycle-convergence of GMRES(m) for normal matrices

Throughout this section we assume (unless otherwise explicitly stated) A to be nonsingular and normal, i.e., A admits the decomposition

$$A = V\Lambda V^*, \tag{2.5}$$

where $\Lambda \in \mathbb{C}^{n \times n}$ is a diagonal matrix with the diagonal elements being the nonzero eigenvalues of A, and $V \in \mathbb{C}^{n \times n}$ is a unitary matrix of the corresponding eigenvectors. $\|\cdot\|$ denotes the 2-norm throughout.

2.1.1 Krylov matrix, its pseudoinverse, and spectral factorization

For a given restart parameter m $(1 \leq m \leq n-1)$, let us denote the kth cycle of GMRES(m) applied to system (2.1), with the initial residual vector $r^{(k-1)}$ as GMRES(A, m, $r^{(k-1)}$). We assume that the residual vector $r^{(k)}$, produced at the end of GMRES(A, m, $r^{(k-1)}$), is nonzero.

A run of GMRES(A, m, $r^{(k-1)}$) generates the Krylov subspace $\mathcal{K}_m\left(A, r^{(k-1)}\right)$ given in (2.4). For each $\mathcal{K}_m\left(A, r^{(k-1)}\right)$ we define a matrix

$$K\left(A, r^{(k-1)}\right) = \begin{bmatrix} r^{(k-1)} & Ar^{(k-1)} & \cdots & A^m r^{(k-1)} \end{bmatrix} \in \mathbb{C}^{n \times (m+1)}, \qquad (2.6)$$

where $k = 1, 2, \ldots, q$, and q is the total number of GMRES(m) cycles.

Matrix (2.6) is called the Krylov matrix. We say that $K\left(A, r^{(k-1)}\right)$ corresponds to the cycle GMRES(A, m, $r^{(k-1)}$). Note that the columns of $K\left(A, r^{(k-1)}\right)$ span the next $(m+1)$-dimensional Krylov subspace $\mathcal{K}_{m+1}(A, r^{(k-1)})$. According to (2.3), the assumption $r^{(k)} \neq 0$ implies that $r^{(k-1)}$ cannot be expressed as a linear combination of vectors in $A\mathcal{K}_m\left(A, r^{(k-1)}\right)$. Thus, the matrix $K\left(A, r^{(k-1)}\right)$ in (2.6) is of the full rank,

$$\text{rank}\left(K\left(A, r^{(k-1)}\right)\right) = m + 1.$$

This equality allows us to introduce the Moore–Penrose pseudoinverse of the matrix $K\left(A, r^{(k-1)}\right)$, i.e.,

$$K^{\dagger}\left(A, r^{(k-1)}\right) = \left(K^*\left(A, r^{(k-1)}\right) K\left(A, r^{(k-1)}\right)\right)^{-1} K^*\left(A, r^{(k-1)}\right) \in \mathbb{C}^{(m+1) \times n},$$

which is well-defined and unique. The following lemma shows that the first column of $\left(K^{\dagger}\left(A, r^{(k-1)}\right)\right)^*$ is the next residual vector $r^{(k)}$ up to a scaling factor.

Lemma 2.3 *Given $A \in \mathbb{C}^{n \times n}$ (not necessarily normal), for any $k = 1, 2, \ldots, q$, we have*

$$\left(K^\dagger \left(A, r^{(k-1)}\right)\right)^* e_1 = \frac{1}{\|r^{(k)}\|^2} r^{(k)}, \tag{2.7}$$

where $e_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}^T \in \mathbb{R}^{m+1}$.

Proof: See Ipsen [43, Theorem 2.1], as well as [17, 65]. ∎

Another important ingredient, first described in [43], is the so-called spectral factorization of the Krylov matrix $K\left(A, r^{(k-1)}\right)$. This factorization is made of three components that encapsulate separately the information on eigenvalues of A, its eigenvectors, and the previous residual vector $r^{(k-1)}$.

Lemma 2.4 *Let $A \in \mathbb{C}^{n \times n}$ satisfy (2.5). Then the Krylov matrix $K\left(A, r^{(k-1)}\right)$, for any $k = 1, 2, \ldots q$, can be factorized as*

$$K\left(A, r^{(k-1)}\right) = V D_{k-1} Z, \tag{2.8}$$

where $d_{k-1} = V^* r^{(k-1)} \in \mathbb{C}^n$, $D_{k-1} = \operatorname{diag}\left(d_{k-1}\right) \in \mathbb{C}^{n \times n}$, and $Z \in \mathbb{C}^{n \times (m+1)}$ is the Vandermonde matrix computed from the eigenvalues of A,

$$Z = \begin{bmatrix} e & \Lambda e & \cdots & \Lambda^m e \end{bmatrix}, \tag{2.9}$$

$e = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}^T \in \mathbb{R}^n$.

Proof: Starting from (2.5) and the definition of the Krylov matrix (2.6),

$$K\left(A, r^{(k-1)}\right) = \begin{bmatrix} r^{(k-1)} & Ar^{(k-1)} & \cdots & A^m r^{(k-1)} \end{bmatrix}$$

$$= \begin{bmatrix} VV^* r^{(k-1)} & V\Lambda V^* r^{(k-1)} & \cdots & V\Lambda^m V^* r^{(k-1)} \end{bmatrix}$$

$$= V \begin{bmatrix} d_{k-1} & \Lambda d_{k-1} & \cdots & \Lambda^m d_{k-1} \end{bmatrix}$$

$$= V \begin{bmatrix} D_{k-1}e & \Lambda D_{k-1}e & \cdots & \Lambda^m D_{k-1}e \end{bmatrix}$$

$$= V D_{k-1} \begin{bmatrix} e & \Lambda e & \cdots & \Lambda^m e \end{bmatrix} = V D_{k-1} Z.$$

∎

It is clear that the statement of Lemma 2.4 can be easily generalized to the case of a diagonalizable (nonnormal) matrix A providing that we define $d_{k-1} = V^{-1} r^{(k-1)}$ in the lemma.

2.1.2 The sublinear cycle-convergence of GMRES(m)

Along with (2.1) let us consider the system

$$A^* x = b \tag{2.10}$$

with the matrix A replaced by its conjugate transpose. Clearly, according to (2.5),

$$A^* = V \overline{\Lambda} V^*. \tag{2.11}$$

It turns out that m steps of GMRES applied to systems (2.1) and (2.10) produce residual vectors of equal norms at each step—provided that the initial residual vector is identical. This observation is crucial for proving the sublinear cycle-convergence of GMRES(m) and is formalized in the following lemma.

Lemma 2.5 *Assume that $A \in \mathbb{C}^{n \times n}$ is a nonsingular normal matrix. Let $r^{(m)}$ and $\hat{r}^{(m)}$ be the nonzero residual vectors obtained by applying m steps of GMRES*

13

to systems (2.1) and (2.10); $1 \leq m \leq n - 1$. Then

$$\left\| r^{(m)} \right\| = \left\| \hat{r}^{(m)} \right\|,$$

provided that the initial approximate solutions of (2.1) and (2.10) induce the same initial residual vector $r^{(0)}$.

Moreover, if $p_m(z)$ and $q_m(z)$ are the (GMRES) polynomials which mini-mize, respectively, $\|p(A)r^{(0)}\|$ and $\|p(A^*)r^{(0)}\|$ over $p(z) \in \mathcal{P}_m$, where $\mathcal{P}_m$ is the set of all polynomials of degree at most m defined on the complex plane such that $p(0) = 1$, then

$$\overline{p}_m(z) = q_m(z),$$

where $\overline{p}(z)$ denotes the polynomial obtained from $p(z) \in \mathcal{P}_m$ by the complex conjugation of its coefficients.

Proof: Let us consider a polynomial $p(z) \in \mathcal{P}_m$. Let $r^{(0)}$ be a nonzero initial residual vector for systems (2.1) and (2.10) simultaneously. Since the matrix A is normal, so is $p(A)$; thus $p(A)$ commutes with its conjugate transpose $p^*(A)$. We have

$$
\begin{aligned}
\|p(A)r^{(0)}\|^2 &= (p(A)r^{(0)}, p(A)r^{(0)}) = (r^{(0)}, p^*(A)p(A)r^{(0)}) \\
&= (r^{(0)}, p(A)p^*(A)r^{(0)}) = (p^*(A)r^{(0)}, p^*(A)r^{(0)}) \\
&= ((Vp(\Lambda)V^*)^* r^{(0)}, (Vp(\Lambda)V^*)^* r^{(0)}) \\
&= (V\overline{p}(\overline{\Lambda})V^*r^{(0)}, V\overline{p}(\overline{\Lambda})V^*r^{(0)}) \\
&= (\overline{p}(V\overline{\Lambda}V^*)r^{(0)}, \overline{p}(V\overline{\Lambda}V^*)r^{(0)}) = \|\overline{p}(V\overline{\Lambda}V^*)r^{(0)}\|^2,
\end{aligned}
$$

14

where $\bar{p}(z)$ is the polynomial obtained from $p(z)$ by conjugating its coefficients. By (2.11) we conclude that

$$\|p(A)r^{(0)}\| = \|\bar{p}(A^*)r^{(0)}\|.$$

We note that the last statement is true for any polynomial p, for any r_0, and for any normal A.

Now, let us look at $\|r^{(m)}\|$ and $\|\hat{r}^{(m)}\|$. On the one hand,

$$\|r^{(m)}\| = \min_{p \in \mathcal{P}_m} \|p(A)r^{(0)}\| = \|p_m(A)r^{(0)}\| = \|\bar{p}_m(A^*)r^{(0)}\|$$
$$\geq \min_{p \in \mathcal{P}_m} \|\bar{p}(A^*)r^{(0)}\| = \min_{p \in \mathcal{P}_m} \|p(A^*)r^{(0)}\| = \|\hat{r}^{(m)}\|.$$

On the other hand,

$$\|\hat{r}^{(m)}\| = \min_{p \in \mathcal{P}_m} \|p(A^*)r^{(0)}\| = \|q_m(A^*)r^{(0)}\| = \|\bar{q}_m(A)r^{(0)}\|$$
$$\geq \min_{p \in \mathcal{P}_m} \|\bar{p}(A)r^{(0)}\| = \min_{p \in \mathcal{P}_m} \|p(A)r^{(0)}\| = \|r^{(m)}\|.$$

Combining both results, we conclude that

$$\|r^{(m)}\| = \|\hat{r}^{(m)}\|,$$

which proves the first part of the lemma.

To prove the second part of the lemma, we consider the following equalities:

$$\|q_m(A^*)r^{(0)}\| = \min_{p \in \mathcal{P}_m} \|p(A^*)r^{(0)}\| = \|\hat{r}^{(m)}\| = \|r^{(m)}\| = \min_{p \in \mathcal{P}_m} \|p(A)r^{(0)}\|$$
$$= \|p_m(A)r^{(0)}\| = \|\bar{p}_m(A^*)r^{(0)}\|.$$

By uniqueness of the GMRES polynomial [36, Theorem 2], we conclude that $\bar{p}_m(z) = q_m(z)$. $\blacksquare$

The previous lemma is a general result for the full GMRES, which states that, given a nonsingular normal matrix A and an initial residual vector $r^{(0)}$, GMRES applied to A with $r^{(0)}$ produces the same convergence curve as GMRES applied to A^* with $r^{(0)}$. In the framework of restarted GMRES, Lemma 2.5 implies that the cycles GMRES$(A, m, r^{(k-1)})$ and GMRES$(A^*, m, r^{(k-1)})$ result in, respectively, residual vectors $r^{(k)}$ and $\hat{r}^{(k)}$ that have the same norm.

Theorem 2.6 (the sublinear cycle-convergence of GMRES(m))

Let $\left\{ r^{(k)} \right\}_{k=0}^{q}$ be a sequence of nonzero residual vectors produced by GMRES(m) applied to system (2.1) with a nonsingular normal matrix $A \in \mathbb{C}^{n \times n}$, $1 \le m \le n - 1$. Then

$$\frac{\|r^{(k)}\|}{\|r^{(k-1)}\|} \le \frac{\|r^{(k+1)}\|}{\|r^{(k)}\|}, \quad k = 1, \ldots, q - 1. \tag{2.12}$$

Proof: Left multiplication of both parts of (2.7) by $K^* \left(A, r^{(k-1)} \right)$ leads to

$$e_1 = \frac{1}{\|r^{(k)}\|^2} K^* \left(A, r^{(k-1)} \right) r^{(k)}.$$

By (2.8) in Lemma 2.4, we factorize the Krylov matrix $K \left(A, r^{(k-1)} \right)$ in the previous equality:

$$e_1 = \frac{1}{\|r^{(k)}\|^2} \left(V D_{k-1} Z \right)^* r^{(k)} = \frac{1}{\|r^{(k)}\|^2} Z^* \overline{D}_{k-1} V^* r^{(k)}$$

$$= \frac{1}{\|r^{(k)}\|^2} Z^* \overline{D}_{k-1} d_k.$$

Applying complex conjugation to this equality (and observing that e_1 is real), we get

$$e_1 = \frac{1}{\|r^{(k)}\|^2} Z^T D_{k-1} \overline{d}_k.$$

According to the definition of D_{k-1} in Lemma 2.4, $D_{k-1} \overline{d}_k = \overline{D}_k d_{k-1}$; thus

$$e_1 = \frac{1}{\|r^{(k)}\|^2} Z^T \overline{D}_k d_{k-1} = \frac{1}{\|r^{(k)}\|^2} \left(Z^T \overline{D}_k V^* \right) r^{(k-1)} = \frac{1}{\|r^{(k)}\|^2} \left(V D_k \overline{Z} \right)^* r^{(k-1)}.$$

From (2.8) and (2.11) we notice that

$$K\left(A^*, r^{(k)}\right) = K\left(V\overline{\Lambda}V^*, r^{(k)}\right) = VD_k\overline{Z},$$

and so therefore

$$e_1 = \frac{1}{\|r^{(k)}\|^2} K^*\left(A^*, r^{(k)}\right) r^{(k-1)}. \tag{2.13}$$

Considering the residual vector $r^{(k-1)}$ as a solution of the underdetermined system (2.13), we can represent the latter as

$$r^{(k-1)} = \|r^{(k)}\|^2 \left(K^*\left(A^*, r^{(k)}\right)\right)^\dagger e_1 + w_k, \tag{2.14}$$

where $w_k \in \text{null}\left(K^*\left(A^*, r^{(k)}\right)\right)$. We note that since $r^{(k+1)}$ is nonzero (assumption in Theorem 2.6), the residual vector $\hat{r}^{(k+1)}$ at the end of the cycle GMRES(A^*, m, $r^{(k)}$) is nonzero as well by Lemma 2.5; hence the corresponding Krylov matrix $K\left(A^*, r^{(k)}\right)$ is of the full rank, and thus the pseudoinverse in (2.14) is well defined. Moreover, since

$$w_k \perp \left(K^*\left(A^*, r^{(k)}\right)\right)^\dagger e_1,$$

using the Pythagorean theorem we obtain

$$\|r^{(k-1)}\|^2 = \|r^{(k)}\|^4 \|\left(K^*\left(A^*, r^{(k)}\right)\right)^\dagger e_1\|^2 + \|w_k\|^2.$$

Now, since $(K^*(A^*, r^{(k)}))^\dagger = (K^\dagger(A^*, r^{(k)})^*)$, we get

$$\|r^{(k-1)}\|^2 = \|r^{(k)}\|^4 \|\left(K^\dagger\left(A^*, r^{(k)}\right)\right)^* e_1\|^2 + \|w_k\|^2,$$

and then by (2.7),

$$= \frac{\|r^{(k)}\|^4}{\|\hat{r}^{(k+1)}\|^2} + \|w_k\|^2$$

$$\geq \frac{\|r^{(k)}\|^4}{\|\hat{r}^{(k+1)}\|^2},$$
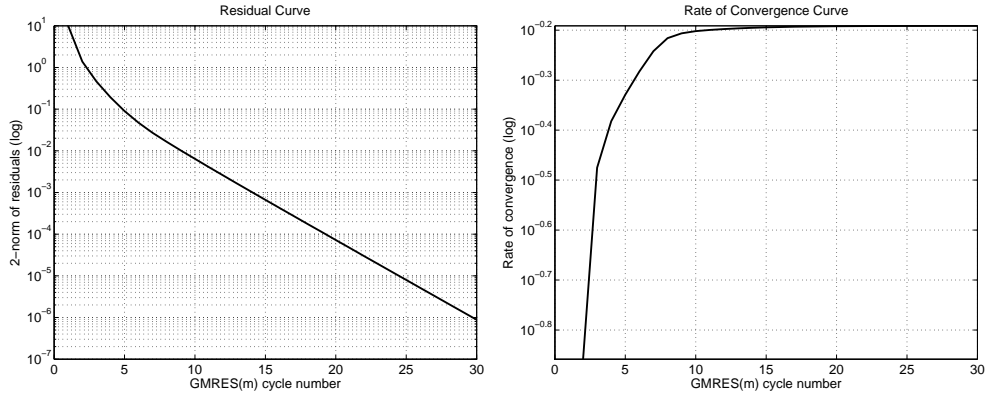
Figure 2.1: Cycle-convergence of GMRES(5) applied to a 100-by-100 normal matrix.

where $\hat{r}^{(k+1)}$ is the residual vector at the end of the cycle GMRES(A^*, m, $r^{(k)}$).

Finally,

$$\frac{\|r^{(k)}\|^2}{\|r^{(k-1)}\|^2} \leq \frac{\|r^{(k)}\|^2 \|\hat{r}^{(k+1)}\|^2}{\|r^{(k)}\|^4} = \frac{\|\hat{r}^{(k+1)}\|^2}{\|r^{(k)}\|^2},$$

so that

$$\frac{\|r^{(k)}\|}{\|r^{(k-1)}\|} \leq \frac{\|\hat{r}^{(k+1)}\|}{\|r^{(k)}\|}. \tag{2.15}$$

By Lemma 2.5, the norm of the residual vector $\hat{r}^{(k+1)}$ at the end of the cycle GMRES(A^*, m, $r^{(k)}$) is equal to the norm of the residual vector $r^{(k+1)}$ at the end of the cycle GMRES(A, m, $r^{(k)}$), which completes the proof of the theorem.

∎

Geometrically, Theorem 2.6 suggests that any residual curve of a restarted GMRES, applied to a system with a nonsingular normal matrix, is nonincreasing and concave up (Figure 2.1).

Corollary 2.7 (cycle-convergence of GMRES(m)) *Let* $\|r^{(0)}\|$ *and* $\|r^{(1)}\|$ *be given. Then, under assumptions of Theorem 2.6, norms of the residual vectors*

$r^{(k)}$ *at the end of each GMRES(m) cycle satisfy the following inequality*

$$\|r^{(k+1)}\| \geq \|r^{(1)}\| \left(\frac{\|r^{(1)}\|}{\|r^{(0)}\|}\right)^k, \quad k = 1, \ldots, q-1. \tag{2.16}$$

Proof: Directly follows from (2.12). ∎

Inequality (2.16) shows that we are able to provide a lower bound for the residual norm at any cycle $k > 1$ after performing only one cycle of GMRES(m), applied to system (2.1) with a nonsingular normal matrix A.

From the proof of Theorem 2.6 it is clear that, for a fixed k, the equality in (2.12) holds if and only if the vector w_k in (2.14) from the null space of the corresponding matrix $K^*\left(A^*, r^{(k)}\right)$ is zero. In particular, if the restart parameter is chosen to be one less than the problem size, i.e., $m = n - 1$, the matrix $K^*\left(A^*, r^{(k)}\right)$ in (2.13) becomes an n-by-n nonsingular matrix, hence with a zero null space, and thus inequality (2.12) is indeed an equality if $m = n - 1$.

We now show that the cycle-convergence of GMRES($n - 1$), applied to system (2.1) with a nonsingular normal matrix A, can be completely determined by norms the of the two initial residual vectors $r^{(0)}$ and $r^{(1)}$.

Corollary 2.8 (the cycle-convergence of GMRES($n - 1$)) *Let us suppose that $\|r^{(0)}\|$ and $\|r^{(1)}\|$ are given. Then, under the assumptions of Theorem 2.6, norms of the residual vectors $r^{(k)}$ at the end of each GMRES(n − 1) cycle obey the following formula:*

$$\|r^{(k+1)}\| = \|r^{(1)}\| \left(\frac{\|r^{(1)}\|}{\|r^{(0)}\|}\right)^k, \quad k = 1, \ldots, q-1. \tag{2.17}$$

Proof: Representation (2.14) of the residual vector $r^{(k-1)}$ for $m = n - 1$ turns into

$$r^{(k-1)} = \|r^{(k)}\|^2 \left(K^*\left(A^*, r^{(k)}\right)\right)^{-1} e_1, \tag{2.18}$$

implying, by the proof of Theorem 2.6, that the equality in (2.12) holds at each GMRES($n - 1$) cycle. Thus,

$$\|r^{(k+1)}\| = \|r^{(k)}\| \frac{\|r^{(k)}\|}{\|r^{(k-1)}\|}, \quad k = 1, \ldots, q - 1.$$

We show (2.17) by induction in k. Using the formula above, it is easy to verify (2.17) for $\|r^{(2)}\|$ and $\|r^{(3)}\|$ ($k = 1, 2$). Let us assume that for some k, $3 \le k \le q - 1$, $\|r^{(k-1)}\|$ and $\|r^{(k)}\|$ can also be computed by (2.17). Then

$$\|r^{(k+1)}\| = \|r^{(k)}\| \frac{\|r^{(k)}\|}{\|r^{(k-1)}\|} = \|r^{(1)}\| \left(\frac{\|r^{(1)}\|}{\|r^{(0)}\|} \right)^{k-1} \frac{\|r^{(1)}\| \left(\frac{\|r^{(1)}\|}{\|r^{(0)}\|} \right)^{k-1}}{\|r^{(1)}\| \left(\frac{\|r^{(1)}\|}{\|r^{(0)}\|} \right)^{k-2}}$$

$$= \|r^{(1)}\| \left(\frac{\|r^{(1)}\|}{\|r^{(0)}\|} \right)^{k-1} \left(\frac{\|r^{(1)}\|}{\|r^{(0)}\|} \right) = \|r^{(1)}\| \left(\frac{\|r^{(1)}\|}{\|r^{(0)}\|} \right)^{k}.$$

Thus, (2.17) holds for all $k = 1, \ldots, q - 1$. ∎

Another observation in the proof of Theorem 2.6 leads to a result from Baker, Jessup, and Manteuffel [6]. In this paper, the authors prove that, if GMRES($n-1$) is applied to a system with Hermitian or skew-Hermitian matrix, the residual vectors at the end of each restart cycle alternate direction in a cyclic fashion [6, Theorem 2]. In the following corollary we (slightly) refine this result by providing the exact expression for the constants α_k in [6, Theorem 2].

Corollary 2.9 (the alternating residuals) *Let $\left\{ r^{(k)} \right\}_{k=0}^{q}$ be a sequence of nonzero residual vectors produced by GMRES($n - 1$) applied to system (2.1) with a nonsingular Hermitian or skew-Hermitian matrix $A \in \mathbb{C}^{n \times n}$. Then*

$$r^{(k+1)} = \alpha_k r^{(k-1)}, \quad \alpha_k = \frac{\left\| r^{(k+1)} \right\|^2}{\left\| r^{(k)} \right\|^2} \in (0, 1], \quad k = 1, 2, \ldots, q - 1. \qquad (2.19)$$

Proof: For the case of a Hermitian matrix A, i.e., $A^* = A$, the proof follows directly from (2.7) and (2.18).

Let A be skew-Hermitian, i.e., $A^* = -A$. Then, by (2.7) and (2.18),

$$r^{(k-1)} = \left(K^* \left(A^*, r^{(k)}\right)\right)^{-1} e_1 = \left(K^* \left(-A, r^{(k)}\right)\right)^{-1} e_1 = \frac{\|r^{(k)}\|^2}{\|\hat{r}^{(k+1)}\|^2} \hat{r}^{(k+1)},$$

where $\hat{r}^{(k+1)}$ is the residual at the end of the cycle GMRES$(-A, n-1, r^{(k)})$.

According to (2.3), the residual vectors $r^{(k+1)}$ and $\hat{r}^{(k+1)}$ at the end of the cycles GMRES$(A, n-1, r^{(k)})$ and GMRES$(-A, n-1, r^{(k)})$ are obtained by orthogonalizing $r^{(k)}$ against the Krylov residual subspaces $A\mathcal{K}_{n-1}\left(A, r^{(k)}\right)$ and $(-A)\mathcal{K}_{n-1}\left(-A, r^{(k)}\right)$, respectively. But $(-A)\mathcal{K}_{n-1}\left(-A, r^{(k)}\right) = A\mathcal{K}_{n-1}\left(A, r^{(k)}\right)$, and hence $\hat{r}^{(k+1)} = r^{(k+1)}$. ∎

In general, for systems with nonnormal matrices, the cycle-convergence behavior of the restarted GMRES is not sublinear. In Figure 2.2, we consider a nonnormal diagonalizable matrix and illustrate the claim. Indeed, for nonnormal matrices, it has been observed that the cycle-convergence of restarted GMRES can be superlinear [81].

In this concluding subsection we restrict our attention to the case of a diagonalizable matrix A,

$$A = V\Lambda V^{-1}, \quad A^* = V^{-*}\overline{\Lambda}V^*. \tag{2.20}$$

The analysis performed in Theorem 2.6 can be generalized for the case of a diagonalizable matrix [79], resulting in inequality (2.15). However, as we depart from normality, Lemma 2.5 fails to hold and the norm of the residual vector $\hat{r}^{(k+1)}$ at the end of the cycle GMRES$(A^*, m, r^{(k)})$ is no longer equal to the norm of the vector $r^{(k+1)}$ at the end of GMRES$(A, m, r^{(k)})$. Moreover, since the eigenvectors
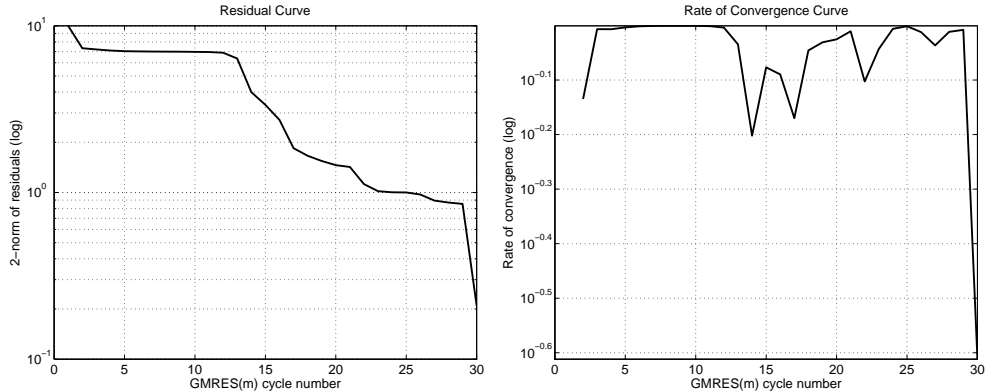
Figure 2.2: Cycle-convergence of GMRES(5) applied to a 100-by-100 diagonalizable (nonnormal) matrix.

of A can be significantly changed by transpose-conjugation, as (2.20) suggests, the matrices A and A^* can have almost nothing in common, so that the norms of $\hat{r}^{(k+1)}$ and $r^{(k+1)}$ are, possibly, far from being equal. This creates an opportunity to break the sublinear convergence of GMRES(m), provided that the subspace $A\mathcal{K}_m\left(A, r^{(k)}\right)$ results in a better approximation (2.3) of the vector $r^{(k)}$ than the subspace $A^*\mathcal{K}_m\left(A^*, r^{(k)}\right)$.

It is natural to expect that the convergence of the restarted GMRES for "almost normal" matrices will be "almost sublinear." We quantify this statement in the following theorem.

Theorem 2.10 *Let $\left\{r^{(k)}\right\}_{k=0}^{q}$ be a sequence of nonzero residual vectors produced by GMRES(m) applied to system (2.1) with a nonsingular diagonalizable matrix $A \in \mathbb{C}^{n \times n}$ as in (2.20), $1 \leq m \leq n - 1$. Then*

$$\frac{\|r^{(k)}\|}{\|r^{(k-1)}\|} \leq \frac{\alpha\left(\|r^{(k+1)}\| + \beta_k\right)}{\|r^{(k)}\|}, \quad k = 1, \ldots, q - 1, \qquad (2.21)$$

where $\alpha = \sigma_{min}^{-2}(V)$, $\beta_k = \|p_k(A)(I - VV^*)r^{(k)}\|$, $p_k(z)$ is the polynomial constructed at the cycle GMRES(A, m, $r^{(k)}$), and where q is the total number of

GMRES(m) cycles. We note that $0 < \alpha \longrightarrow 1$ and $0 < \beta_k \longrightarrow 0$ as $V^*V \longrightarrow I$.

Proof: Let us consider the norm of the residual vector $\hat{r}^{(k+1)}$ at the end of the cycle GMRES(A^*, m, $r^{(k)}$). Then we have

$$\|\hat{r}^{(k+1)}\| = \min_{\hat{p} \in \mathcal{P}_m} \|\hat{p}(A^*)r^{(k)}\| \le \|p(A^*)r^{(k)}\|,$$

where $p(z) \in \mathcal{P}_m$ is any polynomial of degree at most m such that $p(0) = 1$. Then, using (2.20),

$$\begin{aligned}
\|\hat{r}^{(k+1)}\| &\le \|p(A^*)r^{(k)}\| \\
&= \|V^{-*}p(\overline{\Lambda})V^*r^{(k)}\| \\
&= \|V^{-*}p(\overline{\Lambda})(V^{-1}V)V^*r^{(k)}\| \\
&= \|V^{-*}p(\overline{\Lambda})V^{-1}(VV^*)r^{(k)}\| \\
&= \|V^{-*}p(\overline{\Lambda})V^{-1}(I - (I - VV^*))r^{(k)}\| \\
&= \|V^{-*}p(\overline{\Lambda})\left(V^{-1}r_k - V^{-1}(I - VV^*)r^{(k)}\right)\| \\
&\le \|V^{-*}\|\|p(\overline{\Lambda})\left(V^{-1}r^{(k)} - V^{-1}(I - VV^*)r^{(k)}\right)\|.
\end{aligned}$$

We note that

$$\|p(\overline{\Lambda})\left(V^{-1}r^{(k)} - V^{-1}(I - VV^*)r^{(k)}\right)\| = \|\overline{p}(\Lambda)\left(V^{-1}r^{(k)} - V^{-1}(I - VV^*)r^{(k)}\right)\|.$$

Thus,

$$\begin{aligned}
\|\hat{r}^{(k+1)}\| &\le \|V^{-*}\|\|\overline{p}(\Lambda)\left(V^{-1}r^{(k)} - V^{-1}(I - VV^*)r^{(k)}\right)\| \\
&= \|V^{-*}\|\|(V^{-1}V)\overline{p}(\Lambda)\left(V^{-1}r^{(k)} - V^{-1}(I - VV^*)r^{(k)}\right)\| \\
&\le \|V^{-*}\|\|V^{-1}\|\|V\overline{p}(\Lambda)V^{-1}r^{(k)} - V\overline{p}(\Lambda)V^{-1}(I - VV^*)r^{(k)}\| \\
&= \frac{1}{\sigma_{min}^2(V)}\|\overline{p}(V\Lambda V^{-1})r^{(k)} - \overline{p}(V\Lambda V^{-1})(I - VV^*)r^{(k)}\| \\
&\le \frac{1}{\sigma_{min}^2(V)}\left(\|\overline{p}(A)r^{(k)}\| + \|\overline{p}(A)(I - VV^*)r^{(k)}\|\right),
\end{aligned}$$

where σ_{min} is the smallest singular value of V.

Since the last inequality holds for any polynomial $\overline{p}(z) \in \mathcal{P}_m$, it also holds for $\overline{p}(z) = p_k(z)$, where $p_k(z)$ is the polynomial constructed at the cycle GMRES(A, m, $r^{(k)}$). Hence,

$$\|\hat{r}^{(k)}\| \le \frac{1}{\sigma_{min}^2(V)} \left(\|r^{(k+1)}\| + \|p_k(A)(I - VV^*)r^{(k)}\| \right).$$

Setting $\alpha = \frac{1}{\sigma_{min}^2(V)}$, $\beta_k = \|p_k(A)(I - VV^*)r^{(k)}\|$, and observing that $\alpha \longrightarrow 1$, $\beta_k \longrightarrow 0$ as $V^*V \longrightarrow I$, from (2.15) we obtain (2.21). ∎

2.2 Any admissible cycle-convergence behavior is possible for the restarted GMRES at its initial cycles

In the previous section, we have characterized the cycle-convergence of the restarted GMRES applied to system of linear equations (2.1) with a normal coefficient matrix. Now we turn our attention to the general case. The main result of the current section is stated as the following

Theorem 2.11 *Let us be given an integer $n > 0$, a restart parameter m ($0 < m < n$), and a positive sequence $\{f(k)\}_{k=0}^q$, such that $f(0) > f(1) > \cdots > f(s) > 0$, and $f(s) = f(s + 1) = \ldots = f(q)$, where $0 < q < n/m$, $0 \le s \le q$. There exists an n-by-n matrix A and a vector $r^{(0)}$ with $\|r^{(0)}\| = f(0)$, such that $\|r^{(k)}\| = f(k)$, $k = 1, \ldots, q$, where $r^{(k)}$ is the residual at cycle k of restarted GMRES with restart parameter m applied to the linear system $Ax = b$, with initial residual $r^{(0)} = b - Ax^{(0)}$. Moreover, the matrix A can be chosen to have any desired (nonzero) eigenvalues.*

The *full* GMRES has a nonincreasing convergence for any $i \ge 0$, $f(i) \ge f(i + 1)$ and it computes the exact solution in at most n steps ($f(n) = 0$).

We note that the assumptions on $\{f(k)\}_{k=1}^{n-1}$ in Theorem 2.2 do not cover the class of convergence sequences corresponding to the convergence to the exact solution *before* step n. One can see, however, that these assumptions are sufficient to conclude that the theorem holds in this case as well. In this sense it is remarkable that Greenbaum, Pták, and Strakoš are able to prove that any *admissible* convergence behavior is possible for the *full* GMRES at its n steps. At the same time we would like to note that the cycle-convergence of the *restarted* GMRES can have two *admissible* scenarios: either $f(i) > f(i+1)$ for any i, in other words, the cycle-convergence is (strictly) decreasing; or there exists s such that $f(i) > f(i+1)$ for any $i < s$, and then $f(i) = f(s)$ for any $i > s$, in other words, if the restarted GMRES stagnates at cycle $s+1$, it stagnates forever. Thus assumptions on $\{f(k)\}_{k=0}^{q}$ in Theorem 2.11 reflect any *admissible* cycle-convergence behavior of *restarted* GMRES at the first q cycles, except for the case where the convergence to the exact solution happens *within* these q cycles. It turns out that the assumptions are sufficient to guarantee that Theorem 2.11 also holds in the above mentioned case of "early" convergence. In Subsection 2.2.6, we point out how exactly the assumptions of Theorem 2.2 and Theorem 2.11 allow us to conclude that any *admissible* convergence behavior is possible for the *full* and *restarted* GMRES (at its q initial cycles).

As mentioned above, the maximum number of iterations of the *full* GMRES is at most n, and the method delivers the exact solution in a finite number of steps. The *restarted* GMRES, however, may never provide the exact solution. It (hopefully) decreases the residual norm at each cycle, that is, provides a more and more accurate approximation to the exact solution. With n^2 parameters

in A and n parameters in b we are not able to control the convergence for an infinite amount of cycles. For this reason, we consider only the first $q < n/m$ initial GMRES(m) cycles. We note that, in practice, $n >> m$ so q is relatively large.

The rest of this section concerns the proof of Theorem 2.11. The proof we provide is constructive and directly inspired by the article of Greenbaum, Pták, and Strakoš [34]. Although Greenbaum, Pták, and Strakoš laid the path, there are several specific difficulties ahead in the analysis of the *restarted* GMRES.

Let n be a matrix order and m a restart parameter $(m < n)$, $\Lambda = \{\lambda_1, \lambda_2, \ldots \lambda_n\} \subset \mathbb{C} \setminus \{0\}$ be a set of n nonzero complex numbers, and $\{f(k)\}_{k=0}^{q}$ be a positive sequence, such that $f(0) > f(1) > \cdots > f(s) > 0$ and $f(s) = f(s+1) = \ldots = f(q)$, where $0 < q < n/m$, $0 \leq s \leq q$.

In this section we construct a matrix $A \in \mathbb{C}^{n \times n}$ and an initial residual vector $r^{(0)} = b - Ax^{(0)} \in \mathbb{C}^n$ such that GMRES(m) applied to system (2.1) with the initial approximate solution $x^{(0)}$, produces a sequence $\{x^{(k)}\}_{k=1}^{q}$ of approximate solutions with corresponding residual vectors $\{r^{(k)}\}_{k=0}^{q}$ having the prescribed norms: $\|r^{(k)}\| = f(k)$. Moreover the spectrum of A is Λ.

For clarity, we first restrict our attention to the case of the strictly decreasing cycle-convergence, and, in Section 2.2.2, prove Theorem 2.11 under the assumption that $f(0) > f(1) > \cdots > f(q) > 0$ (i.e., we assume that $s = q$). Next, in Section 2.2.3, we complete the proof of Theorem 2.11 by handling the (remaining) case of stagnation, i.e., $f(0) > f(1) > \cdots > f(s) > 0$ and $f(s) = f(s+1) = \ldots = f(q)$, $0 \leq s < q$. This is done by a slight change in the proof for the considered case of the strictly decreasing cycle-convergence.

2.2.1 Outline of the proof of Theorem 2.11

The general approach described in this paper is similar to the approach of Greenbaum, Pták, and Strakoš [34]: we fix an initial residual vector, construct an appropriate basis of $\mathbb{C}^n$, and use this basis to define a linear operator $\mathcal{A}$. This operator is represented by the matrix A in the canonical basis. It has the prescribed spectrum and provides the desired cycle-convergence at the first q cycles of GMRES(m). However, the presence of restarts somewhat complicates the construction: the choice of the basis vectors, as well as the structure of the resulting operator $\mathcal{A}$, becomes less transparent. Below we briefly describe our *three-step construction for the case of the strictly decreasing cycle-convergence* and then suggest its easy modification to prove the general case, which includes stagnation.

At the *first step* we construct q sets of vectors $\mathcal{W}_m^{(k)} = \{w_1^{(k)}, \ldots, w_m^{(k)}\}$, $k = 1, \ldots, q$, each set $\mathcal{W}_m^{(k)}$ is the orthonormal basis of the Krylov residual subspace $A\mathcal{K}_m\left(A, r^{(k-1)}\right)$ generated at the k-th GMRES(m) cycle such that

$$\text{span } \mathcal{W}_j^{(k)} = A\mathcal{K}_j\left(A, r^{(k-1)}\right), \quad j = 1, \ldots, m. \tag{2.22}$$

The orthonormal basis $\mathcal{W}_m^{(k)}$ needs to be chosen in order to generate residual vectors $r^{(k)}$ with the prescribed (strictly decreasing) norms $f(k)$ at the end of each cycle subject to the additional requirement that the set of $mq + 1(\leq n)$ vectors

$$\overline{\mathcal{S}} = \{r^{(0)}, w_1^{(1)}, \ldots, w_{m-1}^{(1)}, r^{(1)}, w_1^{(2)}, \ldots, w_{m-1}^{(2)}, \ldots, r^{(q-1)}, w_1^{(q)}, \ldots, w_{m-1}^{(q)}, r^{(q)}\}$$
$$\tag{2.23}$$

is linearly independent.

Once we have the set $\overline{\mathcal{S}}$, we will complete it to have a basis for $\mathbb{C}^n$. If the number of vectors in $\overline{\mathcal{S}}$ is less than n, a basis $\mathcal{S}$ of $\mathbb{C}^n$ is obtained by completion of $\overline{\mathcal{S}}$ with a set $\widehat{\mathcal{S}}$ of $n - mq - 1$ vectors, i.e., $\mathcal{S} = \{\overline{\mathcal{S}}, \widehat{\mathcal{S}}\}$. This will provide a representation of $\mathbb{C}^n$ as the direct sum

$$\mathbb{C}^n = \operatorname{span} \mathcal{S} = \operatorname{span}\{r^{(0)}, \mathcal{W}_{m-1}^{(1)}\} \oplus \cdots \oplus \operatorname{span}\{r^{(q-1)}, \mathcal{W}_{m-1}^{(q)}\} \oplus \operatorname{span}\{r^{(q)}, \widehat{\mathcal{S}}\}.$$
(2.24)

The latter translates in terms of Krylov subspaces into

$$\mathbb{C}^n = \operatorname{span} \mathcal{S} = \mathcal{K}_m\left(A, r^{(0)}\right) \oplus \cdots \oplus \mathcal{K}_m\left(A, r^{(q-1)}\right) \oplus \operatorname{span}\{r^{(q)}, \widehat{\mathcal{S}}\}.$$

At the *second step* of our construction, we define a linear operator $\mathcal{A}$: $\mathbb{C}^n \longrightarrow \mathbb{C}^n$ with spectrum Λ which generates the Krylov residual subspaces in (2.22) at each GMRES(m) cycle, by its action on the basis vectors $\mathcal{S}$, such that the desired matrix A is the operator $\mathcal{A}$'s representation in the canonical basis. The *third step* accomplishes the construction by a similarity transformation.

In the following subsection we show that this three-step approach indeed allows us to prove Theorem 2.11 in the case of a strictly decreasing positive sequence $\{f(k)\}_{k=0}^q$. In order to deal with the particular case of stagnation, i.e., $f(0) > f(1) > \cdots > f(s) > 0$ and $f(s) = f(s+1) = \ldots = f(q)$, we keep the same framework but set $q = s + 1$ and redefine the vector $r^{(q)}$ ($r^{(q)}$ is the last vector in (2.23)). More details are provided in Subsection 2.2.3.

2.2.2 Proof of Theorem 2.11 for the case of a strictly decreasing cycle-convergence

Throughout this subsection we let the positive sequence $\{f(k)\}_{k=0}^q$ only to be *strictly decreasing*. We also assume here that $q = \max\{z \in \mathbb{Z} : z < n/m\}$. This

28

means that for the given n and m we perform our construction along the largest number of initial cycles where we are able to determine A (having a prescribed spectrum) and $r^{(0)}$ which provide the desired cycle-convergence. Although our proof is formally valid for any $0 < q < n/m$, the assumption emphasizes the extent to which we can take control over the process. We note that any case with $q < \max\{z \in \mathbb{Z} : z < n/m\}$ can be extended to the one assumed above by properly defining a number of additional elements in $\{f(k)\}_{k=0}^{q}$.

Step 1: Construction of a sequence of Krylov subspaces which provide the prescribed cycle-convergence

At the kth GMRES(m) cycle, the residual vector $r^{(k)}$ satisfies minimality condition (2.3). We assume that each set $\mathcal{W}_m^{(k)}$ is an orthonormal basis of a corresponding Krylov residual subspace $A\mathcal{K}_m\left(A, r^{(k-1)}\right)$, therefore condition (2.3) implies

$$r^{(k)} = r^{(k-1)} - \sum_{j=1}^{m} (r^{(k-1)}, w_j^{(k)}) w_j^{(k)}, \quad k = 1, \ldots, q. \tag{2.25}$$

At this stage, in order to simplify the forthcoming justification of the linear independence of the set $\overline{\mathcal{S}}$, we impose a stricter requirement on the residual change inside the cycle. We require that the residual vector $r^{(k-1)}$ remains constant during the first $m-1$ inner steps of GMRES and is reduced only at the last, mth, step. Thus, the equality in (2.25) can be written as

$$r^{(k)} = r^{(k-1)} - (r^{(k-1)}, w_m^{(k)}) w_m^{(k)}, \quad k = 1, \ldots, q. \tag{2.26}$$

This implies that the vectors $w_j^{(k)}$, $j = 1, \ldots, m-1$, are orthogonal to the residual vector $r^{(k-1)}$, i.e.,

$$(r^{(k-1)}, w_j^{(k)}) = 0, \quad j = 1, \ldots, m-1, \quad k = 1, \ldots, q. \tag{2.27}$$

From (2.26), using the fact that $r^{(k)} \perp w_m^{(k)}$ and the Pythagorean theorem, we obtain

$$|(r^{(k-1)}, w_m^{(k)})| = \sqrt{\|r^{(k-1)}\|^2 - \|r^{(k)}\|^2}, \quad k = 1, \ldots, q.$$

By defining (acute) angles $\psi_k \equiv \angle(r^{(k-1)}, w_m^{(k)})$ and the corresponding cosines $\cos \psi_k \equiv \frac{|(r^{(k-1)}, w_m^{(k)})|}{\|r^{(k-1)}\|}$, we can equivalently rewrite the identity above in the following form:

$$\cos \psi_k = \frac{\sqrt{f(k-1)^2 - f(k)^2}}{f(k-1)} \in (0,1), \quad k = 1, \ldots, q, \tag{2.28}$$

where $f(k-1)$ and $f(k)$ are the prescribed values for the norm of the residual vectors $r^{(k-1)}$ and $r^{(k)}$, respectively. Thus, if we are given $r^{(k-1)}$, one way to ensure the desired cycle-convergence at cycle k of GMRES(m) is to choose the unit vectors $w_j^{(k)}$ such that (2.26)–(2.28) holds.

In the following lemma, we show constructively that the described approach (2.26)–(2.28) leads to an appropriate set $\overline{\mathcal{S}}$.

Lemma 2.12 *Given a strictly decreasing positive sequence $\{f(k)\}_{k=0}^q$ and an initial vector $r^{(0)}$, $\|r^{(0)}\| = f(0)$, there exist vectors $r^{(k)}$, $\|r^{(k)}\| = f(k)$ and orthonormal sets $\mathcal{W}_m^{(k)}$ such that (2.26), (2.27), and (2.28) hold, and the set $\overline{\mathcal{S}}$ in (2.23) is linearly independent, $k = 1, \ldots, q$.*

Proof: The proof is by induction. Let $k = 1$. Given the initial vector $r^{(0)}$, $\|r^{(0)}\| = f(0)$, we pick $\mathcal{W}_{m-1}^{(1)} = \{w_1^{(1)}, \ldots, w_{m-1}^{(1)}\}$ an orthonormal set in $r^{(0)\perp}$ in order to satisfy equalities (2.27). The set $\{r^{(0)}, \mathcal{W}_{m-1}^{(1)}\}$ is linearly independent.

In order to choose the unit vector $w_m^{(1)}$ orthogonal to the previously constructed vectors $\mathcal{W}_{m-1}^{(1)}$ and satisfying (2.28), we introduce a unit vector $y^{(1)} \in \{r^{(0)}, \mathcal{W}_{m-1}^{(1)}\}^\perp$, so that

$$w_m^{(1)} = \frac{r^{(0)}}{f(0)}\cos\psi_1 + y^{(1)}\sin\psi_1.$$

We find the vector $r^{(1)}$ by satisfying (2.26). Equality (2.28) guarantees that $\|r^{(1)}\| = f(1)$, as desired. Finally, we append the constructed vector $r^{(1)}$ to $\{r^{(0)}, \mathcal{W}_{m-1}^{(1)}\}$ and get the set $\{r^{(0)}, \mathcal{W}_{m-1}^{(1)}, r^{(1)}\}$, which is linearly independent, since, by construction, $r^{(1)}$ is not in span$\{r^{(0)}, \mathcal{W}_{m-1}^{(1)}\}$.

The induction assumption is that we have already constructed $k-1$ vectors $r^{(1)}, \ldots, r^{(k-1)}$ with the prescribed norms $f(1), \ldots, f(k-1)$ and orthonormal sets $\mathcal{W}_m^{(1)}, \ldots, \mathcal{W}_m^{(k-1)}$, such that equalities (2.26), (2.27) and (2.28) hold, and the set

$$\{r^{(0)}, \mathcal{W}_{m-1}^{(1)}, \ldots, r^{(k-2)}, \mathcal{W}_{m-1}^{(k-1)}, r^{(k-1)}\} \tag{2.29}$$

is linearly independent. We want to show that we can construct the next vector $r^{(k)}$, $\|r^{(k)}\| = f(k)$, and the orthonormal set $\mathcal{W}_m^{(k)}$, satisfying (2.26), (2.27) and (2.28), such that

$$\{r^{(0)}, \mathcal{W}_{m-1}^{(1)}, \ldots, r^{(k-2)}, \mathcal{W}_{m-1}^{(k-1)}, r^{(k-1)}, \mathcal{W}_{m-1}^{(k)}, r^{(k)}\} \tag{2.30}$$

is linearly independent, $k \leq q$.

We start by constructing orthonormal vectors $\mathcal{W}_{m-1}^{(k)} = \{w_1^{(k)}, \ldots, w_{m-1}^{(k)}\}$, satisfying (2.27), with the additional requirement that the set $\mathcal{W}_{m-1}^{(k)}$ is not

31

in the span of the previously constructed vectors given in set (2.29). From these considerations, we choose $\mathcal{W}_{m-1}^{(k)}$ as an orthonormal set in the orthogonal complement of (2.29), i.e.,

$$w_j^{(k)} \in \{r^{(0)}, \mathcal{W}_{m-1}^{(1)}, \ldots, r^{(k-2)}, \mathcal{W}_{m-1}^{(k-1)}, r^{(k-1)}\}^\perp, \quad j = 1, \ldots, m-1.$$

Appending $\mathcal{W}_{m-1}^{(k)}$ to the set (2.29) gives a linearly independent set.

To finish the proof, we need to construct the vector $w_m^{(k)}$, satisfying (2.28) and orthogonal to $\mathcal{W}_{m-1}^{(k)}$. For this reason we introduce a unit vector $y^{(k)}$,

$$y^{(k)} \in \{r^{(0)}, \mathcal{W}_{m-1}^{(1)}, \ldots, r^{(k-2)}, \mathcal{W}_{m-1}^{(k-1)}, r^{(k-1)}, \mathcal{W}_{m-1}^{(k)}\}^\perp,$$

so that
$$w_m^{(k)} = \frac{r^{(k-1)}}{f(k-1)}\cos\psi_k + y^{(k)}\sin\psi_k,$$

where $\cos\psi_k \equiv \frac{|(r^{(k-1)}, w_m^{(k)})|}{\|r^{(k-1)}\|}$.

We define the vector $r^{(k)}$ with (2.26). Equality (2.28) guarantees $\|r^{(k)}\| = f(k)$. Set (2.30) is linearly independent, since, by construction, the vector $r^{(k)}$ is not in span$\{r^{(0)}, \mathcal{W}_{m-1}^{(1)}, \ldots, r^{(k-2)}, \mathcal{W}_{m-1}^{(k-1)}, r^{(k-1)}, \mathcal{W}_{m-1}^{(k)}\}$. ∎

Step 2: Definition of a linear operator with any prescribed spectrum

So far we have shown that, given a strictly decreasing positive sequence $\{f(k)\}_{k=0}^q$ and an initial residual vector $r^{(0)}$, $\|r^{(0)}\| = f(0)$, it is possible to construct vectors $r^{(k)}$, $\|r^{(k)}\| = f(k)$, and orthonormal vectors $\mathcal{W}_m^{(k)}$, $k = 1, \ldots, q$, satisfying equalities (2.26), (2.27) and (2.28), such that the set $\overline{\mathcal{S}}$ of $mq + 1$ vectors in (2.23) is linearly independent.

In order to define a (representation of) unique linear operator, we need to have a valid basis of $\mathbb{C}^n$ at hand. Thus, we expand the set $\overline{\mathcal{S}}$ by linearly

independent vectors $\widehat{\mathcal{S}} = \{\widehat{s}_1, \dots, \widehat{s}_t\}$, $t = n - mq - 1$ ($< m$, since we have assumed that $q = \max \{z \in \mathbb{Z} : z < n/m\}$):

$$\mathcal{S} = \{r^{(0)}, \mathcal{W}_{m-1}^{(1)}, \dots, r^{(q-1)}, \mathcal{W}_{m-1}^{(q)}, r^{(q)}, \widehat{s}_1, \dots, \widehat{s}_t\}, \tag{2.31}$$

so that $\mathcal{S}$ is a basis of $\mathbb{C}^n$.

Before we define a linear operator $\mathcal{A}$, let us consider the set

$$\Lambda = \{\lambda_1, \lambda_2, \dots, \lambda_n\}$$

of nonzero numbers in the complex plane that defines the spectrum of $\mathcal{A}$. We split Λ into $q + 1$ disjoint subsets

$$\Lambda = \{\Lambda_1, \Lambda_2, \dots, \Lambda_q, \Lambda_{q+1}\}, \tag{2.32}$$

such that each Λ_k, $k = 1, \dots, q$, contains m elements of Λ, and the remaining $n - mq$ elements are included into Λ_{q+1}.

For each set Λ_k, $k = 1, \dots, q$, we define a monic polynomial $p_k(x)$, such that the roots of this polynomial are exactly the elements of the corresponding Λ_k:

$$p_k(x) = x^m - \sum_{j=0}^{m-1} \alpha_j^{(k)} x^j, \quad k = 1, \dots, q, \tag{2.33}$$

with $\alpha_j^{(k)}$'s being the coefficients of the respective polynomials, $\alpha_0^{(k)} \neq 0$. Each polynomial $p_k(x)$ in (2.33) can be thought of as the characteristic polynomial of an m-by-m matrix with spectrum Λ_k.

Let us also introduce an arbitrary $(t + 1)$-by-$(t + 1)$ matrix C with the spectrum Λ_{q+1}:

$$C = (\beta_{ij}), \quad \Lambda(C) = \Lambda_{q+1}, \quad i, j = 1, \dots, t + 1 = n - mq. \tag{2.34}$$

33

We define the operator $\mathcal{A} : \mathbb{C}^n \longrightarrow \mathbb{C}^n$ as follows:

$$\mathcal{A}r^{(k-1)} = w_1^{(k)},$$

$$\mathcal{A}w_1^{(k)} = w_2^{(k)},$$

$$\vdots$$

$$\mathcal{A}w_{m-2}^{(k)} = w_{m-1}^{(k)},$$

$$\mathcal{A}w_{m-1}^{(k)} = -\alpha_0^{(k)}r^{(k)} + \alpha_0^{(k)}r^{(k-1)} + \alpha_1^{(k)}w_1^{(k)} + \cdots + \alpha_{m-1}^{(k)}w_{m-1}^{(k)}, \quad k = 1, \ldots q;$$

$$\mathcal{A}r^{(q)} = \beta_{11}r^{(q)} + \beta_{21}\widehat{s}_1 + \cdots + \beta_{t+1,1}\widehat{s}_t,$$

$$\mathcal{A}\widehat{s}_1 = \beta_{12}r^{(q)} + \beta_{22}\widehat{s}_1 + \cdots + \beta_{t+1,2}\widehat{s}_t,$$

$$\vdots \tag{2.35}$$

$$\mathcal{A}\widehat{s}_t = \beta_{1,t+1}r^{(q)} + \beta_{2,t+1}\widehat{s}_1 + \cdots + \beta_{t+1,t+1}\widehat{s}_t,$$

where $\alpha_j^{(k)}$'s are the coefficients of polynomials (2.33) and β_{ij}'s are the elements of the matrix C in (2.34).

The following lemma shows that, given vectors $r^{(k)}$ and orthonormal sets $\mathcal{W}_m^{(k)}$ constructed according to Lemma 2.12, the linear operator $\mathcal{A}$, defined by (2.35) and represented by a matrix A in the canonical basis, generates the desired Krylov residual subspaces given in (2.22); and the spectrum of $\mathcal{A}$ can be chosen arbitrarily.

Lemma 2.13 *Let the initial residual vector $r^{(0)}$, $\|r^{(0)}\| = f(0)$, as well as the residual vectors $r^{(k)}$ and orthonormal sets $\mathcal{W}_m^{(k)}$ be constructed according to Lemma 2.12. Let $\mathcal{S}$ be the basis of $\mathbb{C}^n$ as defined by (2.31) and Λ be an arbitrary set of n nonzero complex numbers. Then the linear operator $\mathcal{A}$ defined according to (2.32)–(2.35) generates the Krylov residual subspaces given in (2.22), where*

the matrix A is a representation of $\mathcal{A}$ in the canonical basis. Moreover, the spectrum of $\mathcal{A}$ is Λ.

Proof: From definition (2.35) of the linear operator $\mathcal{A}$ one can notice that

$$\mathcal{A}r^{(k-1)} = w_1^{(k)},$$

$$\mathcal{A}^2 r^{(k-1)} = w_2^{(k)},$$

$$\vdots$$

$$\mathcal{A}^{m-1} r^{(k-1)} = w_{m-1}^{(k)},$$

$$\mathcal{A}^m r^{(k-1)} = -\alpha_0^{(k)} r^{(k)} + \alpha_0^{(k)} r^{(k-1)} + \alpha_1^{(k)} w_1^{(k)} + \cdots + \alpha_{m-1}^{(k)} w_{m-1}^{(k)}, \quad k = 1, \ldots q.$$

Since, by construction in Lemma 2.12, i.e., equality (2.26),

$$0 \neq -\alpha_0^{(k)} r^{(k)} + \alpha_0^{(k)} r^{(k-1)} \in \mathrm{span}\{w_m^{(k)}\},$$

the above relations immediately imply that for each $k = 1, \ldots q,$

$$\mathrm{span}\{\mathcal{A}r^{(k-1)}, \ldots, \mathcal{A}^j r^{(k-1)}\} = \mathrm{span}\, \mathcal{W}_j^{(k)}, \quad j = 1, \ldots, m.$$

Thus, given the representation A of the linear operator $\mathcal{A}$ in the canonical basis, we have proved that $\mathcal{A}$ generates the Krylov residual subspaces given in (2.22).

To prove that an arbitrarily chosen set Λ can be the spectrum of $\mathcal{A}$, let us consider the matrix $[\mathcal{A}]_{\mathcal{S}}$ of the operator $\mathcal{A}$ in the basis $\mathcal{S}$ (see (2.31) and (2.35)):

$$[\mathcal{A}]_{\mathcal{S}} = \begin{bmatrix} 0\,0 \cdots 0 & \alpha_0^{(1)} & & & \\ 1\,0 \cdots 0 & \alpha_1^{(1)} & & & \\ 0\,1 \cdots 0 & \alpha_2^{(1)} & & & \mathbf{0} \\ \vdots\,\vdots\,\ddots\,\vdots & \vdots & & & \\ 0\,0 \cdots 1 & \alpha_{m-1}^{(1)} & & & \\ & -\alpha_0^{(1)} & \ddots & & \\ & & \ddots & \ddots & \\ & & -\alpha_0^{(q-1)} & 0\,0 \cdots 0 & \alpha_0^{(q)} \\ & & & 1\,0 \cdots 0 & \alpha_1^{(q)} \\ & & & 0\,1 \cdots 0 & \alpha_2^{(q)} \\ & & & \vdots\,\vdots\,\ddots\,\vdots & \vdots \\ & \mathbf{0} & & 0\,0 \cdots 1 & \alpha_{m-1}^{(q)} \\ & & & & -\alpha_0^{(q)} & \beta_{11} & \cdots & \beta_{1,t+1} \\ & & & & & \vdots & \ddots & \vdots \\ & & & & & \beta_{t+1,1} & \cdots & \beta_{t+1,t+1} \end{bmatrix}. \quad (2.36)$$

The matrix $[\mathcal{A}]_{\mathcal{S}}$ has a block lower triangular structure, hence the spectrum of $[\mathcal{A}]_{\mathcal{S}}$ is a union of eigenvalues of all diagonal blocks. The first q blocks are the companion matrices corresponding to the sets Λ_k, $k = 1, \ldots, q$, with characteristic polynomials defined in (2.33). The last block is exactly the matrix $C = (\beta_{ij})$ from (2.34) with the spectrum Λ_{q+1}. Thus, given partition (2.32) of the set Λ, we conclude that spectrum of $\mathcal{A}$ is Λ. ∎

Step 3: Conclusion of the proof of Theorem 2.11 for the case of the strictly decreasing cycle-convergence

Finally, we define A as the representation of the operator $\mathcal{A}$ in the canonical basis $\mathcal{E} = \{e_1, e_2, \ldots, e_n\}$,

$$A = S [\mathcal{A}]_\mathcal{S} S^{-1}, \tag{2.37}$$

where the square matrix S is formed by the vectors given in (2.31) written as columns and $[\mathcal{A}]_\mathcal{S}$ is defined by (2.36). The constructed matrix A provides the prescribed (strictly decreasing) norms of residual vectors at the first q GMRES(m) cycles if starting with $r^{(0)}$ and its spectrum is Λ. We note that the field over which the resulting matrix is defined depends heavily on the partition (2.32) of the set Λ, e.g., A turns out to be (non-Hermitian) complex if a conjugate pair from Λ is not included into the same subset Λ_k.

2.2.3 Extension to the case of stagnation

In the previous subsection, we have proved Theorem 2.11 only for the case of the strictly decreasing positive sequence $\{f(k)\}_{k=0}^q$. Now, in order to conclude the rest of Theorem 2.11, we consider the case of stagnation: $f(0) > f(1) > \cdots > f(s) > 0$ and $f(s) = f(s+1) = \ldots = f(q)$. The latter fits well (after a minor modification) into the framework presented above.

Let us set $q = s + 1$ and, without loss of generality, reduce the problem to constructing a matrix A with a spectrum Λ and an initial residual vector $r^{(0)}$, $\|r^{(0)}\| = f(0)$, for which GMRES(m) produces the following sequence of residual norms: $f(0) > f(1) > \cdots > f(q-1) = f(q) > 0$. We observe that the sequence is strictly decreasing up to the last cycle q. Thus, by Lemma 2.12, at the initial

$q - 1(= s)$ cycles we are able to construct sets $\mathcal{W}_m^{(k)}$ and vectors $r^{(k)}$, such that $\|r^{(k)}\| = f(k)$ and the set

$$\{r^{(0)}, \mathcal{W}_{m-1}^{(1)}, \ldots, r^{(q-1)}, \mathcal{W}_{m-1}^{(q-1)}, r^{(q-1)}\} \tag{2.38}$$

is linearly independent. Then, formally following the construction in Lemma 2.12 at the cycle q, we get orthonormal set $\mathcal{W}_m^{(q)}$ from the orthogonal complement of (2.38) and the residual vector $r^{(q)} = r^{(q-1)}$. This leads to set (2.23), which is no longer linearly independent due to the above mentioned equality of residual vectors. To enforce the linear independence we substitute in (2.23) the "inconvenient" vector $r^{(q)}$ by $w_m^{(q)} + r^{(q-1)}$ and obtain the set

$$\{r^{(0)}, \mathcal{W}_{m-1}^{(1)}, \ldots, r^{(q-2)}, \mathcal{W}_{m-1}^{(q-1)}, r^{(q-1)}, \mathcal{W}_{m-1}^{(q)}, w_m^{(q)} + r^{(q-1)}\}, \tag{2.39}$$

which is linearly independent, due to the fact that the orthonormal set $\mathcal{W}_m^{(q)}$ is chosen, by construction, from the orthogonal complement of (2.38).

The rest of the proof exactly follows the pattern described in Subsection 2.2.2 with $r^{(q)}$ replaced by $w_m^{(q)} + r^{(q-1)}$, $q = s + 1$; see (2.31)–(2.37). The resulting matrix A has the prescribed spectrum Λ and with the initial residual vector $r^{(0)}$, $\|r^{(0)}\| = f(0)$, provides the desired cycle-convergence of GMRES(m) with a stagnation starting after cycle s.

This concludes the proof of Theorem 2.11. In what follows we suggest several remarks and generalizations related to the result.

2.2.4 Difference with the work of Greenbaum, Pták, and Strakoš [34]

For the reader familiar with the work of Greenbaum, Pták, and Strakoš [34], it might be tempting to obtain the present result by pursuing the following scheme: fix $r^{(0)}$ and then consider the first restarted GMRES cycle as the initial

part of a full GMRES run where the convergence is prescribed for the first m iterations (and set arbitrarily for the remaining $n-m$ iterations). Then, similarly, given the starting residual vector $r^{(1)}$ provided by this first cycle, construct the next Krylov residual subspace, which provides the desired convergence following the scheme of Greenbaum, Pták, and Strakoš [34]. Proceed identically for the remaining cycles. This approach, however, does not guarantee the linear independence of the set $\overline{\mathcal{S}}$ in (2.23) and, hence, one meets the problem of defining the linear operator $\mathcal{A}$. These considerations have been the reason for assumptions (2.26), (2.27) on the residual reduction inside a cycle, which have allowed us to quite easily justify the linear independence of the set $\overline{\mathcal{S}}$ and to control the spectrum, as well.

2.2.5 Generating examples with nonzero r_{q+1}

We note from definition (2.35) of the operator $\mathcal{A}$ in Subsection 2.2.2 that $\mathrm{span}\{r^{(q)}, \widehat{s}_1, \ldots, \widehat{s}_t\}$ is an invariant subspace of $\mathcal{A}$ and, hence,

$$r^{(q)} \in A\mathcal{K}_{t+1}\left(A, r^{(q)}\right),$$

where A is the representation of the operator $\mathcal{A}$ in the canonical basis and $t = n - mq - 1$ ($< m$, by the assumption that $q = \max\{z \in \mathbb{Z} : z < n/m\}$ at the beginning of Subsection 2.2.2). This implies that at the end of the $(q+1)$st cycle GMRES(m) converges to the exact solution of system (2.1), i.e., $r^{(q+1)} = 0$. This fact might seem unnatural and undesirable, e.g., for constructing theoretical examples. The "drawback", however, can be easily fixed by a slight correction of the basis $\mathcal{S}$ in (2.31)—somewhat similarly to how we handled the stagnation case in Theorem 2.11.

Given residuals $r^{(k)}$ and orthonormal sets $\mathcal{W}_m^{(k)}$ constructed according to Lemma 2.12, instead of considering the set $\mathcal{S}$, we consider the following basis of $\mathbb{C}^n$:

$$\tilde{\mathcal{S}} = \{r^{(0)}, w_1^{(1)}, \ldots, w_{m-1}^{(1)}, \ldots, r^{(q-1)}, w_1^{(q)}, \ldots, w_{m-1}^{(q)}, r^{(q)} + \gamma r^{(q-1)}, \widehat{s}_1, \ldots, \widehat{s}_t\},$$

(2.40)

where $\gamma \neq -1, 0$. Here we substituted the basis vector $r^{(q)}$ in (2.31) by $r^{(q)} + \gamma r^{(q-1)}$. The vector $r^{(q)} + \gamma r^{(q-1)}$ cannot be represented as a linear combination of other vectors in $\tilde{S}$, since it contains the component $r^{(q)}$, which is not represented by these vectors. Hence, $\tilde{\mathcal{S}}$ is indeed a basis of $\mathbb{C}^n$. Thus we can define the operator $\mathcal{A}$ by its action on $\tilde{\mathcal{S}}$:

$$\mathcal{A}r^{(k-1)} = w_1^{(k)},$$

$$\mathcal{A}w_1^{(k)} = w_2^{(k)},$$

$$\vdots$$

$$\mathcal{A}w_{m-2}^{(k)} = w_{m-1}^{(k)},$$

$$\mathcal{A}w_{m-1}^{(k)} = -\alpha_0^{(k)}r^{(k)} + \alpha_0^{(k)}r^{(k-1)} + \alpha_1^{(k)}w_1^{(k)} + \cdots + \alpha_{m-1}^{(k)}w_{m-1}^{(k)},$$

$$k = 1, \ldots, q-1;$$

$$\mathcal{A}r^{(q-1)} = w_1^{(q)},$$

$$\mathcal{A}w_1^{(q)} = w_2^{(q)},$$

$$\vdots \qquad\qquad\qquad (2.41)$$

$$\mathcal{A}w_{m-2}^{(q)} = w_{m-1}^{(q)},$$

$$\mathcal{A}w_{m-1}^{(q)} = \frac{-\alpha_0^{(q)}}{1+\gamma}(r^{(q)} + \gamma r^{(q-1)}) + \alpha_0^{(q)} r^{(q-1)}$$
$$+ \alpha_1^{(q)} w_1^{(q)} + \cdots + \alpha_{m-1}^{(q)} w_{m-1}^{(q)},$$

$$\mathcal{A}(r^{(q)} + \gamma r^{(q-1)}) = \beta_{11}(r^{(q)} + \gamma r^{(q-1)}) + \beta_{21}\widehat{s}_1 + \cdots + \beta_{t+1,1}\widehat{s}_t,$$

$$\mathcal{A}\widehat{s}_1 = \beta_{12}(r^{(q)} + \gamma r^{(q-1)}) + \beta_{22}\widehat{s}_1 + \cdots + \beta_{t+1,2}\widehat{s}_t,$$

$$\vdots$$

$$\mathcal{A}\widehat{s}_t = \beta_{1,t+1}(r^{(q)} + \gamma r^{(q-1)}) + \beta_{2,t+1}\widehat{s}_1 + \cdots + \beta_{t+1,t+1}\widehat{s}_t,$$

where $\alpha_j^{(k)}$'s are the coefficients of the corresponding characteristic polynomials (2.33) and β_{ij}'s are the elements of the matrix C in (2.34). The fact that the operator $\mathcal{A}$ produces the correct Krylov residual subspace at the cycle q, i.e.,

$$\mathrm{span}\{\mathcal{A}r^{(q-1)}, \ldots, \mathcal{A}^m r^{(q-1)}\} = \mathrm{span}\ \mathcal{W}_m^{(q)},$$

can be observed from the following equalities:

$$\mathcal{A}w_{m-1}^{(q)} = \frac{-\alpha_0^{(q)}}{1+\gamma}(r^{(q)} + \gamma r^{(q-1)}) + \alpha_0^{(q)} r^{(q-1)} + \alpha_1^{(q)} w_1^{(q)} + \cdots + \alpha_{m-1}^{(q)} w_{m-1}^{(q)}$$

$$= \frac{-\alpha_0^{(q)}}{1+\gamma}(r^{(q)} - r^{(q-1)} + (1+\gamma)r^{(q-1)}) + \alpha_0^{(q)} r^{(q-1)}$$
$$+ \alpha_1^{(q)} w_1^{(q)} + \cdots + \alpha_{m-1}^{(q)} w_{m-1}^{(q)}$$

$$= \frac{-\alpha_0^{(q)}}{1+\gamma}(r^{(q)} - r^{(q-1)}) + \alpha_1^{(q)} w_1^{(q)} + \cdots + \alpha_{m-1}^{(q)} w_{m-1}^{(q)},$$

where, by (2.41), $\mathcal{A}w_{m-1}^{(q)} = \mathcal{A}^m r^{(q-1)}$ and, by (2.26), $0 \neq r^{(q)} - r^{(q-1)} \in \mathrm{span}\{w_m^{(q)}\}$.

The matrix $[\mathcal{A}]_{\tilde{\mathcal{S}}}$ of the operator $\mathcal{A}$, defined by (2.41), in the basis $\tilde{\mathcal{S}}$ is identical to (2.36), with the only change of the subdiagonal element $-\alpha_0^{(q)}$ to $\frac{-\alpha_0^{(q)}}{1+\gamma}$, $\gamma \neq -1, 0$. Hence, $\mathcal{A}$ has the desired spectrum Λ. The representation A of the operator $\mathcal{A}$ in the canonical basis is then determined by the similarity transformation in (2.37), with the matrix S formed by vectors from $\tilde{\mathcal{S}}$ in (2.40) written as columns.

Finally, to see that the residual vector $r^{(q+1)}$ is generally nonzero with the new definition of the operator $\mathcal{A}$, we notice from (2.41) that now

$$\text{span}\left\{r^{(q)} + \gamma r^{(q-1)}, \widehat{s}_1, \ldots, \widehat{s}_t\right\}$$

is an invariant subspace of $\mathcal{A}$ and, hence,

$$r^{(q)} + \gamma r^{(q-1)} \in A\mathcal{K}_{t+1}\left(A, r^{(q)} + \gamma r^{(q-1)}\right), \ \gamma \neq -1, 0,$$

or,

$$r^{(q)} \in A\mathcal{K}_{t+1}\left(A, r^{(q)}\right) + \mathcal{K}_{t+2}\left(A, r^{(q-1)}\right), \tag{2.42}$$

where $t+1 \leq m$ by the assumption that $q = \max\left\{z \in \mathbb{Z} : z < n/m\right\}$. Due to the fact that $r^{(q+1)} = 0$ if and only if $r^{(q)} \in A\mathcal{K}_m\left(A, r^{(q)}\right)$, it suffices to show that the component of the vector $r^{(q)}$ from $\mathcal{K}_{t+2}\left(A, r^{(q-1)}\right)$ in representation (2.42) does not generally belong to $A\mathcal{K}_m\left(A, r^{(q)}\right)$. To show this, we observe, since $\gamma \neq 0$, that the term from $\mathcal{K}_{t+2}\left(A, r^{(q-1)}\right)$ in (2.42) contains a nonzero component in the direction $r^{(q-1)}$, which is not in $A\mathcal{K}_m\left(A, r^{(q)}\right)$ unless the initial residual $r^{(0)}$ is chosen from a specific subspace of $\mathbb{C}^n$, i.e., expressing $r^{(q-1)}$ in terms of $r^{(0)}$ and vectors $w_m^{(k)}$ by (2.26),

$$r^{(0)} \notin \mathcal{I}, \ \mathcal{I} = \text{span} \left\{ w_m^{(1)}, \dots, w_m^{(q-1)} \right\} + A\mathcal{K}_m \left(A, r^{(q)} \right),$$

where $\dim \mathcal{I} \le q + m - 1 < n/m + m - 1 \le n$, provided that $0 < m < n$.

2.2.6 Any admissible convergence behavior is possible for full and restarted GMRES (at its q initial cycles)

As we pointed out at the beginning of the current section, the convergence behavior of the *full* GMRES in Theorem 2.2 is restricted to the class of convergence sequences which allow convergence to the exact solution only at step n, i.e., $f(0) \ge f(1) \ge \cdots \ge f(n-1) > 0 \ (f(n) = 0)$. Similarly, the cycle-convergence behavior of *restarted* GMRES in Theorem 2.11 is restricted to cycle-convergence sequences which exclude the possibility of convergence to the exact solution within the initial q cycles, i.e., $f(0) > f(1) > \cdots > f(s) > 0$ and $f(s) = f(s+1) = \ldots = f(q)$. It turns out that the assumptions in Theorem 2.2 and Theorem 2.11 are sufficient for the theorems also to hold if $f(0) \ge f(1) \ge \cdots \ge f(n-1) \ge 0$ and $f(0) > f(1) > \cdots > f(s) \ge 0$, $f(s) = f(s+1) = \ldots = f(q)$, respectively.

Given an integer $n > 0$, assume that we want to construct an n-by-n matrix A with a prescribed spectrum Λ and an initial residual vector $r^{(0)}$ (or, equivalently, a right-hand side b, since $r^{(0)} = b$ after setting $x^{(0)}$ to 0) such that the *full* GMRES applied to the corresponding system (2.1) results in the following convergence pattern: $f(0) \ge f(1) \ge \cdots \ge f(s-1) > f(s) = f(s+1) = \cdots = f(n-1) = 0$, $s < n$, $\|r_k\| = f(k)$. The construction is straight-forward. We first split the set Λ into two disjoint subsets, say, $\Lambda = \Lambda_s \cup \Lambda_{n-s}$, where Λ_s contains s elements from Λ while the remaining $n-s$ elements are included into

Λ_{n-s}. Next, by Theorem 2.2 we construct a matrix $A_s \in \mathbb{C}^{s \times s}$ and a right-hand side $b_s \in \mathbb{C}^s$ ($x^{(0)} = 0 \in \mathbb{C}^s$), such that the *full* GMRES applied to the system $A_s x = b_s$ produces the convergence sequence $f(0) \geq f(1) \geq \cdots \geq f(s-1) > 0$ ($f(s) = 0$), moreover the spectrum of A_s is Λ_s. Finally, we define the resulting matrix $A \in \mathbb{C}^{n \times n}$ and the right-hand side vector $b \in \mathbb{C}^n$ ($x^{(0)} = 0 \in \mathbb{C}^n$) as follows:

$$A = \begin{pmatrix} A_s & 0 \\ 0 & A_{n-s} \end{pmatrix}, \quad b = \begin{pmatrix} b_s \\ 0 \end{pmatrix}, \tag{2.43}$$

where $A_{n-s} \in \mathbb{C}^{(n-s) \times (n-s)}$ is an arbitrary matrix with a spectrum Λ_{n-s}. It is easy to see that the *full* GMRES applied to the system of equations defined by (2.43) produces the desired sequence of residual norms $f(0) \geq f(1) \geq \cdots \geq f(s-1) > f(s) = f(s+1) = \cdots = f(n-1) = 0$, $\|r^{(k)}\| = f(k)$, $r^{(0)} = b$. Clearly the matrix A in (2.43) has the prescribed spectrum $\Lambda = \Lambda_s \cup \Lambda_{n-s}$.

For the *restarted* GMRES the construction of a matrix A with the spectrum Λ and a right-hand side b ($x^{(0)} = 0$) that provide the cycle-convergence sequence $f(0) > f(1) > \cdots > f(s-1) > f(s) = f(s+1) = \ldots = f(q) = 0$ is analogous, $s \leq q$, $\|r^{(k)}\| = f(k)$. Following Theorem 2.11, one constructs a matrix $A_s \in \mathbb{C}^{ms \times ms}$ and a right-hand side vector $b_s \in \mathbb{C}^{ms}$, such that the GMRES(m) applied to the corresponding linear system produces the cycle-convergence curve $f(0) > f(1) > \cdots > f(s-1) > f(s) = 0$. The spectrum of A_s is chosen to coincide with a subset of ms elements of Λ. The construction of the matrix $A \in \mathbb{C}^{n \times n}$ and the right-hand side $b \in \mathbb{C}^n$ is then accomplished by introducing an $(n-ms) \times (n-ms)$ diagonal block with eigenvalues from Λ, which are not in the spectrum of A_s, and expanding the vector b with $(n - ms)$ zeros, similarly to (2.43).

2.2.7 Restarted GMRES with variable restart parameter

The result of Theorem 2.11 can be generalized to the case where the restart parameter m is not fixed, but varies over successive cycles according to an a priori prescribed parameter sequence $\{m_k\}_{k=1}^q$. The proof, basically, repeats the one in Subsection 2.2.2 with the difference that the constructed operator $\mathcal{A}$ in the corresponding basis has block lower triangular structure with varying diagonal block sizes m_k, rather than the constant size $m_k = m$ as in (2.36).

Corollary 2.14 *Let us be given an integer $n > 0$, a sequence $\{m_k\}_{k=1}^q$, $0 < m_k < n$, and a positive sequence $\{f(k)\}_{k=0}^q$, such that $f(0) > f(1) > \cdots > f(s) > 0$ and $f(s) = f(s+1) = \ldots = f(q)$, where q is defined by the condition $\sum_{k=1}^q m_k < n$, $0 \leq s \leq q$. There exists an n-by-n matrix A and a vector $r^{(0)}$ with $\|r^{(0)}\| = f(0)$ such that $\|r^{(k)}\| = f(k)$, $k = 1, \ldots, q$, where $r^{(k)}$ is the residual at cycle k of restarted GMRES with a restart parameter varying according to the sequence $\{m_k\}_{k=1}^q$ applied to the linear system $Ax = b$, with initial residual $r^{(0)} = b - Ax^{(0)}$. Moreover, the matrix A can be chosen to have any desired (nonzero) eigenvalues.*

2.3 Conclusions

In this chapter we have established several results which address the cycle-convergence behavior of the restarted GMRES. First, we have proved that the cycle-convergence of the method applied to a system of linear equations with a normal coefficient matrix is sublinear, and at best linear. Second, in the general case, we have shown that any admissible cycle-convergence behavior is possible

for the restarted GMRES at q initial cycles, regardless of the eigenvalue distribution of the coefficient matrix. This leads to the conclusion that no estimates, which rely solely on the matrix spectrum, can be derived to characterize the cycle-convergence of restarted GMRES at the first q cycles if the method is applied to a linear system with a general nonsingular non-Hermitian matrix. Though in practice q tends to be reasonably large ($q < n/m$), it remains an open question if the above mentioned estimates hold at cycles which follow the n/m-th GMRES(m) cycle.

3. Solution of symmetric indefinite systems with symmetric positive definite preconditioners

We consider a system of linear equations

$$Ax = b, \quad A = A^* \in \mathbb{R}^{n \times n}, \quad b \in \mathbb{R}^n, \tag{3.1}$$

where the coefficient matrix A is nonsingular and symmetric indefinite, i.e., the spectrum of A contains both positive and negative eigenvalues.

Linear systems with large, possibly sparse, symmetric indefinite coefficient matrices arise in a variety of applications. For example, in the form of saddle point problems (see [10] and references therein), such systems may result from mixed finite element discretizations of underlying differential equations of fluid and solid mechanics. In acoustics, large sparse symmetric indefinite systems may be obtained after discretizing the Helmholtz equation [69] for certain media types and boundary conditions. Sometimes the need of solving the indefinite problem (3.1) comes as an auxiliary task within other computational routines, e.g., inner Newton step in the interior point methods in linear and nonlinear optimization, see [53], or solution of the correction equation in the Jacobi-Davidson method [64] for a symmetric eigenvalue problem.

Because of the large problem size, direct methods for solving linear systems may become infeasible, which motivates the use of iterative techniques for finding satisfactory approximations to the exact solutions. There is a number of iterative methods developed specifically to solve symmetric indefinite systems, ranging from modifications of the Richardson's iteration, e.g., [51, 58, 16],

to optimal Krylov subspace methods, see [33, 59]. It is known, however, that in practical problems the coefficient matrix A in (3.1) may be extremely ill-conditioned which, along with the location of the spectrum of A to both sides of the origin, can make the straightforward application of the existing schemes inefficient due to a slow convergence rate. In order to improve the convergence, one can introduce a matrix $T \in \mathbb{R}^{n \times n}$ and consider the *preconditioned system*

$$TAx = Tb. \tag{3.2}$$

If T is not symmetric positive definite (SPD), the matrix TA of the preconditioned system (3.2), in general, is not symmetric with respect to any inner product, implying that the specialized methods for solving symmetric indefinite systems are no longer applicable and need to be replaced by methods for solving nonsymmetric systems, e.g., one of the Krylov subspace methods: GMRES or GMRES(m), BiCG, BiCGstab, QMR, etc (see, e.g., [33, 59]). Though known to be effective for a number of applications, this approach can have several disadvantages. First, in order to maintain the optimality of a Krylov subspace method, one has to allow the increase of the computational work at every new iteration, which can become prohibitive for large problems. Second, the convergence behavior of methods for solving nonsymmetric systems may not rely on possibly accessible (estimated) quantities, such as, e.g., the spectrum of the coefficient matrix (see the corresponding results for GMRES [34, 76]), which makes it difficult or even impossible to estimate the computational costs a priori.

If T is chosen to be SPD ($T = T^* > 0$) then the matrix TA of preconditioned system (3.2) remains symmetric indefinite, however, with respect to the T^{-1}–inner product, i.e., the inner product defined by $(x, y)_{T^{-1}} = (x, T^{-1}y)$

48

for any $x, y \in \mathbb{R}^n$, where $(\cdot, \cdot)$ denotes the Euclidean inner product, in which the matrix A is symmetric. In particular, due to this symmetry preservation (though in a different geometry), system (3.2) can be solved using an *optimal short-term recurrent* Krylov subspace method, e.g., preconditioned MINRES [55], or PMINRES, with the convergence behavior fully described in terms of the (estimates of) spectrum of the preconditioned matrix TA. Therefore, in the light of the discussion above, the choice of a properly defined SPD preconditioner for solving a symmetric indefinite system can be regarded as natural and favorable.

The goal of this chapter is twofold. First, we describe a hierarchy of methods, from a stationary iteration to an optimal Krylov subspace minimal residual method, which allow solving symmetric indefinite linear system (3.1) with an SPD preconditioner T. Second, we suggest a new strategy for constructing SPD preconditioners for *general* symmetric indefinite systems being solved with the described methods. Although the approaches, underlying the methods are mostly known, e.g., the minimization of an appropriate norm of the residual vector over a subspace, several of our observations seem to be new and are, primarily, of a theoretical interest. In particular, we determine the smallest possible Krylov subspace, which can be used to properly restart the preconditioned minimal residual method, applied to a symmetric indefinite linear system. For example, this leads to a scheme which is a natural analogue of the preconditioned steepest descent iteration for solving SPD systems. Independently of a particular implementation scheme, we state and prove simple convergence bounds, and gain an insight into the structure of the local subspaces which determine a new

approximation at each step of the selected method. The results of this chapter will motivate the construction of trial subspaces and SPD preconditioners for symmetric eigenvalue (with a targeted interior eigenpair) and singular value problems in Chapter 4 and Chapter 5.

In Section 3.1, we present the simplest iterative scheme with stationary iteration parameters for solving a symmetric indefinite system with an SPD preconditioner. Other methods are obtained essentially by allowing to vary the parameters at each step of this stationary iteration in a way that a preconditioner-based norm of the residual vector is minimized. In Section 3.2, we present a notion of the *optimal SPD preconditioner* for a symmetric indefinite system, and suggest constructing preconditioners based on an approximation of the inverse of the absolute value of the coefficient matrix (*absolute value preconditioners*). We show on the example of a linear system with a discrete real Helmholtz operator (shifted Laplacian) that such preconditioners can be constructed in practice. Moreover, the use of the preconditioning techniques based, e.g., on multigrid (MG), can make this construction efficient.

3.1 Iterative methods for symmetric indefinite systems with SPD preconditioners

Given a (possibly nonsymmetric) matrix $A \in \mathbb{R}^{n \times n}$, an initial guess $x^{(0)} \in \mathbb{R}^n$, a (nonzero) iteration parameter $\alpha \in \mathbb{R}$, and a (possibly non-SPD) preconditioner $T \in \mathbb{R}^{n \times n}$, the iteration of the form

$$x^{(i+1)} = x^{(i)} + \alpha w^{(i)}, \ w^{(i)} = T r^{(i)}, \ r^{(i)} = b - A x^{(i)}, \ i = 0, 1, \ldots; \qquad (3.3)$$

where $x^{(i)} \in \mathbb{R}^n$ is an approximation to the exact solution of system (3.1) at iteration i, is commonly referred to as a preconditioned *stationary iteration*,

or *Richardson's method* with a stationary iteration parameter, see, e.g., [3, 33, 59]. We note that stationary iteration (3.3) can be considered as a simplest iterative method for solving linear systems, and, if properly preconditioned, can be efficient and computationally inexpensive.

In general, the (asymptotic) convergence rate of method (3.3) is governed by the spectral radius of the iteration matrix $M = I - \alpha TA$, where T and α (sometimes skipped after replacing αT by T) need to be chosen to make the spectral radius of M strictly less than 1, see, e.g., [3, 33]. If both A and T are SPD, one can always find a sufficiently small value of the step-size parameter $\alpha > 0$ to ensure that iteration (3.3) monotonically and linearly reduces the A-norm of error. Moreover, α can be set to the value which provides an optimal convergence rate for the method with an optimal convergence factor $\rho_{opt} = \frac{\kappa(TA)-1}{\kappa(TA)+1} < 1$, where the condition number $\kappa(TA)$ is a ratio of the largest and the smallest eigenvalues of the preconditioned matrix TA (for more details see, e.g., [3]).

If A is symmetric indefinite and T is SPD, stationary iteration (3.3), in general, diverges for any choice of the parameter α.

Proposition 3.1 *Stationary iteration (3.3) applied to linear system (3.1) with a symmetric indefinite matrix A and an SPD preconditioner T diverges for any α, unless the initial guess $x^{(0)}$ is specifically chosen.*

Proof: Let us consider the preconditioned residual, corresponding to iteration (3.3):

$$Tr^{(i+1)} = (I - \alpha TA)^{i+1} Tr^{(0)}, \ i = 0, 1, \ldots . \tag{3.4}$$

Since T is SPD and A is symmetric indefinite, the preconditioned matrix TA is T^{-1}-symmetric and indefinite with eigenpairs (λ_j, y_j), where the (real) eigenvalues λ_j are of both signs and the eigenvectors y_j are T^{-1}-orthonormal, $j = 1, \ldots, n$. Then the eigenpairs of the iteration matrix $I - \alpha TA$ are (μ_j, y_j) where $\mu_j = 1 - \alpha \lambda_j$.

Let c_j be the coordinates of the preconditioned initial residual vector $Tr^{(0)}$ in the basis of the eigenvectors y_j. Assume that the initial guess is chosen in such a way that $Tr^{(0)}$ has at least two nontrivial components in the directions of eigenvectors corresponding to eigenvalues of TA of the opposite sign. Then we can always fix an eigenvalue λ_{j_*} of the matrix TA such that $\mathrm{sign}(\lambda_{j_*}) = -\mathrm{sign}(\alpha)$, for any (nonzero) α, and the vector $Tr^{(0)}$ has a nontrivial component in the direction of y_{j_*}, i.e., $c_{j_*} \neq 0$. Thus, since the corresponding eigenvalue $\mu_{j_*} = 1 - \alpha \lambda_{j_*}$ of $I - \alpha TA$ is strictly greater than 1, using identity (3.4) and the Pythagorean theorem with respect to the T^{-1}-inner product, we obtain:

$$\|r^{(i+1)}\|_T^2 = \|Tr^{(i+1)}\|_{T^{-1}}^2 = \|(I - \alpha TA)^{i+1} Tr^{(0)}\|_{T^{-1}}^2$$

$$= \|(I - \alpha TA)^{i+1} \sum_{j=0}^{n} c_j y_j\|_{T^{-1}}^2 = \|\sum_{j=0}^{n} \mu_j^{i+1} c_j y_j\|_{T^{-1}}^2$$

$$= \sum_{j=0}^{n} \left(\mu_j^{i+1} c_j\right)^2 > \left(\mu_{j_*}^{i+1} c_{j_*}\right)^2 \to \infty,$$

as $i \to \infty$. This proves the divergence of iteration (3.3) with A symmetric indefinite and T SPD for any α, unless the initial guess $x^{(0)}$ is such that the vector Tr_0 has its nontrivial components only in the direction of eigenvectors corresponding to eigenvalues of the same sign. $\blacksquare$

Since iteration (3.3) is not applicable for solving symmetric indefinite systems with SPD preconditioners, we further question what a correct way is to

define a simple scheme with stationary iteration parameters, different from using method (3.3) for the normal equations, that can be applied in the described framework.

3.1.1 Stationary iteration for solving symmetric indefinite systems with SPD preconditioners

Given a symmetric indefinite matrix A and an SPD preconditioner T, we consider the iteration of the form

$$
\begin{aligned}
r^{(i)} &= b - Ax^{(i)}, \ w^{(i)} = Tr^{(i)}, \ s^{(i)} = TAw^{(i)}, \ l^{(i)} = s^{(i)} - \beta w^{(i)}, \\
x^{(i+1)} &= x^{(i)} + \alpha l^{(i)}, \ i = 0, 1, \ldots,
\end{aligned} \tag{3.5}
$$

where α (nonzero) and β are real numbers. Scheme (3.5) can be viewed as preconditioned stationary iteration (3.3) with a search direction $w^{(i)}$ replaced by a modified direction $l^{(i)}$ which is a linear combination of the preconditioned residual $w^{(i)}$ and a vector $s^{(i)} = TAw^{(i)}$. We notice that method (3.5) is exactly stationary iteration (3.3), applied to solve the (preconditioned, T^{-1}-symmetric) system

$$
(TA - \beta I) \, TAx = (TA - \beta I) \, Tb, \tag{3.6}
$$

or, equivalently, the symmetric system $(AT - \beta I) \, Ax = (AT - \beta I) \, b$ with the preconditioner T. Alternatively, (3.6) can be viewed as an instance of a polynomially preconditioned system, see, e.g., [59].

With $\beta = 0$, method (3.5) turns into iteration (3.3) applied to the system of the normal equations $(TA)^2 x = TATb$, or, equivalently, to the system $ATAx = ATb$, with an SPD matrix ATA, preconditioned with T. The optimal choice of the parameter α in this case leads to the (optimal) convergence rate

with a factor $\rho_{opt} = \frac{\kappa((TA)^2)-1}{\kappa((TA)^2)+1}$. We next show that for certain choices of the parameters α and β method (3.5) converges to the solution of system (3.1). Moreover, the (optimal) convergence rate is improved, depending on the eigenvalue distribution of the preconditioned matrix TA, compared to the above discussed approach based on solving the corresponding system of normal equations with method (3.3).

Let us assume that the spectrum of the preconditioned matrix TA, i.e., $\Lambda(TA) = \{\lambda_1, \ldots, \lambda_p, \lambda_{p+1}, \ldots, \lambda_n\}$, is located within the union of the two intervals

$$\mathcal{I} = [a, b] \bigcup [c, d], \tag{3.7}$$

where $a \leq \lambda_1 \leq \lambda_p \leq b < 0 < c \leq \lambda_{p+1} \leq \lambda_n \leq d$, and $\lambda_i \leq \lambda_{i+1}$, $i = 1, \ldots, n-1$. The following theorem holds:

Theorem 3.2 *Let us consider method (3.5) applied to solve linear system (3.1) with a symmetric indefinite coefficient matrix A and an SPD preconditioner T. Let us assume that the spectrum of the matrix TA is only known to be enclosed within the pair of intervals $\mathcal{I}$ in (3.7).*

If $b < \beta < c$ and $0 < \alpha < \tau_\beta$, where $\tau_\beta = 2/\max\limits_{\lambda \in \{a,d\}} (\lambda^2 - \beta\lambda)$, then

$$\frac{\|r^{(i+1)}\|_T}{\|r^{(i)}\|_T} \leq \rho < 1, \quad \text{where } \rho = \max\limits_{\lambda \in \{a,b,c,d\}} \left|1 - \alpha(\lambda^2 - \beta\lambda)\right|, \tag{3.8}$$

i.e., method (3.5) converges to the solution of system (3.1). Moreover, the convergence with the optimal convergence factor

$$\rho = \rho_{opt} = \frac{\tilde{\kappa} - 1}{\tilde{\kappa} + 1}, \quad \text{where } \tilde{\kappa} = \begin{cases} \left(\dfrac{d}{c}\right)\left(\dfrac{|b| + d - c}{|b|}\right), & \text{if } |a| - |b| \leq d - c \\ \left(\dfrac{a}{b}\right)\left(\dfrac{c + |a| - |b|}{c}\right), & \text{if } |a| - |b| > d - c \end{cases} \tag{3.9}$$

corresponds to the choice of parameters $\beta = \beta_{opt} = c - |b|$ *and* $\alpha = \alpha_{opt}$, *where*

$$\alpha_{opt} = \begin{cases} 2/(|b|\, c + d(|b| + d - c)), & \text{if } |a| - |b| \le d - c \\ 2/(|b|\, c + |a|\, (c + |a| - |b|)), & \text{if } |a| - |b| > d - c \end{cases} . \tag{3.10}$$

Proof: As has been mentioned, method (3.5) is exactly stationary iteration (3.3), applied to solve system (3.6), or, equivalently, the symmetric system $(AT - \beta I)\, Ax = (AT - \beta I)\, b$ with the preconditioner T. Thus, in order for method (3.5) to converge, by Proposition 3.1, the parameter β needs to be chosen such that the matrix $S_\beta = (TA - \beta I)\, TA$ in (3.6) is positive definite, i.e., all the eigenvalues μ_j of S_β are positive. Since $\mu_j = \lambda_j^2 - \beta\lambda_j$, where $\lambda_j \in \Lambda(TA)$, we conclude, by enforcing the parabola $\mu(\lambda) = \lambda^2 - \beta\lambda > 0$ on $\mathcal{I}$ (and hence on $\Lambda(TA) \subset \mathcal{I}$), that $\mu_j > 0$ if $b < \beta < c$ for all $j = 1, \ldots, n$.

Next we observe that the preconditioned residual, corresponding to a step of method (3.5), can be written as

$$Tr^{(i+1)} = (I - \alpha TA\, (TA + \beta I))\, Tr^{(i)} = (I - \alpha S_\beta)\, Tr^{(i)}.$$

Thus, using the derivations similar to those in Proposition 3.1, one gets the following inequality for the T-norms of the residual vectors at the consecutive iterations:

$$\|r^{(i+1)}\|_T \le \max_{\lambda_j \in \Lambda(TA)} \left| 1 - \alpha(\lambda_j^2 - \beta\lambda_j) \right| \|r^{(i)}\|_T \le \max_{\lambda \in \mathcal{I}} \left| 1 - \alpha(\lambda^2 - \beta\lambda) \right| \|r^{(i)}\|_T. \tag{3.11}$$

Since $\mu(\lambda) = \lambda^2 - \beta\lambda > 0$ for $\lambda \in \mathcal{I}$, provided that $b < \beta < c$, it is possible to choose a sufficiently small α in (3.11) such that $|1 - \alpha(\lambda^2 - \beta\lambda)| < 1$ on $\mathcal{I}$; i.e.,

$$0 < \alpha < 2/\max_{\lambda \in \mathcal{I}} \left(\lambda^2 - \beta\lambda \right) = 2/\max_{\lambda \in \{a,d\}} \left(\lambda^2 - \beta\lambda \right).$$

Therefore, the choice $b < \beta < c$ and $0 < \alpha < \tau_\beta$, where $\tau_\beta = \max\limits_{\lambda \in \{a,d\}} \left(\lambda^2 - \beta\lambda \right)$, implies, by (3.11), that $\|r^{(i+1)}\|_T / \|r^{(i+1)}\|_T \le \rho < 1$, where

$$\rho = \max_{\lambda \in \mathcal{I}} \left| 1 - \alpha(\lambda^2 - \beta\lambda) \right| = \max_{\lambda \in \{a,b,c,d\}} \left| 1 - \alpha(\lambda^2 - \beta\lambda) \right|.$$

This proves convergence bound (3.8).

Finally we determine the choice of the parameters $\alpha = \alpha_{opt}$ and $\beta = \beta_{opt}$ such that method (3.5) converges to the solution of system (3.1) with an optimal rate, i.e., with the convergence factor

$$\rho = \rho_{opt} = \min_{\alpha,\beta} \max_{\lambda \in \{a,b,c,d\}} \left| 1 - \alpha(\lambda^2 - \beta\lambda) \right| = \max_{\lambda \in \{a,b,c,d\}} \left| 1 - \alpha_{opt}(\lambda^2 - \beta_{opt}\lambda) \right|.$$

We note that, for any $b < \beta < c$, the corresponding optimal value of $\alpha = \alpha_{opt}(\beta)$ is

$$\alpha_{opt}(\beta) = \frac{2}{\min\limits_{\lambda \in \{b,c\}} (\lambda^2 - \beta\lambda) + \max\limits_{\lambda \in \{a,d\}} (\lambda^2 - \beta\lambda)}, \tag{3.12}$$

see, e.g., Axelsson [3, Theorem 5.6] for a detailed explanation. This choice of α leads to the convergence rate with the factor

$$\rho = \rho_{opt}(\beta) = \frac{\tilde{\kappa}(\beta) - 1}{\tilde{\kappa}(\beta) + 1}, \text{ where } \tilde{\kappa}(\beta) = \frac{\max\limits_{\lambda \in \{a,d\}} (\lambda^2 - \beta\lambda)}{\min\limits_{\lambda \in \{b,c\}} (\lambda^2 - \beta\lambda)}, \text{ for any } b < \beta < c. \tag{3.13}$$

Since β is assumed to be arbitrary from the interval (b, c), the above equality allows us to conclude that the optimal convergence rate of method (3.5) applied to solve system (3.1) occurs if β is chosen to minimize $\tilde{\kappa}(\beta)$ in (3.13). The latter is, in fact, equivalent to the observation that method (3.3) with an optimal choice of the parameter α applied to the family of systems (3.6) with (preconditioned) coefficient matrices $\{S_\beta = (TA + \beta I)TA\}$, $b < \beta < c$, delivers the best conver-

gence rate for the matrix $S_{\beta_{opt}}$ corresponding to $\beta = \beta_{opt}$, which minimizes the condition number of S_β.

Now let $\beta = \beta_{opt} = c - |b|$. Then, since $b^2 - \beta_{opt}b = c^2 - \beta_{opt}c = |b|\, c$, we have

$$\tilde{\kappa}(\beta_{opt}) = \begin{cases} \dfrac{d^2 - \beta_{opt}d}{|b|\, c}, & \text{if } |a| - |b| \le d - c \\[2mm] \dfrac{a^2 - \beta_{opt}a}{|b|\, c}, & \text{if } |a| - |b| > d - c \end{cases}$$

$$= \begin{cases} \left(\dfrac{d}{c}\right)\left(\dfrac{|b| + d - c}{|b|}\right), & \text{if } |a| - |b| \le d - c \\[2mm] \left(\dfrac{a}{b}\right)\left(\dfrac{c + |a| - |b|}{c}\right), & \text{if } |a| - |b| > d - c. \end{cases}$$

One can check that the above choice $\beta = \beta_{opt}$ indeed minimizes $\tilde{\kappa}(\beta)$ in (3.13), e.g., by adding an arbitrary perturbation ϵ to β_{opt} and showing that the function $\tilde{\kappa}(\beta_{opt} + \epsilon) \equiv \tilde{\kappa}(\epsilon)$ is increasing for $\epsilon > 0$ and decreasing for $\epsilon < 0$. This proves the optimal convergence rate of method (3.5) given by the factor ρ_{opt} in (3.9) with $\tilde{\kappa} = \tilde{\kappa}(\beta_{opt})$, where $\beta_{opt} = c - |b|$ and, by (3.12), $\alpha_{opt} = \alpha_{opt}(\beta_{opt})$, i.e.,

$$\alpha_{opt} = \frac{2}{|b|\, c + \max\limits_{\lambda \in \{a,d\}} (\lambda^2 - \beta_{opt}\lambda)},$$

which results in expression (3.10). $\blacksquare$

We note that if $|a| - |b| = d - c$, i.e., both intervals in (3.7) are of the same length, then the optimal convergence factor $\rho = \rho_{opt}$ in (3.9) is determined by $\tilde{\kappa} = \dfrac{ad}{bc}$. Although the proof of Theorem 3.2 does not rely on this assumption, it is clear that for the general case, where $|a| - |b| \ne d - c$, the expression for $\tilde{\kappa}$ in (3.9) can be derived after extending the smaller interval to match the length of the larger one by shifting the corresponding endpoint a or d, and then applying the result for the intervals of the equal length. We also note that if $[a, b]$ and $[c, d]$

are located symmetrically with respect to the origin, i.e., $|a| = d$ and $|b| = c$, then $\beta_{opt} = 0$ and method (3.5) turns into stationary iteration (3.3) applied to normal equations with the optimal convergence rate determined by $\tilde{\kappa} = \left(\dfrac{a}{b}\right)^2$ which is essentially a square of the condition number of the matrix TA.

Finally, we remark that the idea of transforming the original symmetric indefinite system (3.1) into an SPD system (3.6) with a minimized condition number, which underlies method (3.5) and Theorem 3.2, has previously appeared in literature, though without a preconditioner, e.g., in [3] in the context of the Chebyshev iteration.

We will use scheme (3.5) as a base for obtaining simple preconditioned residual-minimizing methods to solve system (3.1). Theorem 3.2 will allow us to provide the corresponding convergence estimates.

3.1.2 Simple residual-minimizing methods for solving symmetric indefinite systems with SPD preconditioners

Let us consider the following iterative scheme for solving a symmetric indefinite system (3.1) with an SPD preconditioner T and a fixed parameter β:

$$
\begin{aligned}
l^{(i)} &= s^{(i)} - \beta w^{(i)}, \ s^{(i)} = TAw^{(i)}, \ w^{(i)} = Tr^{(i)}, \ r^{(i)} = b - Ax^{(i)}, \\
x^{(i+1)} &= x^{(i)} + \alpha^{(i)} l^{(i)}, \ \alpha^{(i)} = \tfrac{(w^{(i)}, Al^{(i)})}{(Al^{(i)}, TAl^{(i)})}, \ b < \beta < c, \ i = 0, 1, \ldots,
\end{aligned}
\tag{3.14}
$$

where b and c are the endpoints of the intervals $\mathcal{I}$ in (3.7). Unlike stationary iteration (3.5), method (3.14) allows the parameters $\alpha^{(i)}$ to vary at each step such that the next approximation $x^{(i+1)}$ corresponds to the residual vector with the smallest T-norm in the affine space $r^{(i)} + \text{span}\left\{Al^{(i)}\right\}$, i.e.,

$$
\alpha^{(i)} = \operatorname*{argmin}_{\alpha \in \mathbb{R}} \|r^{(i)} - \alpha Al^{(i)}\|_T.
\tag{3.15}
$$

The following theorem shows that method (3.14) converges to the exact solution of system (3.1) for any $b < \beta < c$, moreover the choice of $\beta = \beta_{opt} = c - |b|$ guarantees that (3.14) converges not slower than stationary iteration (3.5) with optimal parameters.

Theorem 3.3 *Let us consider method (3.14) applied to solve linear system (3.1) with a symmetric indefinite coefficient matrix A and an SPD preconditioner T. We assume that the spectrum of the matrix TA is only known to be enclosed within the pair of intervals $\mathcal{I}$ in (3.7).*

If $b < \beta < c$, then

$$\frac{\|r^{(i+1)}\|_T}{\|r^{(i)}\|_T} \leq \rho < 1, \ \ where \ \rho = \frac{\tilde{\kappa} - 1}{\tilde{\kappa} + 1}, \ \ \tilde{\kappa} = \frac{\max\limits_{\lambda \in \{a,d\}} (\lambda^2 - \beta\lambda)}{\min\limits_{\lambda \in \{b,c\}} (\lambda^2 - \beta\lambda)}. \tag{3.16}$$

Moreover, if $\beta = \beta_{opt} = c - |b|$, then $\tilde{\kappa}$ is defined by (3.9).

Proof: By (3.15) we have

$$\|r^{(i+1)}\|_T = \|r^{(i)} - \alpha^{(i)} A l^{(i)}\|_T \leq \|r^{(i)} - \alpha A l^{(i)}\|_T \ \forall \alpha \in \mathbb{R}.$$

Let us assume that β is fixed, such that $b < \beta < c$. Then, following the proof of Theorem 3.2, the choice of $\alpha = \alpha_{opt}(\beta)$ as in (3.12), by (3.11), leads to expression (3.13) for the convergence factor ρ in (3.16). One can verify that if $\beta = \beta_{opt} = c - |b|$ then $\tilde{\kappa}$ is defined by (3.9). ∎

Given a constant β, e.g., provided by information about the spectrum location or computational experience, scheme (3.14) represents the simplest residual-minimizing method with the minimization at a step i performed over the one-dimensional subspace span $\{Al^{(i)}\}$, moreover the resulting convergence behavior

is in general improved compared to the one of the corresponding methods based on solving normal equations.

If no information for the choice of β is available, one can allow it to vary at each step. For example, let us consider the following iterative scheme:

$$\begin{aligned} l^{(i)} &= s^{(i)} - \beta^{(i)} w^{(i)}, \ \ s^{(i)} = TAw^{(i)}, \ \ w^{(i)} = Tr^{(i)}, \ \ r^{(i)} = b - Ax^{(i)}, \\ x^{(i+1)} &= x^{(i)} + \alpha^{(i)} l^{(i)}, \ \ i = 0, 1, \dots, \end{aligned} \tag{3.17}$$

where the parameters $\alpha^{(i)}$ and $\beta^{(i)}$ are chosen to guarantee the minimality of the T-norm of the next residual vector $r^{(i+1)}$ over the affine space $r^{(i)} + \text{span}\left\{Aw^{(i)}, As^{(i)}\right\}$, i.e., $\alpha^{(i)}$ and $\beta^{(i)}$ in (3.17) are such that

$$\|r^{(i+1)}\|_T = \min_{u \in \text{span}\left\{Aw^{(i)}, As^{(i)}\right\}} \|r^{(i)} - u\|_T. \tag{3.18}$$

Optimality condition (3.18) is equivalent to the following orthogonality conditions

$$(r^{(i+1)}, As^{(i)})_T = (r^{(i+1)}, Aw^{(i)})_T = 0,$$

which provide the expressions for the iteration parameters:

$$\beta^{(i)} = \frac{(s^{(i)}, As^{(i)})(w^{(i)}, As^{(i)}) - (TAs^{(i)}, As^{(i)})(w^{(i)}, Aw^{(i)})}{(w^{(i)}, As^{(i)})^2 - (w^{(i)}, Aw^{(i)})(s^{(i)}, As^{(i)})},$$

$$\alpha^{(i)} = \frac{(w^{(i)}, Al^{(i)})}{(Al^{(i)}, TAl^{(i)})} = \frac{(w^{(i)}, As^{(i)})^2 - (w^{(i)}, Aw^{(i)})(s^{(i)}, As^{(i)})}{(TAs^{(i)}, As^{(i)})(s^{(i)}, Aw^{(i)}) - (s^{(i)}, As^{(i)})^2}. \tag{3.19}$$

We note that along with expressions (3.19) for the choice of the parameters, method (3.17)–(3.18) can admit other implementations, e.g., based on the properly restarted Lanczos procedure. The following theorem provides a bound on the convergence rate of scheme (3.17)–(3.18).

Theorem 3.4 *We consider method (3.17)–(3.18) applied to solve the linear system (3.1) with a symmetric indefinite coefficient matrix A and an SPD preconditioner T. We assume that the spectrum of the matrix TA is only known to be enclosed within the pair of intervals $\mathcal{I}$ in (3.7). Then at each step of the method the T-norm of the residual vector is reduced at least by the factor ρ in (3.9), i.e.,*

$$\frac{\|r^{(i+1)}\|_T}{\|r^{(i)}\|_T} \leq \frac{\tilde{\kappa} - 1}{\tilde{\kappa} + 1}. \tag{3.20}$$

Proof: Since $\alpha^{(i)}$ and $\beta^{(i)}$ are such that $\|r^{(i+1)}\|$ has the smallest T-norm over $r^{(i)} + \text{span}\left\{Aw^{(i)}, As^{(i)}\right\}$, we get

$$\|r^{(i+1)}\|_T = \|r^{(i)} - \alpha^{(i)}Al^{(i)}\|_T = \|r^{(i)} - \alpha^{(i)}As^{(i)} + \alpha^{(i)}\beta^{(i)}Aw^{(i)}\|_T$$

$$\leq \|r^{(i)} - \alpha A\left(s^{(i)} - \beta w^{(i)}\right)\|_T, \ \forall \alpha, \beta \in \mathbb{R}.$$

The choice of $\beta = \beta_{opt} = c - |b|$ and $\alpha = \alpha_{opt}$ in (3.10), by Theorem 3.2, results in convergence factor (3.9) for the reduction of the T-norm of the residual vector at each step of method (3.17)–(3.18). ∎

We remark that method (3.17)–(3.18), for solving symmetric indefinite linear systems with SPD preconditioners, described by convergence estimate (3.9), (3.20), can be viewed as an analogue of the preconditioned steepest descent iteration for solving SPD systems. In the next section we discuss methods, including the optimal minimal residual iterations, which allow us to improve convergence factor (3.9).

3.1.3 The second-order and minimal residual methods for solving indefinite systems with SPD preconditioners

The ideas underlying methods (3.5), (3.14)–(3.15) and (3.17)–(3.18) can be further extended to improve convergence factor (3.9). In particular, applying

the so-called *second-order* stationary iteration, i.e., the iteration of form (3.3) with the additional term in the direction of the difference $p^{(i)} = x^{(i)} - x^{(i-1)}$ of approximations from the current and previous steps, to transformed system (3.6), results in the following scheme for solving system (3.1) with an SPD preconditioner T:

$$l^{(i)} = s^{(i)} - \beta^{(i)} w^{(i)}, \ s^{(i)} = TAw^{(i)}, \ w^{(i)} = Tr^{(i)}, \ r^{(i)} = b - Ax^{(i)},$$

$$x^{(i+1)} = x^{(i)} + \alpha^{(i)} l^{(i)} + (\gamma^{(i)} - 1)p^{(i)}, p^{(i)} = x^{(i)} - x^{(i-1)} , \ p^{(0)} = x^{(0)}, \quad (3.21)$$

$$i = 0, 1, \ldots,$$

where parameters $\alpha^{(i)} = \alpha$, $\beta^{(i)} = \beta$, $\gamma^{(i)} = \gamma \in \mathbb{R}$ are constant throughout iterations. If, similarly to Theorem 3.2 for method (3.5), the parameter β is set to $\beta_{opt} = c - |b|$, then it is possible to show that there exist optimal values for α and γ such that scheme (3.21), with the stationary iteration parameters, converges to the solution of (3.1) with the asymptotically average convergence factor

$$\rho_{avg} = \frac{\sqrt{\tilde{\kappa}} - 1}{\sqrt{\tilde{\kappa}} + 1}, \quad (3.22)$$

where $\tilde{\kappa}$ is defined in (3.9). In particular, the latter can be shown, e.g., by using the convergence bound in Axelsson [3, Theorem 5.9] for the second-order stationary iteration applied to transformed system (3.6) with $\beta = \beta_{opt}$. Thus, the convergence of method (3.21) with the optimal choice of stationary parameters is given by

$$\frac{\|r^{(i)}\|_T}{\|r^{(0)}\|_T} \leq C\rho_{avg}^i, \quad (3.23)$$

where C is a positive constant, and ρ_{avg} is defined in (3.22).

In the same way as stationary scheme (3.5) has been extended to method (3.17)–(3.18), the second-order method (3.21), with $\alpha^{(i)} = \alpha$, $\beta^{(i)} = \beta$ and

$\gamma^{(i)} = \gamma$, can be generalized to have variable iteration parameters, chosen at each step to minimize the T-norm of the next residual vector $r^{(i+1)}$ in the affine space $r^{(i)} + \mathrm{span}\left\{Aw^{(i)}, As^{(i)}, Ap^{(i)}\right\}$, i.e.,

$$\|r^{(i+1)}\|_T = \min_{u \in \mathrm{span}\left\{Aw^{(i)}, As^{(i)}, Ap^{(i)}\right\}} \|r^{(i)} - u\|_T. \qquad (3.24)$$

It is immediately seen that one step of method (3.21), (3.24) results in the reduction of the residual T-norm, which is not worse than that provided by method (3.17)–(3.18), hence, convergence bound (3.9), (3.20) is valid for iteration (3.21), (3.24). We remark that the latter bound is likely to be pessimistic for method (3.21), (3.24), and, in practice, according to estimate (3.23) for iteration (3.21) with optimal stationary parameters, it is reasonable to expect the reduction of the residual norm by a factor of order (3.22).

Methods (3.17)–(3.18) and (3.21), (3.24) are examples of convergent *locally optimal* preconditioned methods for solving a symmetric indefinite system (3.1) with an SPD preconditioner, based on the idea of the residual norm minimization. The local optimality follows from the corresponding conditions (3.18) and (3.24), which, at each step, seek to minimize the residual T-norm over certain low-dimensional, local, subspaces of a fixed size.

As opposed to locally optimal methods, the preconditioned *globally optimal* residual-minimizing methods for solving system (3.1), at each step i, extract a minimizer for the appropriate residual norm from an (expanding) i-dimensional subspace. We now define the (globally optimal) Krylov subspace preconditioned minimal residual methods for solving system (3.1) with a preconditioner T.

Definition 3.5 *We say that a method to solve system (3.1) is a preconditioned minimal residual method, if, at step i, it constructs an approximation $x^{(i)}$ to the*

solution of system (3.1) of the form

$$x^{(i)} \in x^{(0)} + \mathcal{K}_i \left(TA, Tr^{(0)} \right), \tag{3.25}$$

and the corresponding residual vector $r^{(i)} = b - Ax^{(i)}$ is such that

$$\|r^{(i)}\|_S = \min_{u \in A\mathcal{K}_i \left(TA, Tr^{(0)} \right)} \|r^{(0)} - u\|_S, \tag{3.26}$$

where $\mathcal{K}_i \left(TA, Tr^{(0)} \right) = span \left\{ Tr^{(0)}, (TA)Tr^{(0)}, \dots, (TA)^{i-1}Tr^{(0)} \right\}$ is the (preconditioned) Krylov subspace generated by the matrix TA and the vector $Tr^{(0)}$, $A\mathcal{K}_i \left(TA, Tr^{(0)} \right) = span \left\{ (AT)r^{(0)}, \dots, (AT)^i r^{(0)} \right\}$ is the corresponding Krylov residual subspace; $\|x\|_S^2 = (x, Sx)$ for some SPD operator S.

In particular, for general (square) matrices T and A, the preconditioned minimal residual method with $S = T^*T$ in (3.26) is delivered, e.g., by the preconditioned GMRES [61, 33, 59]. The case where A is symmetric indefinite and T is SPD with $S = T$ is commonly fulfilled with the preconditioned MINRES algorithm (PMINRES) [55, 33, 59], which is known to admit a short-term recurrent form while maintaining global optimality (3.26) in exact arithmetic. Scheme (3.17)–(3.18) corresponds to the preconditioned minimal residual method with $S = T$ restarted after every two steps. Iteration (3.21) with variable parameters chosen according to (3.24) can be viewed as the same preconditioned minimal residual method restarted after every two steps with the additional vector $p^{(i)} = x^{(i)} - x^{(i-1)}$.

Finally, let us note that convergence factor (3.22) is commonly used to estimate the convergence rate of the preconditioned minimal residual method (3.25)–(3.26) with $S = T$ (e.g., in PMINRES implementation), once the residual

norms are measured at every other step, i.e.,

$$\frac{\|r^{(i)}\|_T}{\|r^{(0)}\|_T} \leq 2\rho_{avg}^j, \ \ i = 2j, \ j = 1, 2, 3, \ldots, \tag{3.27}$$

see, e.g., [33, 59].

In the next section, we define the optimal SPD preconditioner T for minimal residual methods (3.25)–(3.26), as well as for the locally optimal methods described in the current section, applied to solve system (3.1) with a symmetric indefinite coefficient matrix A.

3.2 Absolute value preconditioners for symmetric indefinite systems

In this section, we propose a novel concept of absolute value preconditioning, where the preconditioner approximates the absolute value of the coefficient matrix. We show, for a model problem, that such a preconditioner can be efficiently constructed in the multigrid framework.

3.2.1 Optimal SPD preconditioners for symmetric indefinite systems

Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix with an eigendecomposition $A = V\Lambda V^*$, where V is an orthogonal matrix of eigenvectors and $\Lambda = \text{diag}\{\lambda_j\}$, $j = 1, \ldots, n$, is a diagonal matrix of eigenvalues of A. We consider the factorization of the form

$$A = |A| \text{sign}(A) = \text{sign}(A) |A|, \tag{3.28}$$

where $|A| = V |\Lambda| V^*$ is an (SPD) absolute value of the matrix A (*matrix absolute value*), $|\Lambda| = \text{diag}\{|\lambda_j|\}$, and $\text{sign}(A) = V\text{sign}(\Lambda)V^*$ is a sign of A (*matrix sign*), $\text{sign}(\Lambda) = \text{diag}\{\text{sign}(\lambda_j)\}$. Factorization (3.28) is, in fact, a *polar decomposition*, see, e.g., [42], of the symmetric matrix A, with the positive (semi) definite factor $|A|$ and the orthogonal factor $\text{sign}(A)$.

The following theorem states that the inverse of the absolute value of the coefficient matrix is an *optimal* SPD preconditioner for the methods described in the previous section, including minimal residual methods (3.25)–(3.26), applied to solve a (general) symmetric indefinite linear system, i.e., $T = T_{opt} = |A|^{-1}$.

Theorem 3.6 *Any minimal residual method (3.25)–(3.26), applied to solve linear system (3.1) with a symmetric indefinite coefficient matrix A and the preconditioner $T = |A|^{-1}$, converges to the exact solution in at most two steps. Further, under the same assumptions on A and T, methods (3.14)–(3.15), (3.17)–(3.18), and (3.21) satisfying (3.24), as well as schemes (3.5) and (3.21) with corresponding optimal stationary iteration parameters, deliver the exact solution in exactly one step.*

Proof: Minimization property (3.26) at a step i of a preconditioned minimal residual method can be equivalently written as

$$\|r^{(i)}\|_S = \min_{p \in \mathcal{P}_i,\ p(0)=1} \|p(AT)r^{(0)}\|_S, \tag{3.29}$$

where $\mathcal{P}_i$ is a set of all polynomials of degree at most i. Then, according to the decomposition (3.28), the choice $T = |A|^{-1}$ results in the matrix $AT = \text{sign}(A)$ with only two distinct eigenvalues: -1 and 1. Hence the minimal polynomial of AT is of the second degree. Thus, by (3.29), $\|r^{(i)}\|_S = 0$ for at most $i = 2$ and any SPD operator S. Thus any preconditioned minimal residual method (3.25)–(3.26) converges to the exact solution of the symmetric indefinite system (3.1) with $T = |A|^{-1}$ in at most two steps.

The one-step convergence of the remaining methods follows from the observation that factors (3.9) and (3.22) are zero, since $\tilde{\kappa} = 1$ if $T = |A|^{-1}$. ∎

66

If A is SPD, the claim of the theorem reduces to the trivial fact that the optimal preconditioner for system (3.1) is the exact inverse of A. We also note that actual implementations of minimal residual methods (3.25)–(3.26), at their two consecutive iterations, perform essentially the same, in terms of matrix-vector multiplications, number of computations as one step of methods (3.14)–(3.15), (3.17)–(3.18), (3.21) with condition (3.24), as well as stationary iterations (3.5) and (3.21) with the optimal choice of iteration parameters. In this sense, the two-step optimality result for a minimal residual method, given by Theorem 3.6, is compatible with the optimal one-step convergence of the above mentioned methods, described in this section. The latter also explains the compatibility of convergence estimates (3.23) and (3.27).

Remark 3.7 *Methods of form (3.17) and (3.21) with $T = |A|^{-1}$ and the corresponding optimality conditions (3.18) and (3.24) replaced by the residual minimization in an arbitrary S-norm, i.e.,*

$$\|r^{(i+1)}\|_S = \min_{u \in span\left\{Aw^{(i)}, As^{(i)}\right\}} \|r^{(i)} - u\|_S, \tag{3.30}$$

and,

$$\|r^{(i+1)}\|_S = \min_{u \in span\left\{Aw^{(i)}, As^{(i)}, Ap^{(i)}\right\}} \|r^{(i)} - u\|_S, \tag{3.31}$$

respectively, also converge to the exact solution of symmetric indefinite system (3.1) in exactly one step for any SPD operator S.

In practical situations the construction of the *optimal* preconditioner $T_{opt} = |A|^{-1}$ becomes prohibitive. We show, however, that the choice of the preconditioner T as some *approximation* of T_{opt}, i.e., $T \approx |A|^{-1}$, may lead to a significant

improvement in the convergence rate of an iterative method. For example, the preconditioners $T \approx |A|^{-1}$ can be constructed by exactly inverting the absolute value of a *symmetric approximation* of the coefficient matrix A, assuming that the latter can be efficiently performed. In particular, if A is diagonally dominant, then T can be chosen to be diagonal,

$$T = \operatorname{diag}\left\{|a_{jj}|^{-1}\right\},$$

where a_{jj} are the diagonal entries of A. Let us agree to call an SPD preconditioner T, such that $T \approx |A|^{-1}$, *an absolute value preconditioner* for a linear system (3.1).

Due to a large problem size, we further assume that an absolute value preconditioner T can be accessed only indirectly, e.g., through a matrix-vector multiplication. In this case, given a vector $r \in \mathbb{R}^n$, there are several ways to approach the construction of T by defining a vector $w = Tr$. As the first option, at each step i of any method described in the previous section, one can attempt to apply the absolute value preconditioner by approximately solving for z the following equation

$$|A|\, z = r, \tag{3.32}$$

where, e.g., $r = r^{(i)}$ or/and $r = ATr^{(i)}$, depending on the selected method.

The coefficient matrix $|A|$ is generally not available. The problem of approximately solving linear system (3.32) can be formally replaced by the problem of finding a vector w which approximates the action of the *matrix function* $f(A) = |A|^{-1}$ on the vector r, i.e., $w \approx f(A)r = |A|^{-1}r$, moreover the construction of w does not require any knowledge of $|A|$ or $|A|^{-1}$. The latter constitutes

a well established task in matrix function computations which is standardly fulfilled by a Krylov subspace method, e.g., [39, 32]. Our numerical experience shows that though the convergence rate of a linear solver can be significantly improved with this approach, the computational costs of the existing methods for approximating $f(A)r = |A|^{-1} r$, e.g., the Lanczos method described in [12], however, remain too high for their direct use in the context of absolute value preconditioners for solving symmetric indefinite linear systems.

Another option to apply an absolute value preconditioner is to use a method, based on a certain preconditioning technique, which is possibly divergent as a stand-alone approximate solver for equation (3.32), e.g., since only limited information about the coefficient matrix $|A|$ is available, however, which (implicitly) results in the construction of an approximation to $|A|^{-1}$ of a reasonably good quality. Below we demonstrate on the example of a model problem that such construction of an efficient absolute value preconditioner is indeed possible, e.g., if based on MG techniques.

3.2.2 An absolute value preconditioner for a model problem

Let us consider the following boundary value problem,

$$
-\Delta u(\mathrm{x},\mathrm{y}) - c^2 u(\mathrm{x},\mathrm{y}) = f(\mathrm{x},\mathrm{y}), (\mathrm{x},\mathrm{y}) \in \Omega = (0,1) \times (0,1),
$$
$$
u|_\Gamma = 0,
$$

(3.33)

where $-\Delta = -\dfrac{\partial^2}{\partial \mathrm{x}^2} - \dfrac{\partial^2}{\partial \mathrm{y}^2}$ is the negative Laplace operator (or, Laplacian), $c \in \mathbb{R}$, $f(\mathrm{x},\mathrm{y}) \in \mathcal{C}(\Omega)$, and Γ denotes the boundary of the domain Ω. Problem (3.33) is, in fact, a particular instance of the Helmholtz equation with Dirichlet boundary conditions, c^2 is a wavenumber; see, e.g., [69].

After introducing a uniform grid of the step size h (mesh size) and using the standard 5-point finite-difference (FD) stencil to discretize continuous problem (3.33), see, e.g., [30], one obtains the corresponding discrete problem, i.e., system of linear equations (3.1) of the form

$$(L - c^2 I)x = b, \tag{3.34}$$

where the coefficient matrix $A = L - c^2 I$ (the discrete Helmholtz operator) represents a discrete negative Laplace operator L, satisfying the Dirichlet boundary condition at the grid points on the boundary, shifted by a scalar c^2 times the identity matrix I. The right-hand side b in (3.34) corresponds to the vector of function values of $f(\mathrm{x}, \mathrm{y})$ calculated at the grid points (numbered in the lexicographical order). In our numerical tests, b is generated randomly. The exact solution $x = x^*$ of system (3.34) then provides an approximation to the solution of the boundary value problem (3.33) evaluated at the grid points.

Further, assuming that c^2 is different from any eigenvalue of the SPD negative Laplacian L and is greater than its smallest, however less than its largest, eigenvalue, i.e., $\lambda_{min}(L) < c^2 < \lambda_{max}(L)$, where $\lambda_{min}(L) = 2\pi^2 + \mathcal{O}(h^2)$ and $\lambda_{max}(L) = 8h^{-2} + \mathcal{O}(1)$, we conclude that the operator $A = L - c^2 I$ is non-singular symmetric indefinite. Thus, in order to solve system (3.34), according to Theorem 3.6, one can choose any of the methods from the previous section with an (absolute value) preconditioner T approximating the operator $|A|^{-1} = |L - c^2 I|^{-1}$. Below we use the MG techniques to provide an examples of such preconditioner. We refer to (3.34) as the *model problem.*

3.2.2.1 Multigrid absolute value preconditioner

In this section we use the ideas underlying the (geometric) MG methods, e.g., [73, 14], to construct a preconditioner for the model symmetric indefinite system (3.34). Combining the MG principles with the idea of the absolute value preconditioners (Theorem 3.6), we construct an efficient preconditioner for the model problem with low wavenumbers c^2, i.e., if the operator $A = L - c^2 I$ in (3.34) is slightly indefinite. We compare the proposed approach with a preconditioning strategy based on the inverse of the Laplacian, which we set as a benchmark to assess the quality of the constructed preconditioner.

Along with the (fine) grid of the mesh size h underlying the discretized Helmholtz equation (3.34) let us consider a (coarse) grid of a mesh size $H > h$. We denote the discretization of the negative Laplacian on this grid by L_H, I_H represents the identity operator of the corresponding dimension. Further, we assume that the fine-level absolute value $|L - c^2 I|$ is not computable, while its coarse-level analogue $|L_H - c^2 I_H|$ can be efficiently constructed and/or inverted, e.g., by the full eigendecomposition. Let us note that in the two-grid framework we use the subscript H to refer to the quantities defined on the coarse grid. No subscript is used for denoting the fine-grid quantities.

We suggest the following scheme as an example of the two-grid absolute value preconditioner for model problem (3.34).

Algorithm 3.8 (Two-grid absolute value preconditioner)

Input r, output w.

1. *Pre-smoothing. Apply ν pre-smoothing steps with the zero initial guess $(w^{(0)} = 0)$:*

$$w^{(i+1)} = w^{(i)} + M^{-1}(r - Lw^{(i)}), \;\; i = 0, \ldots, \nu - 1, \qquad (3.35)$$

   where the (nonsingular) matrix M defines the choice of a smoother. This step results in the pre-smoothed vector $w^{pre} = w^{(\nu)}$, $\nu \geq 1$.

2. *Coarse grid correction. Restrict the vector $r - Lw^{pre}$ to the coarse grid, multiply it by the inverted coarse-level absolute value $|L_H - c^2 I_H|$, and then prolongate the result back to the fine grid. This delivers the coarse-grid correction, which is added to w^{pre} to obtain the corrected vector w^{cgc}:*

$$w_H = \left| L_H - c^2 I_H \right|^{-1} R\left(r - Lw^{pre}\right), \qquad (3.36)$$

$$w^{cgc} = w^{pre} + Pw_H, \qquad (3.37)$$

   where P and R are prolongation and restriction operators, respectively.

3. *Post-smoothing. Apply ν post-smoothing steps with the initial guess $w^{(0)} = w^{cgc}$:*

$$w^{(i+1)} = w^{(i)} + M^{-*}(r - Lw^{(i)}), \;\; i = 0, \ldots, \nu - 1. \qquad (3.38)$$

   This step results in the post-smoothed vector $w^{post} = w^{(\nu)}$. Return $w = w^{post}$.

In (3.36) we have assumed that the coarse-grid operator $|L_H - c^2 I_H|$ is invertible, i.e., c^2 is different from any eigenvalue of L_H. The number of smoothing steps

in (3.35) and (3.38) is the same; the pre-smoother is defined by the nonsingular matrix M, while the post-smoother is delivered by M^*.

We note that once the absolute value of the discrete Helmholtz operator $L - c^2 I$ on the fine level is not available, one can attempt to replace it by an easily accessible SPD approximation, e.g., the negative Laplacian $L = |L| \approx |L - c^2 I|$, as was done in Algorithm 3.8. Although this substitution may result in the divergence of the algorithm as a two-grid method for solving equation (3.32), we show that its use as a preconditioner (along with its multigrid extension) allows us to noticeably accelerate the convergence of the methods described in the previous section applied to solve the symmetric indefinite model problem (3.34) for shifts c^2 of a relatively small size.

One can check that the two-grid Algorithm 3.8 implicitly constructs a mapping $r \mapsto w = T_{tg} r$, where the operator $T = T_{tg}$ has the following structure:

$$T_{tg} = \left(I - M^{-*}L \right)^{\nu} P \left| L_H - c^2 I_H \right|^{-1} R \left(I - LM^{-1} \right)^{\nu} + S, \tag{3.39}$$

with $S = L^{-1} - \left(I - M^{-*}L \right)^{\nu} L^{-1} \left(I - LM^{-1} \right)^{\nu}$. In particular, in the context of methods from the previous section, at each iteration i, the vector r is set to $r^{(i)}$ or/and $ATr^{(i)}$, where $r^{(i)} = b - (L - c^2 I)x^{(i)}$ is the residual vector of problem (3.34) at the i-th step of the corresponding method. The fact that the constructed preconditioner $T = T_{tg}$ is SPD, follows directly from the observation that the first term in (3.39) is symmetric positive semi-definite provided that $P = \alpha R^*$ for some nonzero scalar α, while the second term S is symmetric and positive definite if the spectral radii $\rho(I - M^{-1}L) < 1$ and $\rho(I - M^{-*}L) < 1$. The latter condition, in fact, requires the pre- and post-smoothing iterations (steps 1 and 3 of Algorithm 3.8) to represent convergent methods for system

(3.34) with $c = 0$ and $b = r$ (i.e., for the discrete Poisson's equation) on their own. We note that the above argument for the operator $T = T_{tg}$ to be SPD essentially repeats the corresponding pattern to justify symmetry and positive definiteness of a two-grid preconditioner applied within an iterative scheme, e.g., preconditioned conjugate gradient method (PCG), to solve a system of linear equations with an SPD coefficient matrix; see, e.g., [13, 67].

Now let us consider a hierarchy of $m + 1$ grids numbered by $l = m, m - 1, \dots, 0$ with the corresponding mesh sizes $\{h_l\}$ in the decreasing order ($h_m = h$ corresponds to the finest, and h_0 to the coarsest, grid). For each level l we define the discretization $L_l - c^2 I_l$ of the differential operator in (3.33), where L_l is the discrete negative Laplacian on grid l, and I_l is the identity of the same size.

In order to extend the two-grid absolute value preconditioner given by Algorithm 3.8 to the *multigrid*, instead of inverting the absolute value $|L_H - c^2 I_H|$ in step 2 (formula (3.36)), we recursively apply the algorithm to the restricted vector $R(r - Lw^{pre})$. This pattern is then followed, in the V-cycle "fashion", on all levels, with the exact inversion of the absolute value of the discrete Helmholtz operator on the coarsest grid. The described approach can be viewed as replacing w_H in (3.36) by its approximation, i.e., constructing

$$w_H \approx |L_H - c^2 I_H|^{-1} R(r - Lw^{pre}).$$

If started from the finest grid $l = m$, the following scheme gives the multilevel extension of the two-grid absolute value preconditioner defined by Algorithm 3.8. We note that the subscript l is introduced to match the occurring quantities to the corresponding grid.

Algorithm 3.9 (AVP-MG(r_l): MG absolute value preconditioner)

Input r_l, output w_l.

1. *Pre-smoothing. Apply ν pre-smoothing steps with the zero initial guess $(w_l^{(0)} = 0)$:*

$$w_l^{(i+1)} = w_l^{(i)} + M_l^{-1}(r_l - L_l w_l^{(i)}), \ \ i = 0, \ldots, \nu - 1, \qquad (3.40)$$

where the (nonsingular) matrix M_l defines the choice of a smoother on level l. This step results in the pre-smoothed vector $w_l^{pre} = w_l^{(\nu)}$, $\nu \geq 1$.

2. *Coarse grid correction. Restrict the vector $r_l - L_l w_l^{pre}$ to the grid $l - 1$. If $l = 1$, then multiply the restricted vector by the inverted coarse-level absolute value $|L_0 - c^2 I_0|$,*

$$w_0 = \left| L_0 - c^2 I_0 \right|^{-1} R_0 \left(r_1 - L_1 w_1^{pre} \right), \ \ if \ l = 1. \qquad (3.41)$$

Otherwise, recursively apply AVP-MG to approximate the action of the inverted absolute value $|L_{l-1} - c^2 I_{l-1}|$ on the restricted vector,

$$w_{l-1} = AVP\text{-}MG \left(R_{l-1} \left(r_l - L_l w_l^{pre} \right) \right), \ \ if \ l > 1. \qquad (3.42)$$

Prolongate the result back to the fine grid. This delivers the coarse-grid correction, which is added to w_l^{pre} to obtain the corrected vector w_l^{cgc}:

$$w_l^{cgc} = w_l^{pre} + P_l w_{l-1}, \qquad (3.43)$$

where w_{l-1} is given by (3.41)–(3.42). The operators R_{l-1} and P_l define the restriction from the level l to $l - 1$ and the prolongation from the level $l - 1$ to l, respectively.

3. *Post-smoothing. Apply ν post-smoothing steps with the initial guess $w_l^{(0)} = w_l^{cgc}$:*

$$w_l^{(i+1)} = w_l^{(i)} + M_l^{-*}(r_l - L_l w_l^{(i)}), \ i = 0, \ldots, \nu - 1. \tag{3.44}$$

This step results in the post-smoothed vector $w_l^{post} = w_l^{(\nu)}$. Return $w_l = w_l^{post}$.

The described multigrid absolute value preconditioner implicitly constructs a mapping $r \mapsto w = T_{mg}r$, where the operator $T = T_{mg}$ has the following structure:

$$T_{mg} = \left(I - M^{-*}L\right)^{\nu} PT_{mg}^{(m-1)} R \left(I - LM^{-1}\right)^{\nu} + S, \tag{3.45}$$

with S as in (3.39) and $T_{mg}^{(m-1)}$ defined according to the recursion below,

$$T_{mg}^{(l)} = \left(I_l - M_l^{-*}L_l\right)^{\nu} P_l T_{mg}^{(l-1)} R_{l-1} \left(I_l - L_l M_l^{-1}\right)^{\nu} + S_l, \ l = 1, \ldots, m-1,$$
$$T_{mg}^{(0)} = \left|L_0 - c^2 I_0\right|^{-1}, \tag{3.46}$$

where $S_l = L_l^{-1} - \left(I_l - M_l^{-*}L_l\right)^{\nu} L_l^{-1} \left(I_l - L_l M_l^{-1}\right)^{\nu}$.

Let us note that in (3.45) we skip the subscript in the notation for the quantities associated with the finest level $l = m$. The structure of the multilevel preconditioner $T = T_{mg}$ in (3.45) is similar to that of the two-grid preconditioner $T = T_{tg}$ in (3.39), with $\left|L_H - c^2 I_H\right|^{-1}$ replaced by the recursively defined operator $T_{mg}^{(m-1)}$ in (3.46). If the assumptions on the fine-grid operators M, M^*, R and P, sufficient to ensure that the two-grid preconditioner in (3.39) is SPD, remain valid throughout the coarser levels, i.e., $P_l = \alpha R_{l-1}^*$, $\rho(I_l - M_l^{-1} L_l) < 1$ and $\rho(I_l - M_l^{-*} L_l) < 1$, $l = 1, \ldots, m-1$, then the symmetry and positive definiteness of the multigrid preconditioner $T = T_{mg}$ in (3.45) is easily extended from the same property of a two-grid operator through relations (3.46).

3.2.2.2 Numerical examples

As mentioned before, two-grid Algorithm 3.8, as well as its multilevel extension given by Algorithm 3.9, can be viewed as an attempt to solve equation (3.32) using an MG method, where the absolute value of the discrete Helmholtz operator on finer levels is replaced by its approximation, i.e., the discrete negative Laplacian. Alternatively, the described approach can be interpreted as essentially applying the V-cycle of an MG method to solve the discrete Poisson's problem (i.e., approximating an inverse of the Laplacian), however, with the modified coarse grid solve.

In fact, the use of the inverse of the (shifted) Laplacian as a preconditioner for the Helmholtz equation (with possibly complex c^2), initially introduced in Turkel et al. [7], is well known and remains an object of active research, e.g., [50, 27, 75]. In our numerical tests below we consider the inverted Laplacian preconditioner as a benchmark to assess the quality of the MG absolute value preconditioner delivered by Algorithm 3.9.

Figure 3.1 illustrates several runs of PMINRES with MG absolute value preconditioners applied to solve model problem (3.34), which are compared to the corresponding runs of MINRES preconditioned with an exactly inverted (using matlab "backslash" operator) negative Laplacian. The shifts (wavenumbers) c^2 are chosen to maintain a relatively small number of negative eigenvalues of the Helmholtz operator discretized on the grid of the mesh size $h = 2^{-7}$, i.e., $c^2 = 100, 200, 300$ and 400. The right-hand side vectors b as well as initial guesses x_0 are randomly chosen (same for each shift value); the tolerance for the 2-norm of the residuals (relatively to the 2-norm of right-hand side b) is 10^{-7}.
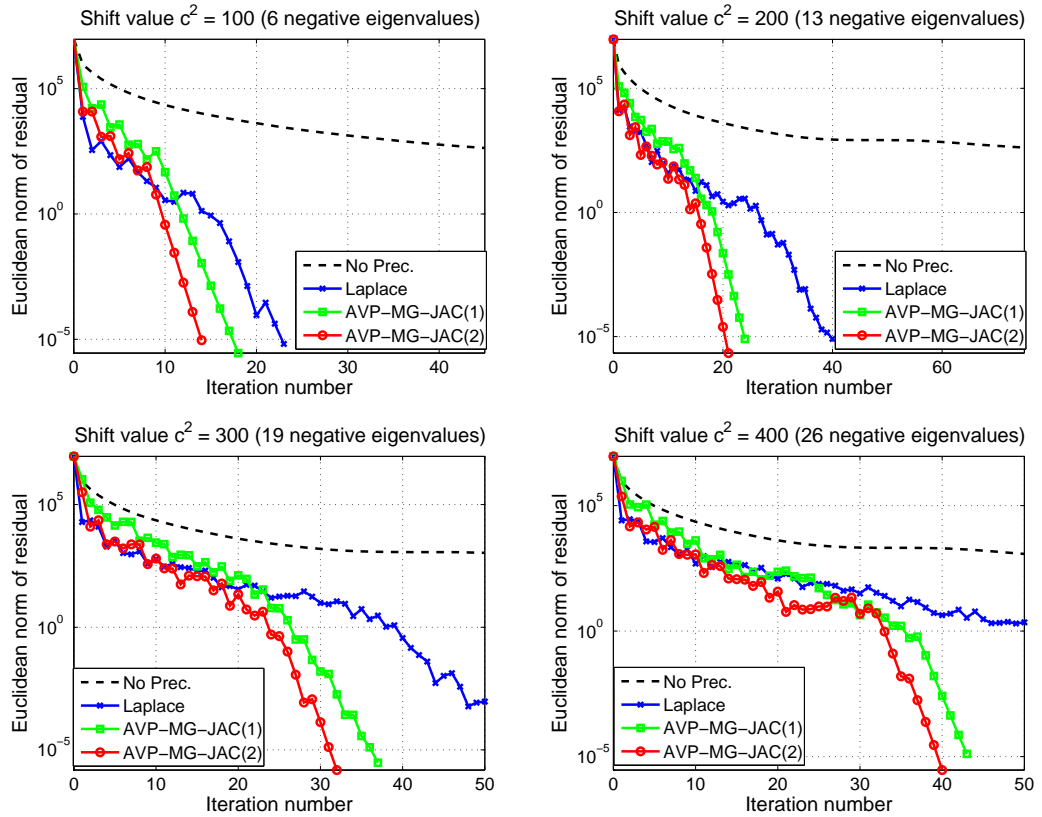
Figure 3.1: Comparison of the MG absolute value and the inverted Laplacian preconditioners for PMINRES applied to the model problem of the size $n = (2^7 - 1)^2 \approx 1.6 \times 10^4$.

The MG components for the absolute value preconditioners are defined in the following way: ω-damped Jacobi iteration as a (pre- and post-) smoother with the damping parameter $\omega = 4/5$, standard coarsening scheme (i.e., $h_{l-1} = 2h_l$) with the coarsest grid of the mesh size 2^{-4} (coarse problem size $n_0 = 225$), full weighting for the restriction, and piecewise multilinear interpolation for the prolongation, see, e.g., Trottenberg et al. [73] for more details. The number of the smoothing steps ν is chosen to be 1 and 2 (these runs are titled "AVP-MG-JAC(1)" and "AVP-MG-JAC(2)", respectively, on Figure 3.1; "Laplace"

corresponds to the case where the inverted Laplacian is used as a preconditioner). We note that the increase in the number of smoothing steps improves the quality of the MG preconditioner and results in the faster (in terms of iterations number) convergence of PMINRES. PMINRES with the absolute value preconditioners is also observed to be more robust with respect to the increase of the shift value compared to the case with the inverted Laplacian.

	$h = 2^{-7}$	$h = 2^{-8}$	$h = 2^{-9}$	$h = 2^{-10}$
$c^2 = 100$	15	14	14	14
$c^2 = 200$	21	21	21	21
$c^2 = 300$	31	32	32	30
$c^2 = 400$	40	39	40	40

Table 3.1: Mesh-independent convergence of PMINRES with the MG absolute value preconditioner

Table 3.1 shows the mesh-independence of the convergence of PMINRES with the MG absolute value preconditioner (one pre- and post-smoothing step) given by Algorithm 3.9. The rows of the table correspond to the shift values c^2 and the columns to the mesh size h. The cell in the intersection contains the number of steps performed to achieve the decrease by the factor 10^{-8} in the error norm. The mesh size of the coarse grid was kept the same throughout all runs, i.e., $h_0 = 2^{-4}$ ($n_0 = 225$).

It can be observed from Table 3.1 that the quality of the MG absolute value preconditioner deteriorates with the increase of the shift value. Figure 3.2, which
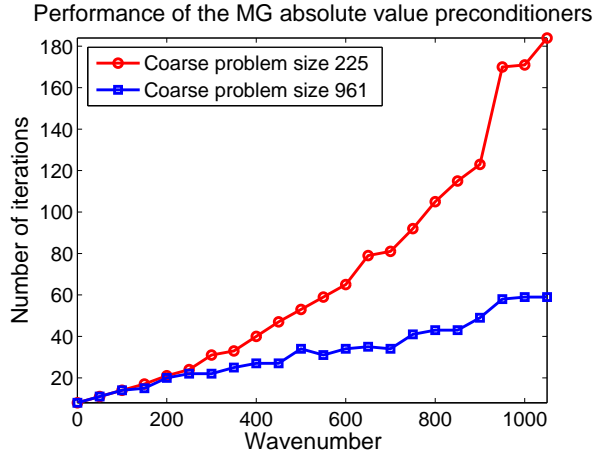
Figure 3.2: Performance of the MG absolute value preconditioners for the model problem with different shift values. The problem size $n = (2^7 - 1)^2 \approx 1.6 \times 10^4$. The number of negative eigenvalues varies from 0 to 75.

shows the number of PMINRES iterations performed to decrease the norm of the initial error by 10^{-8} for a given value of c^2, reflects the speed of this deterioration. The number of pre- and post-smoothing steps is set to one. We note that for higher wavenumbers it may be desirable to have a finer grid on the coarsest level in Algorithm 3.9.

Figure 3.3 compares locally optimal preconditioned methods (3.17)–(3.18), denoted by "AVP-MG-LO1", and (3.21), (3.24), denoted by "AVP-MG-LO2", with (globally optimal) preconditioned ("AVP-MG-MINRES") and unpreconditioned MINRES. The MG absolute value preconditioner, defined according to Algorithm 3.9, is set up as in the previous tests, with one pre- and post-smoothing step. We count the multiplication of a vector by the *preconditioned* matrix TA as one matrix-vector product. In the unpreconditioned case $T = I$. We also assume that methods (3.17)–(3.18) and (3.21), (3.24) are implemented

to perform twice as many matrix-vector multiplications per step as the PMIN-RES algorithm.
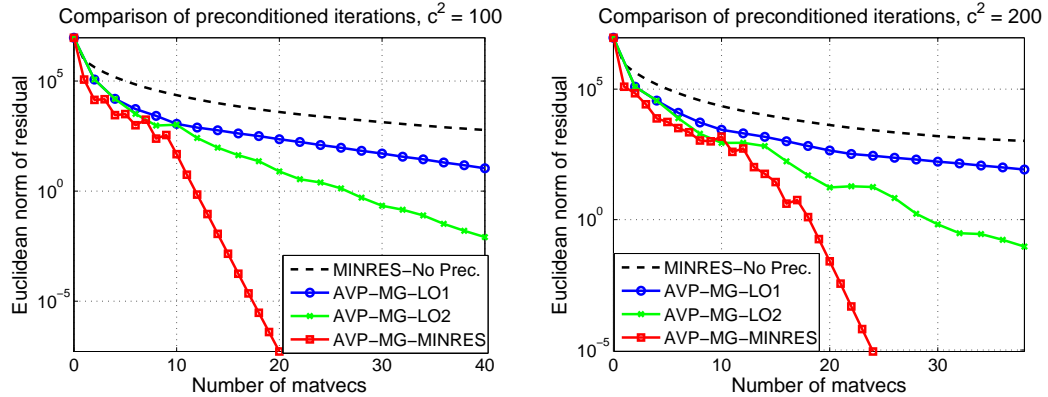

Figure 3.3: Comparison of PMINRES with locally optimal methods (3.17), (3.19) and (3.21), (3.24), all with the MG absolute value preconditioners, applied to the model problem of the size $n = (2^7 - 1)^2 \approx 1.6 \times 10^4$.

As expected, the preconditioned globally optimal method exhibits better convergence rate than methods (3.17), (3.19) and (3.21), (3.24). Method (3.21), (3.24) is noticeably faster than (3.17), (3.19), which demonstrates that the introduction of the vector $p^{(i)}$ in (3.21) indeed improves the convergence. At a number of initial steps iteration (3.21), (3.24) is comparable with PMINRES, however, the latter significantly accelerates at a certain step (possibly with the occurrence of the superlinear convergence), while the former continues to converge essentially at the same rate.

3.3 Conclusions

In this chapter we have introduced a new preconditioning strategy for symmetric indefinite linear systems, which is based on the idea of approximating the inverse of the absolute value of the coefficient matrix. We call SPD preconditioners constructed according to this principle the *absolute value preconditioners*.

We have been able to show that, for the model problem of a linear system with a two-dimensional shifted discrete negative Laplace operator as a coefficient matrix, the construction of an absolute value preconditioner can be efficiently performed using the (geometric) MG techniques. The symmetry and positive definiteness of the suggested preconditioners allow to use them within the optimal short-term recurrent Krylov subspace methods for symmetric indefinite linear systems, e.g., PMINRES. In the next chapter, we show that the absolute value preconditioners can also be used for computing the smallest magnitude eigenvalues and the corresponding eigenvectors of symmetric operators.

The future direction of the related research, as we envision it at the moment, includes the extension of known preconditioning techniques, e.g., the domain decomposition, algebraic multigrid, etc., for constructing absolute value preconditioners. Of our particular interest is the construction of *algebraic* absolute value preconditioners, as opposed, e.g., to the *geometric* MG used to justify the concept in the current chapter. The multilevel methods seem to us quite promising for constructing the efficient preconditioners, since they allow to perform all the intense computations, e.g., the inversion of a matrix absolute value using the full eigendecomposition, on a coarse space of a relatively low dimension. It is also of our interest to relate the multilevel framework to relevant factorizations that can be used for preconditioning symmetric indefinite systems, e.g., Bunch-Parlett factorization, see [29, 15], performed on a coarse space.

The significant part of this chapter has been devoted to the *locally optimal* preconditioned methods for solving symmetric indefinite linear systems. Unlike the preconditioned minimal residual method, e.g., the PMINRES algorithm,

these methods lack the *global* optimality and, hence, demonstrate slower convergence. As will be seen in the next two chapters, the study of the *locally optimal* schemes is of crucial importance for extending the ideas underlying the linear solvers to the eigenvalue and singular value computations. We also note that the understanding of the behavior of the locally optimal methods is important for the "completeness" of theory of the residual-minimizing methods for symmetric indefinite linear systems. In certain frameworks, e.g., if the preconditioner is variable, these schemes can become the methods of choice. The study of the convergence behavior of the locally optimal iterations with variable preconditioning represents one of the directions of the future research.

4. Preconditioned computations of interior eigenpairs of symmetric operators

In this chapter, we consider the *generalized symmetric eigenvalue problem* (eigenproblem)

$$Av = \lambda Bv, \quad A = A^* \in \mathbb{R}^{n \times n}, \quad B = B^* > 0 \in \mathbb{R}^{n \times n}, \tag{4.1}$$

where the targeted eigenpair corresponds to the smallest, *in the absolute value*, eigenvalue of the matrix pencil $A - \lambda B$. It is well known, e.g., [56], that problem (4.1) has all real eigenvalues λ_i, while the corresponding eigenvectors v_i, such that $Av_i - \lambda_i Bv_i = 0$, can be chosen B-orthogonal, i.e., $(v_i, v_j)_B = (v_i, Bv_j) = 0$, $i \neq j$. If $B = I$, then the generalized problem (4.1) reduces to the *standard* symmetric eigenproblem.

Problems of form (4.1) appear in a variety of applications, e.g., analysis of a system's vibration modes, buckling, electronic structure calculations of materials, graph partitioning, etc. The resulting operators A and B are often extremely large, possibly sparse and ill-conditioned. It is usually required to find a small fraction of eigenpairs, which, typically, correspond to neighboring eigenvalues of the pencil $A - \lambda B$.

An important class of symmetric eigenproblems (4.1) seeks to find several extreme, i.e., *algebraically* largest or smallest, eigenvalues and the corresponding eigenvectors (*extreme eigenpairs*). If the problem size is large, there is a number of well-established methods which can be employed to approximate the

extreme eigenpairs: the Lanczos method and its variations [56], the Jacobi-Davidson method (JD) [64], the family of preconditioned conjugate gradient (PCG) iterations, surveyed, e.g., in [54], etc. Though different in their formulations, many of the methods, in fact, follow the same framework, i.e., they perform the *Rayleigh-Ritz procedure*, see, e.g., [56], on certain low-dimensional subspaces, further called the *trial subspaces*. The choice of the trial subspaces essentially constitutes the main difference between such methods, also called *projection* methods. For example, the Lanczos method performs the Rayleigh-Ritz procedure on the Krylov subspaces, the JD relies on the subspaces obtained by solving correction equations, locally optimal (block) PCG methods [48] use spans of the current eigenvector approximations, the preconditioned residuals and the "conjugate" directions. For the comprehensive review of the relevant algorithms we refer the reader to [4].

Another important class of eigenproblems (4.1) aims at finding several eigenpairs corresponding to the eigenvalues in the interior of the spectrum of the pencil $A - \lambda B$ (*interior eigenpairs*). In particular, the important case is to find a number of eigenpairs corresponding to the eigenvalues with the smallest absolute values of a symmetric indefinite matrix. Large problems of this type frequently appear in applications, e.g., in the electronic structure calculations, see [68, 60], where a number of eigenpairs of a Hamiltonian matrix around a given energy level need to be found. The standard approaches for finding the interior eigenpairs are typically based on the shift-and-invert (SI), e.g., [4], or on the folded spectrum (FS), e.g., [71] and the references therein, transformations, and the subsequent application of one of the above mentioned methods, e.g.,

PCG, for finding extreme eigenpairs of the transformed problem. Both of the approaches, however, have potential disadvantages. To apply SI, at each step of a method, one needs to solve a large linear system involving the shifted matrix A. The FS-based methods worsen the conditioning of the problem, possibly increase the clustering in the targeted (transformed) eigenvalues, and are not easily applicable to generalized eigenproblems, i.e., $B \neq I$.

In this chapter, we introduce a method, that we refer to as the *Preconditioned Locally Minimal Residual* method (PLMR), which allows us to compute an eigenpair, corresponding to the smallest, in the absolute value, eigenvalue of problem (4.1). The described approach does not require any preliminary transformation of the eigenproblem and is applied directly to the pencil $A - \lambda B$. The PLMR method uses an SPD (absolute value) preconditioner to improve its convergence rate and robustness, and is based on the so-called *refined procedure* [44], performed in the *preconditioner-based inner product*, to extract eigenvector approximations from *four-dimensional* trial subspaces. Although the current work is concerned with finding only one eigenpair, the computation of several eigenpairs can be done similarly, either by using the method on properly deflated subspaces, or by generalizing the presented ideas to the subspace iteration.

The present chapter is organized as following. In Section 4.1, we discuss a concept of an idealized short-term recurrent preconditioned method (eigensolver) for finding an interior eigenpair. We establish a connection between solution of symmetric indefinite systems and eigenproblems, which allows us to extend the results of the previous chapter, including the idea of the absolute value preconditioning, to the case of the eigenvalue computations. In Section 4.2, we describe

the PLMR method for computing the smallest magnitude eigenvalue and the corresponding eigenvector. The numerical results, on the example of a model problem, involving a shifted Laplace operator, are presented in Section 4.3.

4.1 Idealized preconditioned methods for finding an interior eigenpair

Let us assume that the smallest, in the absolute value, eigenvalue $\lambda = \lambda_q$ is located in the interior of the spectrum of the pencil $A - \lambda B$ in (4.1), and is *a priori known*. Under the last, idealized, assumption, instead of eigenproblem (4.1) we can consider the problem of finding a null space vector:

$$(A - \lambda_q B)x = 0. \tag{4.2}$$

The link between methods for solving linear systems and eigenvalue problems has been emphasized, e.g., in Knyazev [47], or [48], where it is shown that the choice of a proper linear solver (null space finder) for (4.2) can lead to efficient methods (eigensolvers) for finding eigenpairs of the pencil $A - \lambda B$ in problem (4.1). We follow this approach here.

In order to skip unnecessary complications, we assume that all eigenvalues of $A - \lambda B$ are distinct. The solution of the singular homogeneous symmetric system (4.2) determines a vector $x = v_q$, which is the eigenvector corresponding to the eigenvalue λ_q. We also assume that the vector v_q is normalized to have the unit B-norm, and hence is unique up to a sign. We further consider the (preconditioned iterative) methods for solving linear system (4.2), and regard them as the *idealized* methods for finding the eigenpair (λ_q, v_q)—or, since λ_q is trivially found, the eigenvector v_q— of eigenproblem (4.1).

Since the coefficient matrix $A - \lambda_q B$ of linear system (4.1) is singular symmetric indefinite, we would like to construct iterative schemes, which are suitable for symmetric problems, converge to a nonzero solution, and allow us using a preconditioner $T \in \mathbb{R}^{n \times n}$ to accelerate their convergence. In Sections 3.1 of Chapter 3, we described a hierarchy of methods designed specifically to solve symmetric indefinite, though nonsingular, linear systems. It is easy to show, however, that the techniques of Chapter 3 can also be used for (consistent) singular systems. In particular, if applied to the singular homogeneous system (4.2), with an SPD preconditioner T, the methods deliver the (approximate) eigenvector v_q, which lets us consider them as *idealized* methods for finding the eigenpair (λ_q, v_q). We further restrict our attention only to the residual-minimizing methods, i.e., (3.17)–(3.18), (3.21) satisfying (3.24), and preconditioned minimal residual method (3.25)–(3.26) with the inner product defined by $S = T$, applied to system (4.2).

Proposition 4.1 *Let λ_q be an eigenvalue of the matrix pencil $A - \lambda B$ of eigenproblem (4.1). Then, given an SPD preconditioner T, methods (3.17)–(3.18), (3.21) satisfying (3.24) and (3.25)–(3.26) with the inner product generated by $S = T$, applied to the singular homogeneous system (4.2), converge to a nontrivial solution, provided that the initial guess $x^{(0)}$ has a nonzero component from the null space of $A - \lambda_q B$ in the expansion using the basis of the eigenvectors of the pencil $A - \lambda B$.*

Proof: We prove the proposition only for the case of the preconditioned minimal residual method (3.25)–(3.26) with $S = T$. The convergence of methods

(3.17)–(3.18) and (3.21), (3.24) can be shown by analogy.

Let $x^{(0)} = x_N^{(0)} + x_R^{(0)}$ be the T^{-1}-orthogonal decomposition of the initial guess vector $x^{(0)}$, such that $x_N^{(0)} \in \mathcal{N}\{T(A - \lambda_q B)\}$ and $x_R^{(0)} \in \mathcal{R}\{T(A - \lambda_q B)\}$, where $\mathcal{N}\{T(A - \lambda_q B)\}$ and $\mathcal{R}\{T(A - \lambda_q B)\}$ are the null space and the range of the operator $T(A - \lambda_q B)$, respectively. Since $\mathcal{N}\{T(A - \lambda_q B)\} = \mathcal{N}\{A - \lambda_q B\}$, we have $x_N^{(0)} \in \mathcal{N}\{A - \lambda_q B\}$ and, by assumption, $x_N^{(0)} \neq 0$.

From relation (3.25) in the definition of the preconditioned minimal residual method, we observe that, at any iteration i, the approximation $x^{(i)}$ to the solution of system (4.2) is of the form

$$x^{(i)} = x_N^{(0)} + x_R^{(i)}, \qquad x_R^{(i)} \in x_R^{(0)} + \mathcal{K}_i\left(T(A - \lambda_q B), Tr_R^{(0)}\right), \qquad (4.3)$$

where $r_R^{(0)} = (A - \lambda_q B)x_R^{(0)} = (A - \lambda_q B)(x_N^{(0)} + x_R^{(0)}) = (A - \lambda_q B)x^{(0)} = r^{(0)}$, and $x_R^{(i)} \in \mathcal{R}\{T(A - \lambda_q B)\} = \mathcal{N}\{A - \lambda_q B\}^{\perp_{T^{-1}}}$. In this case, minimization (3.26) in the definition of the preconditioned minimal residual method with $S = T$ gives

$$\|r_R^{(i)}\|_T = \min_{u \in (A - \lambda_q B)\mathcal{K}_i\left(T(A - \lambda_q B), Tr_R^{(0)}\right)} \|r_R^{(0)} - u\|_T, \qquad (4.4)$$

where $r_R^{(i)} = (A - \lambda_q B)x_R^{(i)} = (A - \lambda_q B)(x_N^{(0)} + x_R^{(i)}) = (A - \lambda_q B)x^{(i)} = r^{(i)}$, by (4.3). Expression (4.3) shows that the preconditioned minimal residual method, applied to solve system (4.2), preserves the null space component $x_N^{(0)}$ for all the iterates $x^{(i)}$, while, by (4.3)–(4.4), the range components $x_R^{(i)}$ converge to zero at the rate delivered by the method applied to find the (unique) zero solution of the (nonsingular) restricted system

$$T(A - \lambda_q B)|_{\mathcal{R}\{T(A - \lambda_q B)\}} x = 0, \qquad (4.5)$$

with the initial guess $x_R^{(0)} \in \mathcal{R}\{T(A - \lambda_q B)\}$ and the preconditioner T. Thus, the approximations $x^{(i)}$ to the solution of (4.2) converge to $x_N^{(0)} \in \mathcal{N}\{A - \lambda_q B\}$.

∎

The proof of Proposition 4.1 shows that the preconditioned methods, described in the previous chapter, applied to system (4.2), deliver a nonzero solution by annihilating the component in the range of $T(A - \lambda_q B)$ of the initial guess, at the rate delivered by the selected method, applied to the restricted system (4.5). This observation, along with bound (3.9), (3.20), suggests the following convergence estimate for method (3.17)–(3.18), applied to solve the singular homogeneous system (4.2):

$$\frac{\|Ax^{(i+1)} - \lambda_q Bx^{(i+1)}\|_T}{\|Ax^{(i)} - \lambda_q Bx^{(i)}\|_T} \leq \frac{\tilde{\kappa} - 1}{\tilde{\kappa} + 1} < 1, \tag{4.6}$$

where, assuming that $\mu_1 < \mu_{q-1} < \mu_q = 0 < \mu_{q+1} < \mu_n$ are the nonzero eigenvalues of the preconditioned operator $T(A - \lambda_q B)$, the expression for $\tilde{\kappa}$ is given by

$$\tilde{\kappa} = \begin{cases} \left(\dfrac{\mu_n}{\mu_{q+1}}\right)\left(1 + \dfrac{\mu_n - \mu_{q+1}}{|\mu_{q-1}|}\right), & \text{if } |\mu_1| - |\mu_{q-1}| \leq \mu_n - \mu_{q+1} \\[4mm] \left(\dfrac{\mu_1}{\mu_{q-1}}\right)\left(1 + \dfrac{|\mu_1| - |\mu_{q-1}|}{\mu_{q+1}}\right), & \text{if } |\mu_1| - |\mu_{q-1}| > \mu_n - \mu_{q+1}. \end{cases} \tag{4.7}$$

Bound (4.6)–(4.7) can also be used to estimate the convergence rate of method (3.21), (3.24), applied to system (4.2), however, as has been discussed in Subsection 3.1.3 of the previous chapter, it is likely to be pessimistic, and we can expect, in practice, the reduction in the residual T-norm in (4.6) by a factor of order $\dfrac{\sqrt{\tilde{\kappa}} - 1}{\sqrt{\tilde{\kappa}} + 1}$, with $\tilde{\kappa}$ given in (4.7). If the preconditioned minimal residual

method (3.25)–(3.26) with $S = T$ is applied to solve system (4.2), by (3.27), the following estimate holds:

$$\frac{\|Ax^{(i)} - \lambda_q Bx^{(i)}\|_T}{\|Ax^{(0)} - \lambda_q Bx^{(0)}\|_T} \leq 2 \left(\frac{\sqrt{\tilde{\kappa}} - 1}{\sqrt{\tilde{\kappa}} + 1} \right)^j, \ \ i = 2j, \ j = 1, 2, \ldots. \quad (4.8)$$

Remark 4.2 *Proposition 4.1, along with bounds (4.6)–(4.7) and (4.7)–(4.8), is also valid for a symmetric positive semi-definite preconditioner T, such that*

$$\mathcal{R}\{T\} = \mathcal{R}\{A - \lambda_q B\}.$$

In this case, the nonzero solution of (4.2) is delivered by annihilating the component in the range of $A - \lambda_q B$ of the initial guess $x^{(0)}$. The T-norm is formally replaced by the T-seminorm.

The preconditioned minimal residual method (3.25)–(3.26) with $S = T$ (possibly symmetric positive semi-definite, by Remark 4.2), applied to solve (4.2), with the convergence rate given by (4.8), represents the *globally optimal* idealized method (in the class of the preconditioned Krylov subspace methods) for finding the eigenpair (λ_q, v_q). The method is known to admit a short-term recurrent implementation, e.g., in the form of preconditioned MINRES (PMINRES), orthodir(3) or orthomin(2) [78, 33, 59]. Typically, these implementations of the preconditioned minimal residual method require one matrix-vector multiplication and one application of a preconditioner per iteration, or, equivalently, according to (3.27), or (4.8), two matrix-vector multiplications and two applications of a preconditioner to guarantee the reduction of the residual T-norm at every other step.

As a *base* version of the *idealized* eigenvalue solver we suggest to choose a *preconditioned* method, which is *locally optimal* and *convergent* for any initial

guess, such that the convergence rate and the amount of the involved computational work mimic those of the *globally optimal* preconditioned minimal residual method (in one of its short-term recurrent formulations). For such *base* idealized method we choose (3.21), (3.24), applied to the singular homogeneous system (4.2). The scheme can be written in the form of the four-term recurrence:

$$x^{(i+1)} = x^{(i)} + \alpha^{(i)} w^{(i)} + \beta^{(i)} T \left(Aw^{(i)} - \lambda_q Bw^{(i)} \right) + \gamma^{(i)} p^{(i)},$$
$$w^{(i)} \;\; = T \left(Ax^{(i)} - \lambda_q Bx^{(i)} \right), \;\; p^{(i)} = x^{(i)} - x^{(i-1)}, \;\; p^{(0)} = 0; \;\; i = 0, 1, \ldots,$$

$$(4.9)$$

where the iteration parameters $\alpha^{(i)}$, $\beta^{(i)}$ and $\gamma^{(i)}$ are chosen to minimize the T-norm of the new residual vector over the corresponding low-dimensional subspace, as in (3.24). The approximation to the eigenvector v_q is obtained after a suitable normalization of the last iterate $x^{(i)}$. The preconditioner T is assumed to be SPD or, by Remark 4.2, symmetric positive semi-definite. In practice, however, we only consider preconditioners, which are SPD. The semi-definite case, given by Remark 4.2, is introduced mainly for theoretical purposes, e.g., for defining an optimal preconditioner below. In the next section we use the *base* method (4.9), with an SPD preconditioner T, as a starting point for deriving preconditioned methods for computing interior eigenpairs.

As has been noted above, it is reasonable to expect that idealized scheme (4.9) can attain the convergence factor of order $\dfrac{\sqrt{\tilde{\kappa}} - 1}{\sqrt{\tilde{\kappa}} + 1}$, with $\tilde{\kappa}$ given in (4.7), which, in a sense, according to (4.8), reflects the convergence behavior of the *globally optimal* method, moreover, the amount of computations, required to achieve the reduction in the residual norm, is also essentially the same for both methods (assuming that the preconditioned minimal residual method is implemented in a short-term recurrent form). If vectors $p^{(i)}$ are removed from (4.9),

then one gets the scheme corresponding to method (3.17)–(3.18), applied to solve system (4.2). The latter represents a less computationally expensive idealized method for finding the eigenpair (λ_q, v_q), which is generally expected to exhibit a slower convergence. The corresponding bound is given by (4.6).

We remark here that our original intent was to choose the *base* idealized solver to be one of the short-term recurrent implementations of the preconditioned minimal residual method, applied to system (4.2). However, the choice of the preconditioned orthomin(2) as the *base* method lacked robustness, in the sense that the algorithm admitted break-downs, or stagnations (if written in the form of the three-term recurrent relation, as opposed to the standard version based on two linked two-term recurrences, similarly to PCG, e.g., in [3, 48]), which is a known drawback of the orthomin family of methods, see, e.g., [33]. At the same time, the robust implementations, i.e., PMINRES and the preconditioned orthodir(3), failed to provide a proper insight into the structure of local subspaces, used to determine the improved approximations—the recurrent relations, underlying the algorithms, involve, e.g., the Lanczos vectors for the Krylov subspaces generated by $T(A - \lambda_q B)$ (PMINRES), or the orthogonal direction vectors (orthodir), which can neither be computed, nor approximated, once we depart from the *idealized* framework (with λ_q already known) considered in the current section. Therefore, the choice of (4.9) as a *base* idealized method for finding the eigenpair (λ_q, v_q) can be viewed as a reasonable compromise. Scheme (4.9) is no longer *globally* optimal. However, it reveals the structure of the local subspaces, used to determine the next iterate, and is convergent for any initial guess. Moreover, the convergence rate and the amount of computational work

can be expected to mimic those (up to the possible occurrence of effects apparently attributed to the superlinear convergence, see, e.g., Figure 3.3) of the globally optimal preconditioned minimal residual method in one of its robust short-term recurrent formulations.

Finally, let us discuss the choice of the preconditioner T for the *base* idealized method (4.9). The following proposition defines the *optimal* preconditioner.

Proposition 4.3 *Let $T = |A - \lambda_q B|^\dagger$, where λ_q is the eigenvalue of the matrix pencil $A - \lambda B$ in (4.1). Then method (4.9) converges to the eigenvector, corresponding to λ_q, in exactly one step, provided that the initial guess has a nontrivial component from the null space of $A - \lambda_q B$.*

Proof: The proof follows from Proposition 4.1, Remark 4.2 and Theorem 3.6. ∎

We note that Proposition 4.3 is valid if $p^{(i)} = 0$ for all i in (4.9), i.e., if the idealized solver is given by method (3.17)–(3.18), applied to system (4.2). We further use the notion of the *optimal* (symmetric positive semi-definite) preconditioner to obtain *practical* SPD preconditioners for the base idealized scheme (4.9).

The targeted eigenvalue λ_q is the smallest in the absolute value. In some applications, its magnitude can be considered relatively negligible, e.g., compared to a norm of $B^{-1}A$. In such cases, it is worth trying to replace the theoretically optimal preconditioner $T_{opt} = |A - \lambda_q B|^\dagger$ by $T = |A|^\dagger \approx T_{opt} = |A - \lambda_q B|^\dagger$. Let us remark that a similar strategy to justify the choice of a preconditioner has been successfully used for the LOBPCG method, see [48], for approximating

extreme eigenpairs of symmetric matrix pencils.

If A is nonsingular, then $T = |A|^{-1}$, otherwise one can, e.g., introduce a (relatively small) regularization parameter $\alpha \in \mathbb{R}$, and (instead of $T = |A|^{\dagger}$) set $T = |A + \alpha I|^{-1}$, which is SPD. Since the computation of the exact inverse of the matrix absolute value may be prohibitive for practical problem sizes, the actual SPD preconditioners, used in (4.9), can be constructed as some *approximations* of $|A|^{-1}$ (or, $|A + \alpha I|^{-1}$, if A is (close to) singular). Such preconditioners, along with their examples for a model problem, have been introduced in Chapter 3 of the present manuscript, and are referred to as the *absolute value precondition-ers*. In the next sections, we show that exactly the same (SPD) absolute value preconditioners that are used for solving symmetric indefinite systems can be utilized for computing interior eigenpairs, corresponding to the smallest, in the absolute value, eigenvalues of symmetric matrix pencils.

4.2 The Preconditioned Locally Minimal Residual method for computing interior eigenpairs

In this section, we describe an iterative scheme for computing an eigenpair, corresponding to the smallest, in the absolute value, eigenvalue of the pencil $A - \lambda B$ in (4.1). The proposed method is based on a *four-term recurrent re-lation*, which means that, at each step, the new eigenvector approximation is extracted from a four-dimensional trial subspace. The extraction of the ap-proximate eigenvector represents, essentially, the *refined procedure*, also called the *refined projection procedure*, originally introduced in Jia [44], however, per-formed in the inner product generated by a properly chosen SPD preconditioner T. We call the method the *Preconditioned Locally Minimal Residual method,* or

PLMR, and discuss its possible variants.

4.2.1 PLMR: The choice of trial subspaces

In the previous section, we have considered several *idealized* methods for finding an interior eigenpair, derived assuming that the targeted eigenvalue is known, as null space finders based on the preconditioned schemes for symmetric indefinite systems described in Chapter 3. As a *base* idealized method we have suggested scheme (4.9), which represents the four-term recurrent relation with the next approximation $x^{(i+1)}$ determined as an element of

$$\text{span}\left\{x^{(i)}, w^{(i)}, T(Aw^{(i)} - \lambda_q Bw^{(i)}), x^{(i-1)}\right\}, \ w^{(i)} = T(Ax^{(i)} - \lambda_q Bx^{(i)}), \ (4.10)$$

where λ_q is the known (smallest in the absolute value) eigenvalue of the pencil $A - \lambda B$, $x^{(-1)} = 0$. The sequence of approximations $x^{(i)}$ in (4.9) converges (under mild assumptions on the initial guess $x^{(0)}$, see Proposition 3.1) to a nonzero vector from the null space of $A - \lambda_q B$. After being normalized to have a unit B-norm, the (approximate) null space vector delivers the (approximate) eigenvector v_q, corresponding to the eigenvalue λ_q.

Our goal is to obtain a preconditioned method for finding an eigenpair, corresponding to the smallest, in the absolute value, eigenvalue of the pencil $A - \lambda B$ in (4.1), which is similar, in terms of the convergence behavior and the computational cost, to the *base* idealized method (4.9). Thus, it is desirable that, at each step, the new approximation $v^{(i+1)}$ to the eigenvector v_q is extracted from the recurrently defined low-dimensional subspace of form (4.10), with $x^{(i)} = v^{(i)}$ and $x^{(i-1)} = v^{(i-1)}$ being the current and the previous eigenvector approximations, respectively. In practice, however, since the exact value of λ_q is unknown, the computation of subspaces (4.10) is, generally, impossible. Instead, at step $(i+1)$,

we suggest to replace the targeted eigenvalue λ_q in (4.10) by its (asymptotically quadratic) approximation, i.e., the Rayleigh quotient

$$\lambda^{(i)} = \frac{(v^{(i)}, Av^{(i)})}{(v^{(i)}, Bv^{(i)})}, \tag{4.11}$$

and, given $v^{(i)}$, $v^{(i-1)}$ and an SPD preconditioner T, extract the new eigenvector approximation $v^{(i+1)}$ from

$$\text{span}\left\{v^{(i)}, w^{(i)}, T(Aw^{(i)} - \lambda^{(i)}Bw^{(i)}), v^{(i-1)}\right\}, \quad w^{(i)} = T(Av^{(i)} - \lambda^{(i)}Bv^{(i)}), \tag{4.12}$$

where $v^{(-1)} = 0$. This can be translated, e.g., into the recurrence of the following form:

$$v^{(i+1)} = \alpha^{(i)}v^{(i)} + \beta^{(i)}w^{(i)} + \gamma^{(i)}T\left(Aw^{(i)} - \lambda^{(i)}Bw^{(i)}\right) + \delta^{(i)}p^{(i)},$$
$$w^{(i)} = T\left(Av^{(i)} - \lambda^{(i)}Bv^{(i)}\right), \ p^{(i)} = v^{(i)} - \alpha^{(i-1)}v^{(i-1)}, \ p^{(0)} = 0, \tag{4.13}$$
$$i = 0, 1, \dots;$$

where $\alpha^{(i)}$, $\beta^{(i)}$, $\gamma^{(i)}$ and $\delta^{(i)}$ are some iteration parameters, $v^{(0)}$ is the initial guess. By (4.13), at step $(i + 1)$, the eigenvector approximation $v^{(i+1)}$ is determined as an element of the subspace

$$\mathcal{V}^{(i+1)} = \text{span}\left\{v^{(i)}, w^{(i)}, T(Aw^{(i)} - \lambda^{(i)}Bw^{(i)}), v^{(i)} - \alpha^{(i-1)}v^{(i-1)}\right\}, \tag{4.14}$$

which is the same (in the exact arithmetic) as (4.12), $v^{(-1)} = 0$. The choice of the vector $p^{(i)}$ in (4.13)–(4.14) as a weighted difference of the two consecutive eigenvector approximations has been motivated by implementational considerations, mainly, to obtain a stable formula, see [48], for computation of trial subspaces, by calculating $p^{(i)}$ implicitly, i.e.,

$$p^{(i+1)} = \beta^{(i)}w^{(i)} + \gamma^{(i)}T\left(Aw^{(i)} - \lambda^{(i)}Bw^{(i)}\right) + \delta^{(i)}p^{(i)}. \tag{4.15}$$

We further discuss the selection of the iteration parameters in (4.13).

4.2.2 PLMR: The choice of iteration parameters

Given a k-dimensional subspace $\mathcal{V} \subseteq \mathbb{R}^n$, we want to extract an approximation $v \in \mathcal{V}$ to the eigenvector v_q, corresponding to the smallest, in the absolute value, eigenvalue λ_q of (4.1).

Let us assume that $\tilde{\lambda} \in \mathbb{R}$ is some approximation to the targeted eigenvalue λ_q, i.e., $\tilde{\lambda} \approx \lambda_q$. In this case, one can attempt to extract the corresponding eigenvector approximation $v \in \mathcal{V}$ by satisfying the following optimality condition:

$$v = \operatorname*{argmin}_{z \in \mathcal{V}, \|z\|_B = 1} \|Az - \tilde{\lambda} Bz\|, \tag{4.16}$$

where $\|z\|_B^2 = (z, Bz)$, and $\|z\| = \|z\|_I$ is the 2-norm. The minimization principle in (4.16), in fact, defines the *refined procedure*, also called the *refined projection procedure*, which is straightforwardly extended to the case of the generalized eigenproblem (the original condition in [44] was formulated for $B = I$). The minimizer v in (4.16) is called the *refined approximate eigenvector*.

Given an SPD preconditioner T, we modify condition (4.16) to perform the minimization in the preconditioner-based T-norm, rather than in the standard 2-norm, i.e.,

$$v = \operatorname*{argmin}_{z \in \mathcal{V}, \|z\|_B = 1} \|Az - \tilde{\lambda} Bz\|_T, \tag{4.17}$$

where $\|z\|_T^2 = (z, Tz)$. Assuming that the matrix $V \in \mathbb{R}^{n \times k}$ is such that $\operatorname{col}(V) = \mathcal{V}$, where $\operatorname{col}(V)$ denotes the column space of V, and, hence, any $z \in \mathcal{V}$ is of the

form $z = Vy$, for some $y \in \mathbb{R}^k$, we get

$$\|Az - \tilde{\lambda}Bz\|_T^2 = (Az - \tilde{\lambda}Bz, T(Az - \tilde{\lambda}Bz)) = (z, (A - \tilde{\lambda}B)T(A - \tilde{\lambda}B)z)$$

$$= (Vy, (A - \tilde{\lambda}B)T(A - \tilde{\lambda}B)Vy)$$

$$= (y, V^*(A - \tilde{\lambda}B)T(A - \tilde{\lambda}B)Vy). \tag{4.18}$$

Thus, (4.17) can be replaced by the problem of finding the minimizer $y_{min} \in \mathbb{R}^k$, such that

$$y_{min} = \operatorname*{argmin}_{y \in \mathbb{R}^k, \|Vy\|_B = 1} (y, V^*(A - \tilde{\lambda}B)T(A - \tilde{\lambda}B)Vy)$$

$$= \operatorname*{argmin}_{y \in \mathbb{R}^k} \frac{(y, V^*(A - \tilde{\lambda}B)T(A - \tilde{\lambda}B)Vy)}{(Vy, BVy)}$$

$$= \operatorname*{argmin}_{y \in \mathbb{R}^k} \frac{(y, V^*(A - \tilde{\lambda}B)T(A - \tilde{\lambda}B)Vy)}{(y, V^*BVy)},$$

which is equivalent to the problem of finding the eigenvector y_{min}, corresponding to the smallest eigenvalue θ_{min}^2, of the k-by-k generalized symmetric eigenvalue problem

$$(V^*(A - \tilde{\lambda}B)T(A - \tilde{\lambda}B)V)y = \theta^2 (V^*BV)y. \tag{4.19}$$

The square root of the smallest eigenvalue in (4.19), i.e., θ_{min}, gives the minimal value of norm (4.17), while the eigenvector y_{min} determines the corresponding minimizer

$$v = Vy_{min}, \quad \|v\|_B = 1, \tag{4.20}$$

which we set as the new eigenvector approximation. The value of θ_{min} is typically discarded.

As has been previously discussed, at a general step $(i + 1)$, according to (4.14), we choose the trial subspaces as spans of four vectors, i.e., $\mathcal{V} = \mathcal{V}^{(i+1)}$,

and $k = 4$. The new eigenvector approximation $v = v^{(i+1)}$, satisfying (4.17) for a (presumably) given $\tilde{\lambda} = \tilde{\lambda}^{(i)}$, is determined by (4.20) after finding the eigenvector y_{min}, corresponding to the smallest eigenvalue of the 4-by-4 generalized eigenvalue problem (4.19), where the matrix V has vectors from (4.14) as columns. The iteration parameters $\alpha^{(i)}$, $\beta^{(i)}$, $\gamma^{(i)}$ and $\delta^{(i)}$ in (4.13) are then given as the components of the vector y_{min}. We note that, at the initial step $(i = 0)$, the trial subspace (4.14) is spanned by three vectors, i.e., $k = 3$, and, hence, the described extraction of the eigenvector approximation reduces to the solution of the 3-by-3 eigenvalue problem (4.19). Let us remark that at each step minimization principle (4.17) with $\mathcal{V} = \mathcal{V}^{(i+1)}$ and $\tilde{\lambda} = \tilde{\lambda}^{(i)} \approx \lambda_q$ mimics optimality condition (3.24) underlying the base idealized method. The remaining question is how to find the eigenvalue approximations $\tilde{\lambda} = \tilde{\lambda}^{(i)}$ in (4.17)?

If the current approximation $v^{(i)}$ is already close to the desired eigenvector v_q, then one can choose to set $\tilde{\lambda} = \tilde{\lambda}^{(i)}$ in (4.17) into the corresponding value of the Rayleigh quotient (4.11), i.e.,

$$\tilde{\lambda}^{(i)} = \lambda^{(i)}.$$

In general, however, the approximation $v^{(i)}$ can be far from the targeted eigenvector. In this case, assuming that the SPD operator B in (4.1) can be efficiently inverted, prior to fulfilling (4.17), we suggest to find an estimate $\tilde{\lambda} = \tilde{\lambda}^{(i)}$ by performing the Rayleigh-Ritz procedure for the pencil

$$AB^{-1}Av = \lambda^2 Bv, \tag{4.21}$$

on the trial subspace $\mathcal{V} = \mathcal{V}^{(i+1)}$, defined in (4.14), and available at the step $(i + 1)$. Then, if $\tilde{v} = \tilde{v}^{(i)}$ is the *Ritz vector*, corresponding to the smallest Ritz

value of (4.21) on $\mathcal{V}$, we set $\tilde{\lambda} = \tilde{\lambda}^{(i)}$ to the value of the Rayleigh quotient for problem (4.1), evaluated at $\tilde{v}^{(i)}$, i.e.,

$$\tilde{\lambda}^{(i)} = \frac{\left(\tilde{v}^{(i)}, A\tilde{v}^{(i)}\right)}{\left(\tilde{v}^{(i)}, B\tilde{v}^{(i)}\right)}, \tag{4.22}$$

and discard the Ritz value. We note that if $B = I$, i.e., (4.1) is the standard eigenproblem, the above described approach for estimating $\tilde{\lambda}$ is the Rayleigh-Ritz procedure for A^2 on the given subspace.

We now summarize the whole approach in the following algorithm.

Algorithm 4.4 (The PLMR method)

Input: starting vector $v^{(0)}$, functions to compute Av, Bv, $B^{-1}v$ and Tv

Output: approximation to the eigenpair (λ_q, v_q), such that $|\lambda_q| = \min\limits_{j} |\lambda_j|$

1. *Start: Select $v^{(0)}$ and set $p^{(0)} = 0$*

2. *Iterate: For $i = 0, 1, \ldots$, Until Convergence Do:*

3. $\lambda^{(i)} := \left(v^{(i)}, Av^{(i)}\right) / \left(v^{(i)}, Bv^{(i)}\right)$, $r := Av^{(i)} - \lambda^{(i)} Bv^{(i)}$

4. $w^{(i)} := Tr$, $s^{(i)} := T(Aw^{(i)} - \lambda^{(i)} Bw^{(i)})$

5.   *Use the Rayleigh-Ritz method for (4.21) on the trial subspace*
   span $\left\{v^{(i)}, w^{(i)}, s^{(i)}, p^{(i)}\right\}$

6. $\tilde{\lambda} := (\tilde{v}, A\tilde{v}) / (\tilde{v}, B\tilde{v})$

   ($\tilde{v}$ is the Ritz vector corresponding to the smallest Ritz value in 5.)

7.   *If $i > 0$, then $V := \left[v^{(i)}; w^{(i)}; s^{(i)}; p^{(i)}\right] \in \mathbb{R}^{n \times 4}$;*
   else $V := \left[v^{(i)}; w^{(i)}; s^{(i)}\right] \in \mathbb{R}^{n \times 3}$

8. *Solve (4.19) and set* $(\alpha^{(i)}\ \beta^{(i)}\ \gamma^{(i)}\ \delta^{(i)}) := y^*_{min}$

9. $v^{(i+1)} := \alpha^{(i)}v^{(i)} + \beta^{(i)}w^{(i)} + \gamma^{(i)}s^{(i)} + \delta^{(i)}p^{(i)}$

10. $p^{(i+1)} := \beta^{(i)}w^{(i)} + \gamma^{(i)}s^{(i)} + \delta^{(i)}p^{(i)}$

11. *EndDo*

As has been previously suggested, if the pair $(\lambda^{(i)}, v^{(i)})$ is near the exact solution (λ_q, v_q), then one can skip step 5 of Algorithm 4.4, and set $\tilde{\lambda}$ to the current value of the Rayleigh quotient $\lambda^{(i)}$ at step 6. We also remark here that, as the approximations $v^{(i)}$ get closer to the eigenvector v_q, it may become necessary to perform B-orthogonalization on the trial subspaces to achieve a better numerical stability and a higher attainable accuracy of the method. We demonstrate this in our numerical tests of the next section.

We, finally, note that, instead of satisfying the T-norm optimality condition (4.17), a possible approach to extract an eigenvector approximation in Algorithm 4.4 could be to use the *Rayleigh-Ritz* procedure for (4.21) on the trial subspace (4.14). The resulting algorithm, however, did not bring any improvements to the method, given by Algorithm 4.4, and, in many cases, led to significantly less satisfactory convergence behavior, e.g., demonstrating a lower convergence rate or stagnations. The observed robustness of Algorithm 4.4 with a suitable preconditioner, on the example of the model problem below, can, in part, be attributed to the properly chosen, convergent, *base* null space finder (4.9), discussed in Section 4.1.

4.3 Numerical examples

In this section we apply the PLMR method, given by Algorithm 4.4, to the model problem of approximating an eigenpair of the discrete negative Laplace operator L, which corresponds to the eigenvalue, closest to a given shift value c^2. We assume that the operator is discretized using the 5-point FD stencil on the unit square domain, with Dirichlet boundary conditions. This problem, in fact, corresponds to the task of finding an approximation to the, generally, interior eigenpair (λ_q, v_q) of the shifted negative Laplacian, $L - c^2 I$, which corresponds to its smallest, in the absolute value, eigenvalue. In other words, we consider the symmetric eigenvalue problem

$$\left(L - c^2 I\right) v = \lambda v, \tag{4.23}$$

where the desired eigenpair (λ_q, v_q) is such that $|\lambda_q| = \min_j |\lambda_j|$, and λ_j are the eigenvalues of $L - c^2 I$.

Since the exact eigenvalues of the (shifted) Laplacian in (4.23) can be found using explicit expressions, see, e.g., [33], for our theoretical purposes, we can fix the desired value of λ_q and, in the spirit of Section 4.1, consider the problem of finding a nonzero null space vector of operator $(L - c^2 I) - \lambda_q I$. In order to control the performance of the PLMR algorithm, we suggest to compare its convergence behavior versus PMINRES, applied to the singular homogeneous system

$$\left((L - c^2 I) - \lambda_q I\right) x = 0, \tag{4.24}$$

and versus the *base* idealized eigensolver (4.9), with A replaced by $L - c^2 I$ (i.e., versus method (3.21), (3.24), applied to (4.24)). In this framework, the globally optimal PMINRES algorithm provides the pattern of the theoretically optimal convergence (in the class of the Krylov subspace methods), while the locally optimal method (4.9), with $A = L - c^2 I$, delivers a benchmark for the actual convergence rate of the PLMR algorithm. We further refer to the version of PMINRES as "MINRESNULL" and call method (4.9) "LO-BASE". We remark that the code for "MINRESNULL", used in our numerical experiments, has been obtained by modifying the matlab funcion "minres.m" to skip the check for the zero right-hand side and deliver the residual norms computed at iterates, normalized to have a unit length.

As has been discussed in Section 4.1, as an SPD preconditioner T for the PLMR method, applied to eigenproblem (4.23), as well as for the introduced above "control" methods, applied to system (4.24), one can choose an approximation to $|L - c^2 I|^{-1}$. We recall that such (absolute value) preconditioner has already been constructed for the model linear system (3.34) with the coefficient matrix $L - c^2 I$, in Chapter 3, using the MG approach, see Algorithm 3.9. It is remarkable that the absolute value preconditioners, constructed for solving symmetric indefinite linear systems, can be used within the PLMR method for approximating the eigenpairs, corresponding to the interior eigenvalues, of the respective operators. Thus, as a preconditioner for problem (4.23), we use Algorithm 3.9, described in the previous chapter.

Figure 4.1 illustrates the convergence of the PLMR method, applied to problem (4.23) with different values of the shifts c^2, which give different smallest
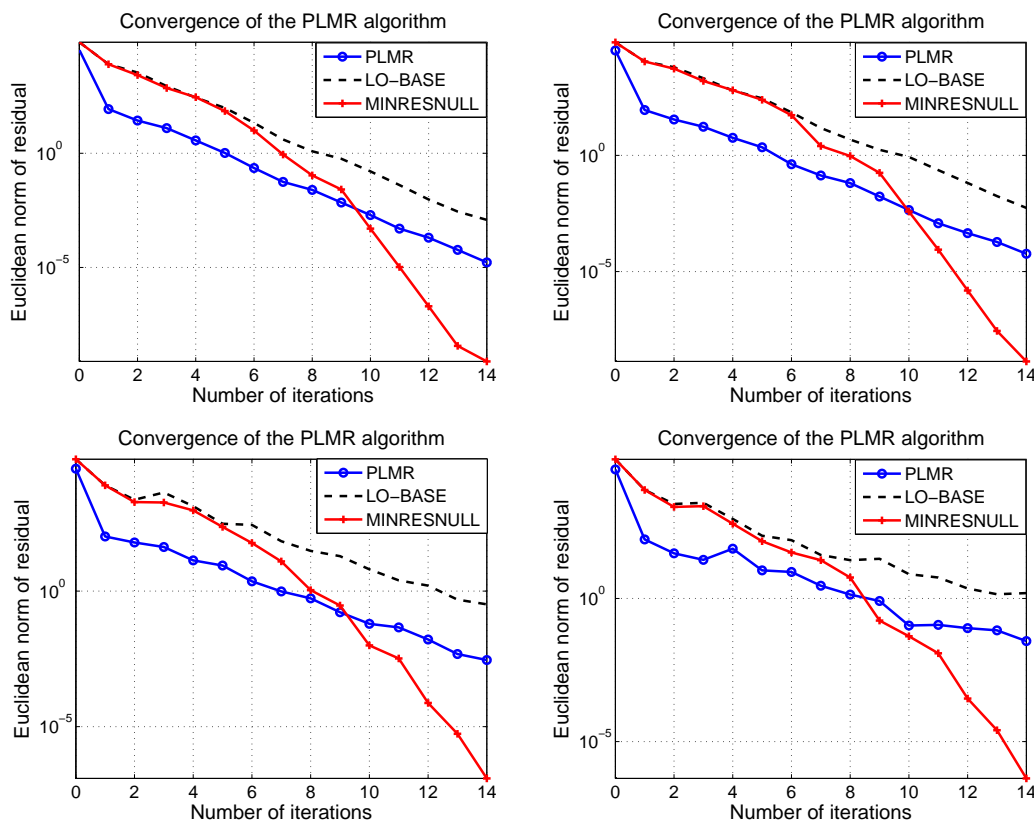
Figure 4.1: Comparison of the PLMR method with the MG absolute value pre-conditioner versus the idealized eigenvalue solvers, applied to the model eigen-problem of the size $n = (2^7 - 1)^2 \approx 1.6 \times 10^4$. The targeted eigenpairs correspond to the smallest magnitude eigenvalues of the shifted discrete negative Laplacian (from top left to bottom left, clockwise): $\lambda_{13} \approx -6.33 \times 10^{-4}$, $\lambda_{13} \approx -2.7426$, $\lambda_{15} \approx -3.4268$ and $\lambda_{17} \approx 7.19 \times 10^{-4}$, given by shift values $c^2 = 197.258, 200, 250$ and 256.299, respectively.

magnitude eigenvalues. The Laplace operator is discretized on the grid of the mesh size $h = 2^{-7}$, initial eigenvector approximations are randomly chosen. The MG components for the absolute value preconditioner are defined similarly to Subsection 3.2.2.2, with one step of the 4/5-damped Jacobi iteration as a (pre- and post-) smoother, standard coarsening scheme with the coarsest grid of the mesh size 2^{-4}, full weighting for the restriction, and piecewise multilinear interpolation for the prolongation. The norms $\|(L - c^2 I)v^{(i)} - \lambda^{(i)}v^{(i)}\|$ of the residual vectors for problem (4.23), generated at each step of the PLMR algorithm, are compared to the norms of the residuals $\dfrac{\|((L - c^2 I) - \lambda_q I)x^{(i)}\|}{\|x^{(i)}\|}$ for problem (4.24), with the corresponding value of λ_q, evaluated at the normalized iterates $\dfrac{x^{(i)}}{\|x^{(i)}\|}$, produced by "MINRESNULL" and "LO-BASE". We note that, by (3.27), or (4.8), PMINRES generally requires at least two steps to guarantee the reduction in the residual norm. Therefore, the plotted values of the "MIN-RESNULL" residual norms in Figure 4.1 are obtained by measuring the norms, produced by the PMINRES method, after every other step. In other words, the "MINRESNULL" residual norm at step i, in Figure 4.1, corresponds to the norm of the PMINRES algorithm, applied to (4.24), at step $j = 2i$, evaluated at the normalized iterate.

In Figure 4.1 we observe that the PLMR method converges, essentially, at the same rate as the idealized base eigensolver "LO-BASE". The globally optimal "MINRESNULL", at a number of its initial steps, demonstrates the similar convergence behavior, however, accelerates, possibly, with the occurrence of the superlinear convergence, which is frequently noticed for preconditioned globally optimal Krylov subspace methods, e.g., [9, 62]. In fact, it is generally

hard to expect the superlinear convergence for the PLMR algorithm, which is likely to be the price, paid for the departure from the *global* optimality.

In the next set of tests, illustrated in Figure 4.2, we examine the effects of orthogonalization on the trial subspaces, prior to performing steps 5 and 7 of Algorithm 4.4. We denote the version of the PLMR method with the orthogonalization by "PLMR-ORTH". Similarly to the numerical examples above, we seek to approximate the smallest, in the absolute value, eigenvalue, and the corresponding eigenvector, of the shifted discrete negative Laplacian in (4.23), with different shift values. Algorithm 3.9 is used as a preconditioner for both PLMR versions, with the same MG components as in the previous test.
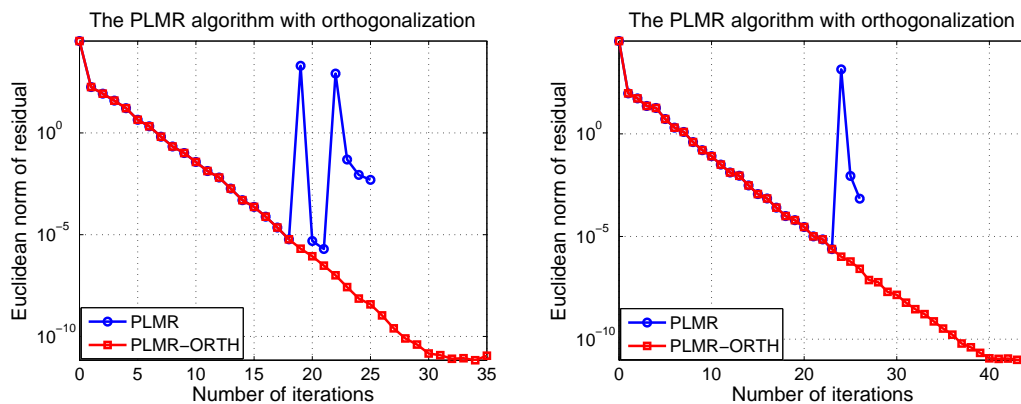
Figure 4.2: Comparison of the PLMR method with and without orthogonalization on the trial subspaces. Both versions of the method are applied to the model eigenproblem of the size $n = (2^7 - 1)^2 \approx 1.6 \times 10^4$ and use the MG absolute value preconditioner. The targeted eigenpairs correspond to the smallest magnitude eigenvalues of the shifted discrete negative Laplacian: $\lambda_{13} \approx -2.7426$ (left) and $\lambda_{15} \approx -3.4268$ (right), given by shift values $c^2 = 200$ and 250, respectively.

In Figure 4.2 we observe that, as the approximate eigenpairs get close to the exact solution of (4.23), the PLMR method, given by Algorithm 4.4, begins to exhibit instability, which can be fixed, however, by orthogonalizing the trial

subspaces. We relate this phenomenon to the one observed for the LOBPCG algorithm, and addressed, e.g., in [48, 38]. The nature of the possible instability is explained by an increasing ill-conditioning of the chosen basis of trial subspaces (4.14), as approximations $v^{(i)}$ converge to the desired eigenvector.

We note, however, that for problems, arising from discretizations of the underlying equations of mathematical physics, the required accuracy of the solution of the discrete problem is limited by the discretization error. For this reason, in practice, the PLMR algotithm can be expected to deliver the desired approximations without orthogonalizing the trial subspaces, i.e., as given by Algorithm 4.4.

4.4 Conclusions

In this chapter we have proposed a novel approach, which we call the PLMR method, for computing an approximation of the smallest, in the absolute value, eigenvalue and the corresponding eigenvector of a symmetric matrix pencil. The method represents a four-term recurrent iterative scheme, with iteration parameters determined by solving small auxiliary eigenvalue problems. The method is preconditioned. It requires an SPD preconditioner, which can be constructed according to the idea of the absolute value preconditioning described in the context of symmetric indefinite linear systems in the previous chapter. In fact, this allows to use the same SPD preconditioners for both symmetric indefinite linear systems and the corresponding interior eigenvalue problems.

We have applied the PLMR method to approximate an eigenpair of the two-dimensional discrete negative Laplace operator, which corresponds to the eigenvalue, closest to a given shift value. As a preconditioner we have reused

the (geometric) MG absolute value preconditioner, constructed for the corresponding linear system (the model problem) in the previous chapter. For a significant number of the initial steps, the PLMR method has demonstrated convergence behavior, comparable to that of an *idealized optimal* preconditioned eigenvalue solver.

The current and future work includes the extension of the present version of the PLMR algorithm to the block (subspace) iteration, its theoretical study, development of relevant codes, and investigation of their performance, including comparisons with the existing techniques, for various application areas.

5. Preconditioned singular value computations

Let us consider the problem of finding triplets (σ, v, u), such that

$$\begin{cases} A^* u = \sigma v \\ Av = \sigma u \end{cases}, \qquad A \in \mathbb{R}^{m \times n}, \ \sigma \in \mathbb{R}, \ u \in \mathbb{R}^m, \ v \in \mathbb{R}^n, \ \|u\| = \|v\| = 1. \quad (5.1)$$

We call problem (5.1) the *singular value problem*, and assume, without loss of generality, that $m \geq n$. Throughout the chapter, $\| \cdot \|$ denotes the Euclidean norm, defined on the vector space of the corresponding dimension.

The existence of the solution of problem (5.1) follows directly from the *singular value decomposition* (SVD) of a matrix A, see, e.g., [42], and is given by n triplets (σ_j, v_j, u_j), corresponding to the *singular values* σ_j of A, such that

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0.$$

The unit vectors v_j and u_j are called the *right and left singular vectors*, corresponding to the singular value σ_j, respectively, and are such that $(v_i, v_j) = (u_i, u_j) = 0$, $i \neq j$. $(\cdot, \cdot)$ denotes the standard inner product. We further call (σ_j, v_j, u_j) the *singular triplets*.

Problems of computing singular triplets, or finding the SVD of a matrix, are known to appear in a number of application areas, such as information retrieval, image and signal processing, seismic tomography; see [11]. Sometimes the singular value problems appear as a part of more complex computational tasks, e.g., such as constructing low-rank matrix approximations, solving least-squares problems, estimating matrix rank, computation of pseudospectrum, etc.

For small and dense matrices A, there exists a variety of efficient methods, which allow computing the SVD, i.e., delivering (possibly all) singular triplets of A in (5.1). Examples of such methods include: QR algorithm, DQDS, divide-and-conquer, Jacobi's method, etc.; see, e.g., [18, 28, 57, 19, 20, 1, 4]. In the present work, however, we assume A to be large and sparse. Moreover, only a tiny fraction of singular triplets, corresponding to the extreme singular values, is required. In this framework, the above mentioned methods can be inapplicable, which motivates the search for novel techniques.

Standard approaches for approximating singular triplets of large sparse matrices are based on substituting singular value problem (5.1) by a symmetric eigenvalue problem. As the first option, (5.1) can be replaced by the problem of finding eigenpairs of the matrix A^*A, i.e.,

$$A^*Av = \sigma^2 v. \tag{5.2}$$

In this case, the eigenvalues of problem (5.2) are the squared singular values of A in (5.1), while the corresponding eigenvectors, normalized to have a unit norm, are the *right* singular vectors. The *left* singular vectors are then computed as following,

$$u = \frac{Av}{\|Av\|} = \frac{Av}{\sigma}. \tag{5.3}$$

As the second option, instead of (5.1), one can consider the symmetric eigen-problem

$$\underbrace{\begin{bmatrix} 0 & A^* \\ A & 0 \end{bmatrix}}_{C} \begin{pmatrix} v \\ u \end{pmatrix} = \lambda \begin{pmatrix} v \\ u \end{pmatrix}. \tag{5.4}$$

111

The relation between (5.4) and singular value problem (5.1) is given by Jordan–Wielandt theorem, see, e.g., [66, 42].

Theorem 5.1 (Jordan-Wielandt) *The augmented matrix $C \in \mathbb{R}^{(m+n)\times(m+n)}$ in problem (5.4) has eigenvalues*

$$\left\{ -\sigma_1, \ldots, -\sigma_n, \underbrace{0, \ldots, 0}_{m-n}, \sigma_n, \ldots, \sigma_1 \right\}, \tag{5.5}$$

and eigenvectors, normalized to have a unit norm,

$$\frac{1}{\sqrt{2}} \begin{pmatrix} v_j \\ \pm u_j \end{pmatrix}, \qquad j = 1, \ldots, n, \tag{5.6}$$

corresponding to $\pm\sigma_j$, where v_j and u_j are the right and left singular vectors of A in (5.1), respectively.

Additionally, if $m > n$, then eigenvectors corresponding to the remaining $(m - n)$ zero eigenvalues are of the form

$$\begin{pmatrix} 0 \\ \tilde{u}_j \end{pmatrix}, \qquad n+1, \ldots, m, \tag{5.7}$$

where vectors $\tilde{u}_j \in \mathbb{R}^m$ can be chosen to be orthonormal.

Theorem 5.1 shows that the eigenvalues of C are plus and minus the singular values of A in (5.1). The singular vectors can be extracted from the corresponding eigenvectors in (5.6). The additional zero eigenvalue of multiplicity $(m - n)$ in (5.5) is entailed by the (part of) null space of the matrix $A^* \in \mathbb{R}^{n\times m}$, which naturally arises if $n < m$, regardless of the matrix rank. This (part of) null space of A^* is spanned by vectors $\tilde{u}_j$, which determine the eigenvectors in (5.7), corresponding to the zero eigenvalue.

Many of the existing algorithms for finding singular triplets, corresponding to the extreme singular values of large matrices, are based on the methods for solving symmetric eigenvalue problems, e.g., the Lanczos methods, the Davidson methods, shift-and-invert, and trace minimization techniques, *specifically tuned* for problem (5.2) or (5.4); see [11]. Usually, such algorithms are able to produce satisfactory results for computing the *largest* singular values and the corresponding singular vectors (the *largest* singular triplets). However, their application for finding the triplets corresponding to the *smallest* singular values (the *smallest* singular triplets), may often result in slow convergence and a lack of robustness. In this chapter, we focus on approximating the *smallest* singular triplet (σ_n, v_n, u_n) of the matrix A in (5.1), which typically represents a challenging computational problem.

If the *smallest* singular triplet is approximated by a standard approach, i.e., using one of the formulations based on a symmetric eigenvalue problem, it may be considered favorable to replace singular value problem (5.1) by (5.2)–(5.3). In this case, the *smallest* eigenvalue of the SPD matrix A^*A is σ_n^2, and the corresponding eigenvector v_n is, simultaneously, the right singular vector of A. The left singular vector u_n is computed by (5.3).

In practice, the approach based on computing the eigenpair (σ_n^2, v_n) of the matrix A^*A, using one of the available iterative eigenvalue solvers, may not lead to a satisfactory algorithm. The reasons for possible failures are commonly related to the observation that eigenproblem (5.2) can suffer from the increased clustering of its smallest eigenvalues, compared to the distribution of the corresponding singular values of A. The latter adversely effects the convergence be-

havior of many iterative eigenvalue solvers, e.g., based on the Lanczos method, see [56]. Further, forming the matrix of the normal equations A^*A squares the condition number of the initial problem, i.e., (5.1), which is generally undesirable for numerical algorithms, and can prohibit obtaining approximate solutions of the required accuracy. Also, the increased ill-conditioning may noticeably slow down the eigenvalue solvers, whose convergence depends on the condition number of the coefficient matrix, e.g., CG methods, see [47, 54].

A possible remedy may be to use a *preconditioned* method for finding the eigenpair (σ_n^2, v_n) of (5.2), e.g., the locally optimal preconditioned CG method, see [48], if a suitable preconditioner for A^*A is available. However, even if the method delivers a satisfactory approximation to the targeted vector v_n, formula (5.3) may result in an inaccurate approximation to the left singular vector u_n, see [40] for an example. The subsequent refinement procedures, e.g., based on a shift-and-invert approach, can be computationally expensive or inapplicable.

At the same time, let us note that eigenvalue problem (5.4) avoids potential difficulties caused by the squaring of small singular values. In fact, formulation (5.4) can be viewed as a matrix form of singular value problem (5.1), with σ, formally, replaced by λ. In this sense, the approach, based on finding the eigenpair corresponding to the eigenvalue σ_n (or, $-\sigma_n$) of the augmented matrix C in (5.4), may be regarded as more natural. The main complication here, however, lies in the fact that the desired eigenvalues $\lambda = \pm\sigma_n$ are far in the interior of the spectrum of the augmented matrix C.

In Chapter 4, we have described the preconditioned method PLMR for finding the smallest magnitude eigenvalue of a symmetric matrix. In fact, given a

suitable preconditioner, for square, or rectangular and rank-deficient, matrices A, the method can be *directly* applied to find the smallest, in the absolute value, eigenvalue of (5.4), i.e., $\pm\sigma_n$. The eigenpair, corresponding to this eigenvalue, then delivers the smallest singular triplet. We note that in the case, where A is full-rank and rectangular, the smallest magnitude eigenvalue in (5.4) is not $\pm\sigma_n \neq 0$, but zero of multiplicity $(m-n)$, see (5.5). This makes PLMR no longer *directly* applicable. Nevertheless, if ran on a properly chosen subspace, i.e., the orthogonal complement of the null space of the augmented matrix C, or, possibly, generalized to a subspace iteration, if $m \approx n$, the method can also deliver the targeted smallest triplet.

The known disadvantage of the approach for computing the singular triplets, which is based on using a method for finding eigenpairs of problem (5.4), is related to the treatment of the so-called "unbalanced" eigenvector approximations for problem (5.4), i.e., (unit) vectors of form

$$\begin{pmatrix} v \\ u \end{pmatrix}, \qquad v \in \mathbb{R}^n, \ u \in \mathbb{R}^m, \tag{5.8}$$

such that the norms of the decoupled components (subvectors) v and u are significantly different, e.g., $\|v\| >> \|u\|$. Usually, at each step of a method, the eigenvector approximations of the above form are chosen from a certain subspace of $\mathbb{R}^{m+n}$, using an *optimization principle*, e.g., (4.17) for the PLMR algorithm. The specificity of the singular value problem (5.1) typically requires obtaining *unit* approximations to the corresponding singular vectors at each iteration. In other words, instead of approximate eigenvectors in (5.8), it requires obtaining

vectors

$$\begin{pmatrix} \alpha v \\ \beta u \end{pmatrix}, \ \alpha = 1/\|v\|, \ \beta = 1/\|u\|, \ \alpha \neq \beta,$$

which no longer satisfy the optimization principle, used for extracting the approximate eigenvector for problem (5.4). The effects, caused by this departure from the optimization principle, can be especially noticeable at initial steps, where the approximate eigenvector is significantly far from the exact solution.

Finally, we note that there exists a number of methods for computing singular triplets of large matrices, which are not based on applying techniques for solving symmetric eigenvalue problems to (5.2) and (5.4), and directly target singular value problem (5.1). Many of such methods are based on the idea of Golub-Kahan-Lanczos bidiagonalization, introduced in [31]; see, e.g., [45, 49, 37]. Several recent works consider extensions of the Jacobi-Davidson type approach for computing a number of singular triplets; e.g., [40, 41]. Though some progress has been recently reported, the problem of finding the smallest singular triplets still remains computationally challenging and difficult.

In this chapter, we describe a new technique, that we call PLMR-SVD, to compute the smallest singular triplet. The proposed approach is based on the idea of using *two* separate low-dimensional trial subspaces for extracting approximations to the right and left singular vectors. Importantly, the suggested method is preconditioned, i.e., it allows using *two* operators, which, if properly chosen, can noticeably improve the convergence rate and robustness of the suggested scheme.

We note that the framework for introducing the method is similar to the one described in Chapter 4, where we derived the PLMR algorithm for computing an interior eigenpair. Thus, first, in Section 5.1, we describe the base idealized method for computing the singular triplet. Next, in Section 5.2, we use ideas, underlying the base method, to obtain the PLMR-SVD algorithm. Finally, in Section 5.3, we apply the new method to find the smallest singular triplet of the two-dimensional discrete gradient operator, using a multilevel SPD preconditioner.

5.1 Idealized preconditioned methods for finding a singular triplet

As has been previously observed, problem (5.4) can be formally viewed as singular value problem (5.1), written in the matrix form, with σ replaced by λ. Then, following the approach of Section 4.1 of the previous chapter, we assume that the singular value $\sigma = \sigma_n$ is *a priori* known, and consider the problem of finding a null space vector of the augmented matrix C, shifted by the value of σ_n, i.e.,

$$\underbrace{\begin{bmatrix} -\sigma_n I_n & A^* \\ A & -\sigma_n I_m \end{bmatrix}}_{C-\sigma_n I} \underbrace{\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}}_{x} = 0, \tag{5.9}$$

where $x_1 \in \mathbb{R}^n$, $x_2 \in \mathbb{R}^m$; $I_n \in \mathbb{R}^{n \times n}$, $I_m \in \mathbb{R}^{m \times m}$, and $I \in \mathbb{R}^{(m+n) \times (m+n)}$ are the identity matrices of corresponding dimensions.

By Theorem 5.1, σ_n is an eigenvalue of the augmented matrix C in (5.4), so the shifted matrix $C - \sigma_n I \in \mathbb{R}^{(m+n) \times (m+n)}$ is singular, and hence homogeneous system (5.9) has a nontrivial solution, which (after a suitable normalization) delivers the right and left singular vectors corresponding to σ_n, i.e., u_n and v_n.

In order to simplify the presentation, we assume that the singular values σ_j of A are distinct.

The coefficient matrix $C - \sigma_n I$ in (5.9) is symmetric and (highly) indefinite. We consider preconditioned iterative methods for finding a nonzero solution of (5.9) as the *idealized* methods for computing a singular triplet. In Chapter 3, we have introduced a range of methods, e.g., (3.17)–(3.18), (3.21) and (3.24), (3.25)–(3.26) with $S = T$, for solving nonsingular symmetric indefinite systems with SPD preconditioners. According to Proposition 4.1 of Chapter 4, under a mild assumption on the initial guess, these methods can be used to find a nonzero null space vector in (5.9). In this case, given an SPD, or, by Remark 4.2, possibly a symmetric positive semi-definite, preconditioner $T \in \mathbb{R}^{(m+n)\times(m+n)}$, the corresponding convergence bounds, with A replaced by C, λ_q by σ_n, and $B = I$, are stated in (4.6), (4.8).

Similarly to Section 4.1 of the previous chapter, as the *base* idealized method for computing the smallest singular triplet, we choose (3.21), (3.24) applied to solve homogeneous system (5.9). We first consider a correct setting for the chosen method and then define a proper preconditioner. We note that the discussion below is also valid for other idealized methods, i.e., (3.17)–(3.18) and (3.25)–(3.26) with $S = T$, applied to find a nonzero solution of system (5.9).

Let us assume that $\sigma_n \neq 0$ and $m > n$, i.e., the matrix A in (5.9), or in (5.1), is of full-rank and rectangular. In this case, according to Theorem 5.1, the spectrum of the augmented matrix C in (5.4) contains the zero eigenvalue of multiplicity $(m - n)$, entailed by the natural null space of A^*. Shifting C by the value of σ_n then generates the small eigenvalue $-\sigma_n$ of multiplicity $(m - n)$.

The presence of this eigenvalue in the spectrum of the matrix $C - \sigma_n I$ in (5.9) is merely an artifact of the rectangular structure of A. In general, unless a preconditioner of an extremely high quality is available, the small eigenvalue $-\sigma_n$ negatively effects the convergence of an idealized method by significantly increasing the effective condition number of $C - \sigma_n I$. The drawback, however, can be avoided by forcing the chosen *base* idealized iteration, i.e., method (3.21), (3.24), applied to solve system (5.9), to run on the orthogonal complement of the eigenspace corresponding to the unwanted eigenvalue $-\sigma_n$ of $C - \sigma_n I$. This orthogonal complement is, in fact, the range of the augmented matrix C in (5.4),

$$\mathcal{R}\{C\} = \left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} : x_1 \in \mathbb{R}^n, x_2 \in \mathcal{R}\{A\} \right\} \subset \mathbb{R}^{m+n}. \qquad (5.10)$$

Since the eigenspace corresponding to the eigenvalue $-\sigma_n$ of $C - \sigma_n I$ is the same as the null space of C, their orthogonality to the range of C is a direct consequence of the fact that

$$\mathcal{R}\{C\} = \mathcal{N}\{C\}^{\perp}.$$

We remark that the restriction of iterations to the range of the augmented matrix C, i.e., to subspace $\mathcal{R}\{C\}$ in (5.10), is natural, since the solution of system (5.9) itself belongs to $\mathcal{R}\{C\}$. In practice, this restriction can be fulfilled by choosing the initial guess from $\mathcal{R}\{C\}$, and requiring the preconditioner T to have $\mathcal{R}\{C\}$ as its invariant subspace. We note that, in this case, it is sufficient for T to be SPD, or, by Remark 4.2, possibly, symmetric positive semi-definite, at least on $\mathcal{R}\{C\}$. We further consider only the choice of preconditioners, which

are SPD on $\mathcal{R}\{C\}$, i.e., such that

$$\mathcal{R}\left\{T|_{\mathcal{R}\{C\}}\right\} = \mathcal{R}\{C\} \text{ and } T|_{\mathcal{R}\{C\}} = \left(T|_{\mathcal{R}\{C\}}\right)^* > 0, \qquad T \in \mathbb{R}^{(m+n)\times(m+n)}. \tag{5.11}$$

The semi-definite case, given by Remark 4.2, is used mainly for theoretical purposes, e.g., for defining an optimal preconditioner for problem (5.9).

We note that the considered case, where $\sigma_n \neq 0$ and $m > n$, is probably the most frequently addressed in singular value computations. In other cases, i.e., if $m = n$, or $\sigma_n = 0$ and $m > n$, the above discussion simplifies, since no "spurious" small eigenvalues are generated for $C - \sigma_n I$. Hence, for these problem parameters, the base idealized method for computing the smallest singular triplet is scheme (3.21), (3.24), straightforwardly applied to solve (5.9), with a random initial guess and an SPD preconditioner T. We now discuss the construction of the appropriate preconditioners.

According to Proposition 4.3 in the previous chapter, the optimal preconditioner for the base idealized method (3.21), (3.24), applied to solve (5.9), regardless of the matrix rank and dimensions, is defined as

$$T_{opt} = |C - \sigma_n I|^\dagger. \tag{5.12}$$

The exact computation of the optimal preconditioner T_{opt} in (5.12) is generally infeasible. For many practical cases, the value of σ_n is relatively small, e.g., compared to the norm of A. In these situations we suggest replacing the theoretically optimal preconditioner $T_{opt} = |C - \sigma_n I|^\dagger$ by $|C|^\dagger \approx T_{opt}$. Therefore, as reasonable (absolute value) preconditioners T for the base idealized method for finding the smallest singular triplet, we can choose *approximations* to $|C|^\dagger$.

In the case, where $\sigma_n \neq 0$ and $m > n$, such approximations need to satisfy assumptions in (5.11). In particular, this requires preconditioners T to be SPD at least on the range of the augmented matrix, i.e., on subspace (5.10). In the remaining cases, if $m = n$, or $\sigma_n = 0$ and $m > n$, the preconditioners T are only required to be SPD.

Let us observe that the absolute value of the augmented matrix $|C|$ has a block-diagonal form,

$$|C| = \left(C^2\right)^{\frac{1}{2}} = \begin{bmatrix} (A^*A)^{\frac{1}{2}} & 0 \\ 0 & (AA^*)^{\frac{1}{2}} \end{bmatrix}. \tag{5.13}$$

We note that the diagonal element $(A^*A)^{\frac{1}{2}}$ is, in fact, a symmetric positive (semi-) definite factor in the polar decomposition of A (further referred to as "the polar factor"); see, e.g., [42]. If A is square, then $(AA^*)^{\frac{1}{2}}$ is also a positive (semi-) definite polar factor, however, coming from the so-called *left* polar decomposition of A.

Structure (5.13) of $|C|$ motivates the following block-diagonal form of a preconditioner T for the base idealized method,

$$T = \begin{bmatrix} T_1 & 0 \\ 0 & T_2 \end{bmatrix}, \tag{5.14}$$

where $T_1 \in \mathbb{R}^{n \times n}$ and $T_2 \in \mathbb{R}^{m \times m}$. Since the preconditioner T needs to be chosen to approximate $|C|^\dagger$, the diagonal blocks in (5.14) are such that $T_1 \approx \left((A^*A)^{\frac{1}{2}}\right)^\dagger$ and $T_2 \approx \left((AA^*)^{\frac{1}{2}}\right)^\dagger$.

If $\sigma_n \neq 0$ and $m > n$, according to (5.10) and (5.11), the block T_1 must be SPD, while T_2 has to be SPD at least on the range of A, which represents its

121

invariant subspace, i.e.,

$$\mathcal{R}\left\{T_2|_{\mathcal{R}\{A\}}\right\} = \mathcal{R}\left\{A\right\} \text{ and } T_2|_{\mathcal{R}\{A\}} = \left(T_2|_{\mathcal{R}\{A\}}\right)^* > 0, \qquad T_2 \in \mathbb{R}^{m \times m}. \quad (5.15)$$

Thus, if $\sigma_n \neq 0$ and $m > n$, then $T_1 \approx (A^*A)^{-\frac{1}{2}}$, while, for example, $T_2 \approx P(AA^* + \alpha I_m)^{-\frac{1}{2}}P$, where P is an orthogonal projector on the range of A, and $\alpha \in \mathbb{R}$ is a small regularization parameter.

We note that in other cases, i.e., if $m = n$, or $\sigma_n = 0$ and $m > n$, blocks T_1 and T_2 must be SPD. In particular, if both A^*A and AA^* are nonsingular, then $T_1 \approx (A^*A)^{-\frac{1}{2}}$ and $T_2 \approx (AA^*)^{-\frac{1}{2}}$, otherwise, e.g., one can choose $T_1 \approx (A^*A + \alpha I_n)^{-\frac{1}{2}}$ and $T_2 \approx (AA^* + \alpha I_m)^{-\frac{1}{2}}$.

Finally, the block-diagonal structure of the preconditioner T in (5.14) allows us to write the chosen *base* idealized method for finding the smallest singular triplet, i.e., scheme (3.21), (3.24), applied to solve (5.9), in the following "decoupled" form,

$$
\begin{aligned}
x_1^{(i+1)} &= x_1^{(i)} + \alpha^{(i)} w_1^{(i)} + \beta^{(i)} T_1 \left(A^* w_2^{(i)} - \sigma_n w_1^{(i)}\right) + \gamma^{(i)} p_1^{(i)}, \ \ p_1^{(i)} = x_1^{(i)} - x_1^{(i-1)}, \\
x_2^{(i+1)} &= x_2^{(i)} + \alpha^{(i)} w_2^{(i)} + \beta^{(i)} T_2 \left(A w_1^{(i)} - \sigma_n w_2^{(i)}\right) + \gamma^{(i)} p_2^{(i)}, \ \ p_2^{(i)} = x_2^{(i)} - x_2^{(i-1)}, \\
w_1^{(i)} &= T_1 \left(A^* x_2^{(i)} - \sigma_n x_1^{(i)}\right), \ \ w_2^{(i)} = T_2 \left(A x_1^{(i)} - \sigma_n x_2^{(i)}\right), \ \ p_1^{(0)} = 0, \ p_2^{(0)} = 0, \\
&\quad i = 0, 1, \dots;
\end{aligned}
$$
$$(5.16)$$

where the parameters $\alpha^{(i)}$, $\beta^{(i)}$ and $\gamma^{(i)}$ are chosen to minimize the T-norm of the residual vector for the problem (5.9) over the corresponding low-dimensional subspace, as in (3.24). The iteration of form (5.16) has been obtained by splitting the terms in (3.21) according to the partitioning

$$x^{(i)} = \begin{pmatrix} x_1^{(i)} \\ x_2^{(i)} \end{pmatrix}, \qquad x_1^{(i)} \in \mathbb{R}^n, \ x_2^{(i)} \in \mathbb{R}^m, \qquad (5.17)$$

of the approximation $x^{(i)}$ to the solution of the augmented system (5.4).

As has been discussed above, the matrix (block) T_1 in (5.16) is assumed to be SPD. If $\sigma_n \neq 0$ and $m > n$, then T_2 must satisfy (5.15), and the initial guess $x^{(0)}$ has to be chosen from the range of the augmented matrix C, i.e.,

$$\begin{pmatrix} x_1^{(0)} \\ x_2^{(0)} \end{pmatrix}, \qquad x_1^{(0)} \in \mathbb{R}^n, \ x_2^{(0)} \in \mathcal{R}\{A\}.$$

The latter assumptions on T_2 and $x^{(0)}$ guarantee that iteration (5.16) is performed on the range of C. If $m = n$, or $\sigma_n = 0$ and $m > n$, then T_2 is taken to be SPD, $x^{(0)} \in \mathbb{R}^{m+n}$.

Thus, in the described setting, base idealized scheme (5.16) delivers a non-trivial solution of system (5.9), provided that the initial guess has a nonzero component from the null space of the augmented matrix $C - \sigma_n I$. In the next section, we use idealized method (5.16) as a starting point for deriving a practical algorithm for computing the smallest singular triplet.

5.2 The Preconditioned Locally Minimal Residual method for computing the smallest singular triplet

In this section, we introduce an iterative method, that we call the Preconditioned Locally Minimal Residual method for computing the smallest singular triplet, or PLMR-SVD. The method is based on two four-term recurrent relations for approximating the right and left singular vectors, respectively. Thus, at each step, the proposed scheme extracts singular vector approximations from two separate four-dimensional subspaces. The underlying extraction procedure is similar to the refined procedure for the augmented matrix, performed in the preconditioner-based norm.

Importantly, the PLMR-SVD algorithm can use *two* preconditioners to accelerate the convergence rate and improve the robustness. One of the preconditioner is required to be SPD, while the other needs to be either SPD, or, for rectangular matrices of the full rank, SPD at least on a certain subspace.

5.2.1 PLMR-SVD: The choice of trial subspaces

In the previous section, assuming that the targeted singular value σ_n is already known, we have described the *base* idealized method for computing the smallest singular triplet. The resulting scheme is given in (5.16). The latter corresponds to method (3.21), (3.24), applied to solve homogeneous system (5.9), with the involved terms decoupled into the "top" and "bottom" parts.

According to the discussion in Section 5.1, the parts $x_1^{(i)}$ and $x_2^{(i)}$ of the augmented iterates $x^{(i)}$ in (5.17), deliver, after suitable normalizations, the approximations of singular vectors v_n and u_n, respectively. We observe that at each step of base idealized method (5.16) the improved approximations $x_1^{(i+1)}$ and $x_2^{(i+1)}$ are chosen as elements of the following four-dimensional subspaces,

$$\mathrm{span}\left\{x_1^{(i)}, w_1^{(i)}, T_1\left(A^* w_2^{(i)} - \sigma_n w_1^{(i)}\right), x_1^{(i-1)}\right\}, \quad w_1^{(i)} = T_1\left(A^* x_2^{(i)} - \sigma_n x_1^{(i)}\right), \quad \text{and}$$
$$\mathrm{span}\left\{x_2^{(i)}, w_2^{(i)}, T_2\left(A w_1^{(i)} - \sigma_n w_2^{(i)}\right), x_2^{(i-1)}\right\}, \quad w_2^{(i)} = T_2\left(A x_1^{(i)} - \sigma_n x_2^{(i)}\right),$$
$$(5.18)$$

respectively, where $x_1^{(-1)} = 0$, $x_2^{(-1)} = 0$, $x_1^{(0)} \in \mathbb{R}^n$, and σ_n is the known smallest singular value. In this context, where we distinguish between the subspaces, which provide the new singular vector approximations, blocks T_1 and T_2 of the (augmented) preconditioner T in (5.14) can be viewed as *two separate* preconditioners. Thus, as discussed, the first preconditioner T_1 is chosen to be SPD. If $\sigma_n \neq 0$ and $m > n$, then the second preconditioner T_2 must satisfy (5.15), i.e.,

be SPD at least on $\mathcal{R}\{A\}$, moreover, the initial guess $x_2^{(0)}$ also needs to be from $\mathcal{R}\{A\}$. Otherwise, if $m = n$, or $\sigma_n = 0$ and $m > n$, the preconditioner T_2 is SPD and $x_2^{(0)} \in \mathbb{R}^m$.

Our goal is to construct a method for computing the smallest singular triplet, that mimics the behavior of the base idealized scheme, and extracts approximate singular vectors from two separate low-dimensional trial subspaces. Ideally, at step $(i + 1)$, we would like the subspaces for extracting the new approximations $v^{(i+1)}$ and $u^{(i+1)}$ to the right and left singular vectors v_n and u_n, respectively, to be of the same form as (5.18), with $x_1^{(i)} = v^{(i)}$ and $x_2^{(i)} = u^{(i)}$. In practice, however, since σ_n is not known exactly, subspaces (5.18), generally, cannot be computed. Instead, we replace σ_n by its (asymptotically quadratic) approximation, i.e., the (singular value) Rayleigh quotient

$$\sigma^{(i)} = \frac{(u^{(i)}, Av^{(i)})}{\|u^{(i)}\|\|v^{(i)}\|}. \tag{5.19}$$

Given current and previous (unit) singular vector approximations $v^{(i)}$, $u^{(i)}$, and $v^{(i-1)}$, $u^{(i-1)}$, SPD preconditioners T_1 and T_2 (the latter possibly needs to satisfy assumptions (5.15)), at step $(i + 1)$, we consider the following subspaces for extracting the new approximate singular vectors $v^{(i+1)}$ and $u^{(i+1)}$,

$$\text{span}\left\{v^{(i)}, w_1^{(i)}, T_1\left(A^*w_2^{(i)} - \sigma^{(i)}w_1^{(i)}\right), v^{(i-1)}\right\}, \quad w_1^{(i)} = T_1\left(A^*u^{(i)} - \sigma^{(i)}v^{(i)}\right),$$
$$\text{span}\left\{u^{(i)}, w_2^{(i)}, T_2\left(Aw_1^{(i)} - \sigma^{(i)}w_2^{(i)}\right), u^{(i-1)}\right\}, \quad w_2^{(i)} = T_2\left(Av^{(i)} - \sigma^{(i)}u^{(i)}\right), \tag{5.20}$$

respectively, where $v^{(-1)} = 0$, $u^{(-1)} = 0$, and, if $\sigma_n \neq 0$ and $m > n$, $u^{(0)} \in \mathcal{R}\{A\}$. We note that vectors $w_1^{(i)}$ and $w_2^{(i)}$ are the preconditioned residuals of the singular value problem (5.1), and are, in fact, the partial gradients of the singular value Rayleigh quotient in (5.19), evaluated at point $(v^{(i)}, u^{(i)})$.

125

Subspaces (5.20) suggest the preconditioned iteration, e.g., of the form

$$
\begin{aligned}
v^{(i+1)} &= \alpha_1^{(i)} v^{(i)} + \beta_1^{(i)} w_1^{(i)} + \gamma_1^{(i)} T_1 \left(A^* w_2^{(i)} - \sigma^{(i)} w_1^{(i)} \right) + \delta_1^{(i)} p_1^{(i)}, \\
u^{(i+1)} &= \alpha_2^{(i)} u^{(i)} + \beta_2^{(i)} w_2^{(i)} + \gamma_2^{(i)} T_2 \left(A w_1^{(i)} - \sigma^{(i)} w_2^{(i)} \right) + \delta_2^{(i)} p_2^{(i)}, \\
p_1^{(i)} &= v^{(i)} - \alpha_1^{(i-1)} v^{(i-1)}, \; p_2^{(i)} = u^{(i)} - \alpha_2^{(i-1)} u^{(i-1)}, \; p_1^{(0)} = 0, \; p_2^{(0)} = 0, \\
w_1^{(i)} &= T_1 \left(A^* u^{(i)} - \sigma^{(i)} v^{(i)} \right), \; w_2^{(i)} = T_2 \left(A v^{(i)} - \sigma^{(i)} u^{(i)} \right), \; i = 0, 1, \ldots ;
\end{aligned}
\tag{5.21}
$$

where $\alpha_l^{(i)}$, $\beta_l^{(i)}$, $\gamma_l^{(i)}$, and $\delta_l^{(i)}$, $l = 1, 2$, are some iteration parameters. The preconditioner T_1 in scheme (5.21) is SPD, $v^{(0)} \in \mathbb{R}^n$. According to the discussion in the previous section, in order to properly set up iteration (5.21), one must have a certain information about the specificity of the problem (5.1) under consideration. In particular, if $\sigma_n \neq 0$ and $m > n$, i.e., A is of full-rank and rectangular, the preconditioner T_2 in (5.21) needs to satisfy assumptions (5.15), with the initial guess for the right singular vector $u^{(0)} \in \mathcal{R}\{A\}$. Otherwise, i.e., for square or rectangular rank-deficient matrices, T_2 needs to be SPD with $u^{(0)} \in \mathbb{R}^m$. The choice of vectors $p_1^{(i)}$ and $p_2^{(i)}$ as weighted differences of the two consecutive singular vector approximations has been motivated by implementational considerations, mainly, to obtain a more stable calculation of trial subspaces, as has been pointed out in Subsection 4.2.2 of the previous chapter, where similar, implicitly computed, vectors have been introduced for finding interior eigenpairs.

Finally, let us note that, that at step $(i+1)$, scheme (5.21) searches the new approximate singular vectors in trial subspaces $\mathcal{V}^{(i+1)}$ and $\mathcal{U}^{(i+1)}$, such that

$$
\begin{aligned}
\mathcal{V}^{(i+1)} &= \operatorname{span}\left\{ v^{(i)}, w_1^{(i)}, T_1 \left(A^* w_2^{(i)} - \sigma^{(i)} w_1^{(i)} \right), v^{(i)} - \alpha_1^{i-1} v^{(i-1)} \right\}, \\
\mathcal{U}^{(i+1)} &= \operatorname{span}\left\{ u^{(i)}, w_2^{(i)}, T_2 \left(A w_1^{(i)} - \sigma^{(i)} w_2^{(i)} \right), u^{(i)} - \alpha_2^{(i-1)} u^{(i-1)} \right\},
\end{aligned}
\tag{5.22}
$$

which are the same (in the exact arithmetic) as (5.20), $v^{(-1)} = 0$, $u^{(-1)} = 0$.

We next describe the extraction procedure, i.e., the choice of the iteration parameters $\alpha_l^{(i)}$, $\beta_l^{(i)}$, $\gamma_l^{(i)}$, and $\delta_l^{(i)}$, $l = 1, 2$ in (5.21).

5.2.2 PLMR-SVD: The choice of iteration parameters

Given two k-dimensional subspaces $\mathcal{V} \subseteq \mathbb{R}^n$ and $\mathcal{U} \subseteq \mathbb{R}^m$, we want to extract unit approximations $v \in \mathcal{V}$ and $u \in \mathcal{U}$ to the right and left singular vectors corresponding to the smallest singular value σ_n, respectively.

Let us consider the subspace

$$\mathcal{I} = \left\{ \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} : z_1 \in \mathcal{V}, z_2 \in \mathcal{U} \right\} \subseteq \mathbb{R}^{m+n}, \tag{5.23}$$

generated by vectors from $\mathcal{V}$ and $\mathcal{U}$. Thus, our goal is to extract a vector $\begin{pmatrix} v \\ u \end{pmatrix}$ from $\mathcal{I}$ in (5.23), such that $\|v\| = \|u\| = 1$, with v and u approximating the right and left singular vectors, respectively.

Now let us assume that $\tilde{\sigma} \geq 0$ is some approximation to the smallest singular value σ_n, i.e., $\tilde{\sigma} \approx \sigma_n$, and consider the following vector $r \in \mathbb{R}^{m+n}$,

$$r = \begin{pmatrix} A^* z_2 - \tilde{\sigma} z_1 \\ A z_1 - \tilde{\sigma} z_2 \end{pmatrix}, \qquad \tilde{\sigma} \geq 0, \tag{5.24}$$

where $z_1 \in \mathcal{V}$ and $z_2 \in \mathcal{U}$. We note that if $\tilde{\sigma} = (z_2, A z_1)$, where $\|z_1\| = \|z_2\| = 1$, i.e., $\tilde{\sigma}$ is a (singular value) Rayleigh quotient (5.19) with $v^{(i)}$ and $u^{(i)}$ replaced by z_1 and z_2, respectively, then the vector r in (5.24) represents the residual vector of singular value problem (5.1), evaluated at z_1 and z_2. A norm of this vector is a reasonable quantity that can be used to assess the quality of approximations z_1 and z_2 to right and left singular vectors, respectively.

If $\tilde{\sigma} = \sigma_n$, the vector r in (5.24) turns into the corresponding residual vector of homogeneous system (5.9), which has been used as a starting point to derive base idealized method (5.16), i.e., scheme (3.21), (3.24), applied to solve (5.9). In order to mimic optimality principle (3.24) underlying the base idealized method, which minimizes the residual in a preconditioner-based norm, we suggest to extract approximations $v \in \mathcal{V}$ and $u \in \mathcal{U}$ to the right and left singular vectors v_n and u_n, respectively, which satisfy the following condition,

$$
\begin{pmatrix} v \\ u \end{pmatrix} = \operatorname*{argmin}_{\substack{z = (z_1^* \; z_2^*)^* \in \mathcal{I}, \\ \|z_1\| = \|z_2\| = 1}} \|r\|_T, \qquad r = \begin{pmatrix} A^* z_2 - \tilde{\sigma} z_1 \\ A z_1 - \tilde{\sigma} z_2 \end{pmatrix}, \; \tilde{\sigma} \geq 0 \qquad (5.25)
$$

where T is taken to be of block-diagonal form (5.14), the subspace $\mathcal{I}$ is defined in (5.23). According to the discussion in the previous sections, in the case where $\sigma_n \neq 0$ and $m > n$, we additionally assume that the subspace $\mathcal{U} \subseteq \mathcal{R}\{A\}$. This implies that the corresponding subspace $\mathcal{I}$ in (5.23) is in the range of the augmented matrix C, i.e., in the subspace $\mathcal{R}\{C\}$ in (5.10). The latter guarantees that the vector r from (5.24), whose norm is minimized in (5.25), is also an element of $\mathcal{R}\{C\}$, since $A^* z_2 - \tilde{\sigma} z_1 \in \mathbb{R}^n$ and $A z_1 - \tilde{\sigma} z_2 \in \mathcal{R}\{A\}$, provided that $z_1 \in \mathcal{V}$ and $z_2 \in \mathcal{U} \subseteq \mathcal{R}\{A\}$. The choice of T_1 to be SPD and T_2 to satisfy (5.15), leads to the operator T in (5.14), such that (5.11) holds, i.e., T is SPD on $\mathcal{R}\{C\}$, which is the invariant subspace of T. This means that on $\mathcal{R}\{C\}$, T generates a norm that can indeed be used for the minimization in (5.25). If $m = n$, or $\sigma_n = 0$ and $m > n$, the discussion simplifies, since, according to the previous section, both T_1 and T_2 are chosen to be SPD, hence, the operator T in (5.14) is SPD on $\mathbb{R}^{m+n}$, and generates a norm in (5.25).

Let matrices $V \in \mathbb{R}^{n \times k}$ and $U \in \mathbb{C}^{m \times k}$ be such that $\mathrm{col}(V) = \mathcal{V}$ and $\mathrm{col}(U) = \mathcal{U}$. Then vectors $z_1 \in \mathcal{V}$ and $z_2 \in \mathcal{U}$ can be represented as $z_1 = V y_1$ and $z_2 = U y_2$, where $y_1, y_2 \in \mathbb{R}^k$. This allows us to write the vector r in (5.24) in the following form,

$$
\begin{aligned}
r &= \begin{pmatrix} A^* z_2 - \tilde{\sigma} z_1 \\ A z_1 - \tilde{\sigma} z_2 \end{pmatrix} = \begin{bmatrix} -\tilde{\sigma} I_n & A^* \\ A & -\tilde{\sigma} I_m \end{bmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \\
&= \begin{bmatrix} -\tilde{\sigma} I_n & A^* \\ A & -\tilde{\sigma} I_m \end{bmatrix} \begin{bmatrix} V & 0 \\ 0 & U \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{bmatrix} -\tilde{\sigma} V & A^* U \\ A V & -\tilde{\sigma} U \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}.
\end{aligned}
$$

Thus, by (5.14), we get

$$
\begin{aligned}
\|r\|_T^2 &= (r, T r) \\
&= (\begin{bmatrix} -\tilde{\sigma} V & A^* U \\ A V & -\tilde{\sigma} U \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \begin{bmatrix} T_1 & 0 \\ 0 & T_2 \end{bmatrix} \begin{bmatrix} -\tilde{\sigma} V & A^* U \\ A V & -\tilde{\sigma} U \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}) \\
&= (\underbrace{\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}}_{y}, \underbrace{\begin{bmatrix} -\tilde{\sigma} V^* & V^* A^* \\ U^* A & -\tilde{\sigma} U^* \end{bmatrix} \begin{bmatrix} T_1 & 0 \\ 0 & T_2 \end{bmatrix} \begin{bmatrix} -\tilde{\sigma} V & A^* U \\ A V & -\tilde{\sigma} U \end{bmatrix}}_{D} \underbrace{\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}}_{y}).
\end{aligned} \tag{5.26}
$$

The obtained expression for the (squared) norm of the vector r allows us to substitute minimization problem (5.25) by finding vectors $y_{1,min}, y_{2,min} \in \mathbb{R}^k$, such that

$$
\begin{pmatrix} y_{1,min} \\ y_{2,min} \end{pmatrix} = \mathop{\mathrm{argmin}}_{\substack{y = (y_1^* \ y_2^*)^* \in \mathbb{R}^{2k}, \\ y_1 \in \mathbb{R}^k, \ y_2 \in \mathbb{R}^k, \\ \|V y_1\| = \|U y_2\| = 1}} (y, D y), \tag{5.27}
$$

where $D \in \mathbb{R}^{2k \times 2k}$, after multiplying the matrices in (5.26), is given by

$$D = \begin{bmatrix} V^*A^*T_2AV + \tilde{\sigma}^2 V^*T_1V & -\tilde{\sigma}V^*T_1A^*U - \tilde{\sigma}V^*A^*T_2U \\ -\tilde{\sigma}U^*AT_1V - \tilde{\sigma}U^*T_2AV & U^*AT_1A^*U^* + \tilde{\sigma}^2U^*T_2U \end{bmatrix}. \quad (5.28)$$

Thus, the solution of the quadratically constrained quadratic optimization problem (5.27)–(5.28) determines the corresponding minimizer in (5.25), i.e., vectors $v \in \mathcal{V}$ and $u \in \mathcal{U}$, by

$$v = Vy_{1,min}, \text{ and } u = Uy_{2,min}, \qquad \|v\| = \|u\| = 1, \quad (5.29)$$

which deliver the new singular vector approximations.

As has been discussed in the previous section, at a general step $(i + 1)$, according to (5.21) and (5.22), we choose the two trial subspaces for approximating the right and left singular vectors as spans of four vectors, i.e., $\mathcal{V} = \mathcal{V}^{(i+1)}$ and $\mathcal{U} = \mathcal{U}^{(i+1)}$, respectively, $k = 4$. The choice of the trial subspaces generates the corresponding subspace $\mathcal{I}$ in (5.23) with $\mathcal{V} = \mathcal{V}^{(i+1)}$ and $\mathcal{U} = \mathcal{U}^{(i+1)}$. The new approximate singular vectors $v = v^{(i+1)}$ and $u = u^{(i+1)}$, satisfying condition (5.25) for a (presumably) given $\tilde{\sigma} = \tilde{\sigma}^{(i)}$, are determined by (5.29) after finding the solutions $y_{1,min}, y_{2,min}$ of optimization problem (5.27)–(5.28) with 8 unknowns. The matrices V and U have vectors from $\mathcal{V}^{(i+1)}$ and $\mathcal{U}^{(i+1)}$ in (5.22) as their columns, respectively. The iteration parameters $\alpha_l^{(i)}$, $\beta_l^{(i)}$, $\gamma_l^{(i)}$, and $\delta_l^{(i)}$ in (5.21) are determined by the components of the corresponding vectors $y_{l,min}$, $l = 1, 2$. At the initial step $(i = 0)$, the trial subspaces (5.22) are spanned by three vectors, i.e., $k = 3$, and, hence, the extraction of the singular vector approximations reduces to the solution of problem (5.27)–(5.28) with 6 unknowns. We next consider the choice of the singular value approximations $\tilde{\sigma} = \tilde{\sigma}^{(i)}$ in (5.25).

Let us first note that in a vicinity of the solution of problem (5.1), i.e., if $v^{(i)}$ and $u^{(i)}$ are already close to the targeted singular vectors v_n and u_n, a good choice of the parameter $\tilde{\sigma} = \tilde{\sigma}^{(i)}$ in (5.25) can be obtained by setting $\tilde{\sigma}^{(i)}$ to the value of the Rayleigh quotient (5.19), i.e.,

$$\tilde{\sigma}^{(i)} = \sigma^{(i)}.$$

In the general case, we suggest to obtain the smallest singular value estimates $\tilde{\sigma} = \tilde{\sigma}^{(i)}$ by performing two separate Rayleigh-Ritz procedures, one for problem (5.2), and the other for problem

$$AA^*u = \sigma^2 u, \tag{5.30}$$

on the already available trial subspaces $\mathcal{V}^{(i+1)}$ and $\mathcal{U}^{(i+1)}$, defined in (5.22), respectively. Then, assuming that $\tilde{v}^{(i)}$ and $\tilde{u}^{(i)}$ are the (unit) Ritz vectors, corresponding to the smallest Ritz values of (5.2) and (5.30) on $\mathcal{V}^{(i+1)}$ and $\mathcal{U}^{(i+1)}$, respectively, we set $\tilde{\sigma} = \tilde{\sigma}^{(i)}$ to the absolute value of the Rayleigh quotient (5.19), evaluated at $\tilde{v}^{(i)}$ and $\tilde{u}^{(i)}$, i.e.,

$$\tilde{\sigma}^{(i)} = \left| (\tilde{u}^{(i)}, A\tilde{v}^{(i)}) \right|, \ \|\tilde{v}^{(i)}\| = \|\tilde{u}^{(i)}\| = 1, \tag{5.31}$$

and discard the corresponding Ritz values. We note that, as discussed, in the case where $\sigma_n \neq 0$ and $m > n$, at each step, the trial subspace $\mathcal{U}^{(i+1)}$ for approximating the left singular vector is in the range of A, i.e., $\mathcal{U}^{(i+1)} \subseteq \mathcal{R}\{A\}$, which is orthogonal to the null space of A^*, and to the null space of AA^*. The Ritz vector, given by the Rayleigh-Ritz procedure for (5.30) on $\mathcal{U}^{(i+1)}$, delivers an approximation to the eigenvector corresponding to the smallest *nonzero* eigenvalue of AA^*, and can indeed be used for estimating the smallest singular value $\sigma_n \neq 0$ by (5.31). The described approach results in the following algorithm.

Algorithm 5.2 (The PLMR-SVD algorithm)

Input: starting vectors $v^{(0)}$ and $u^{(0)}$, functions to compute Av, A^*u, T_1v, T_2v

      If $m \neq n$ and A is of full rank, then $u^{(0)} \in \mathcal{R}\{A\}$ and T_2 satisfies (5.15)

Output: approximation to the smallest singular triplet (σ_n, v_n, u_n)

1. *Start: Normalize $v^{(0)}$, $u^{(0)}$ and set $p_1^{(0)} = 0$, $p_2^{(0)} = 0$*

2. *Iterate: For $i = 0, 1, \ldots,$ Until Convergence Do:*

3. $\sigma^{(i)} := \left(u^{(i)}, Av^{(i)}\right)$, $r_1 := A^*u^{(i)} - \sigma^{(i)}v^{(i)}$, $r_2 := Av^{(i)} - \sigma^{(i)}u^{(i)}$

4. $w_1^{(i)} := T_1 r_1$, $w_2^{(i)} := T_2 r_2$,
    $s_1^{(i)} := T_1(A^*w_2^{(i)} - \sigma^{(i)}w_1^{(i)})$, $s_2^{(i)} := T_2(Aw_1^{(i)} - \sigma^{(i)}w_2^{(i)})$

5.   *Use the Rayleigh-Ritz method for (5.2) and (5.30) on the trial subspaces*
    $\mathrm{span}\left\{v^{(i)}, w_1^{(i)}, s_1^{(i)}, p_1^{(i)}\right\}$ *and* $\mathrm{span}\left\{u^{(i)}, w_2^{(i)}, s_2^{(i)}, p_2^{(i)}\right\}$, *respectively*

6. $\tilde{\sigma} := |(\tilde{u}, A\tilde{v})|$ *($\tilde{v}$ and $\tilde{u}$ are the Ritz vectors corresponding to the smallest*
    Ritz values in 5)

7.   *If $i > 0$, then $V := \left[v^{(i)}; w_1^{(i)}; s_1^{(i)}; p_1^{(i)}\right]$, $U := \left[u^{(i)}; w_2^{(i)}; s_2^{(i)}; p_2^{(i)}\right]$;*
    else $V := \left[v^{(i)}; w_1^{(i)}; s_1^{(i)}\right]$, $U := \left[u^{(i)}; w_2^{(i)}; s_2^{(i)}\right]$

8.   *Solve (5.27). Set $(\alpha_1^{(i)}\ \beta_1^{(i)}\ \gamma_1^{(i)}\ \delta_1^{(i)}) := y_{1,min}^*$, $(\alpha_2^{(i)}\ \beta_2^{(i)}\ \gamma_2^{(i)}\ \delta_2^{(i)}) := y_{2,min}^*$*

9. $v^{(i+1)} := \alpha_1^{(i)}v^{(i)} + \beta_1^{(i)}w_1^{(i)} + \gamma_1^{(i)}s_1^{(i)} + \delta_1^{(i)}p_1^{(i)}$
    $u^{(i+1)} := \alpha_2^{(i)}u^{(i)} + \beta_2^{(i)}w_2^{(i)} + \gamma_2^{(i)}s_2^{(i)} + \delta_2^{(i)}p_2^{(i)}$

10. $p_1^{(i+1)} := \beta_1^{(i)}w_1^{(i)} + \gamma_1^{(i)}s_1^{(i)} + \delta_1^{(i)}p_1^{(i)}$
    $p_2^{(i+1)} := \beta_2^{(i)}w_2^{(i)} + \gamma_2^{(i)}s_2^{(i)} + \delta_2^{(i)}p_2^{(i)}$ *EndDo*

We finally note that similar to the PLMR algorithm for eigenvalue computations, introduced in the previous chapter, the PLMR-SVD method may require orthogonalization on the trial subspaces as the approximations $v^{(i)}$ and $u^{(i)}$ get closer to the targeted singular vectors. Also, as has been previously suggested, if the triplet $(\sigma^{(i)}, v^{(i)}, u^{(i)})$ is near the exact solution (σ_n, v_n, u_n), then one can skip step 5 of Algorithm 5.2, and set $\tilde{\sigma}$ to the current value of the (singular value) Rayleigh quotient $\sigma^{(i)}$ at step 6.

In the next section we apply the PLMR-SVD method to compute the smallest singular triplet of a two-dimensional discrete gradient operator.

5.3 Numerical example

In this concluding section we use the PLMR-SVD method, given by Algorithm 5.2, to compute the smallest singular triplet of the gradient operator, discretized on a unit square, assuming Dirichlet boundary conditions. The discretization is performed using finite differences, in such a way that the matrix of the normal equations of the resulting operator G, i.e., $L = G^*G$, is exactly the discrete negative Laplacian, considered in the model problems of Chapters 3 and 4. It is clear that the number of rows of the matrix G is approximately twice the number of its columns. The transpose of G represents the corresponding discrete divergence operator.

Let us note that, in fact, computing the singular triplets of the gradient G can possibly be a reasonable alternative to finding the respective eigenpairs of the negative Laplacian $L = G^*G$, in the case where both the eigenmodes of the latter *and their gradients* are desired. In particular, as has been pointed out at the beginning of this chapter, once an approximate eigenvector of the matrix

$L = G^*G$, corresponding to the smallest eigenvalue is found, the computation of its gradient, i.e., in the operator terms, the multiplication by G, may deliver highly inaccurate results. This complication may be avoided, e.g., by replacing the eigenvalue problem by the corresponding singular value problem.

As has been shown in the previous sections, the PLMR-SVD algorithm can simultaneously use two preconditioners T_1 and T_2. For problem (5.1), with A chosen to be the discrete gradient G, the preconditioner T_1 is such that

$$T_1 \approx (G^*G)^{-\frac{1}{2}} = (L)^{-\frac{1}{2}}, \tag{5.32}$$

where L is the discrete negative Laplacian, $(G^*G)^{\frac{1}{2}}$ is the polar factor of G. We further call preconditioners T_1, which are constructed according to the idea of approximation of the inverted polar factor, the *polar factor preconditioners*. The preconditioner T_2 needs to approximate, e.g.,

$$T_2 \approx (GG^* + \alpha I)^{-\frac{1}{2}},$$

where α is a small real parameter. Moreover, since G is rectangular and of full rank, T_2 needs to satisfy (5.15).

In our experiment we show that even the introduction of only one preconditioner, i.e., the polar factor preconditioner T_1 in (5.32), can give significantly improved results, compared, e.g., to the *unpreconditioned* idealized methods for computing the smallest singular triplet, discussed in Section 5.1.

In order to construct the preconditioner T_1 in (5.32), we apply the MG technique similar to Algorithm 3.9, which has been used as the absolute value preconditioner for the discrete Helmholtz equation. In particular, we suggest to perform the smoothing steps using the negative Laplace operator, i.e., the matrix

$L = G^*G$ of the normal equations, and invert the polar factor $(L)^{\frac{1}{2}} = (G^*G)^{\frac{1}{2}}$ of the gradient G on the coarse grid. For consistency, we state the corresponding two-grid scheme and its multilevel extension.

In the two-grid context, we use the subscript H to refer to the coarse-grid quantities. For example, G_H and L_H denote the gradient and the negative Laplace operator, discretized on the coarse grid of mesh size H, respectively. No subscript is used for denoting the fine-grid components.

Algorithm 5.3 (Two-grid polar factor preconditioner T_1)

Input r, output w.

1. *Pre-smoothing. Apply ν pre-smoothing steps (for the problem $Lw = r$) with the zero initial guess ($w^{(0)} = 0$):*

$$w^{(i+1)} = w^{(i)} + M^{-1}(r - Lw^{(i)}), \ i = 0, \ldots, \nu - 1,$$

where the (nonsingular) matrix M defines the choice of a smoother. This step results in the pre-smoothed vector $w^{pre} = w^{(\nu)}$, $\nu \geq 1$.

2. *Coarse grid correction. Restrict the vector $r - Lw^{pre}$ to the coarse grid, multiply it by the inverted coarse-level polar factor $(G_H^*G_H)^{\frac{1}{2}} = (L_H)^{\frac{1}{2}}$ of the gradient, and then prolongate the result back to the fine grid. This delivers the coarse-grid correction, which is added to w^{pre} to obtain the corrected vector w^{cgc}:*

$$w_H = (G_H^*G_H)^{-\frac{1}{2}} R\left(r - Lw^{pre}\right), \tag{5.33}$$

$$w^{cgc} = w^{pre} + Pw_H, \tag{5.34}$$

where P and R are prolongation and restriction operators, respectively.

3. *Post-smoothing. Apply ν post-smoothing steps (for the problem $Lw = r$) with the initial guess $w^{(0)} = w^{cgc}$:*

$$w^{(i+1)} = w^{(i)} + M^{-*}(r - Lw^{(i)}), \ \ i = 0, \ldots, \nu - 1.$$

This step results in the post-smoothed vector $w^{post} = w^{(\nu)}$. Return the vector $w = w^{post}$.

The two-grid preconditioner $T_1 = T_{1,tg}$, constructed by Algorithm 5.3, has the following structure,

$$T_{1,tg} = \left(I - M^{-*}L\right)^\nu P \left(G_H^* G_H\right)^{-\frac{1}{2}} R \left(I - LM^{-1}\right)^\nu + S, \tag{5.35}$$

with $S = L^{-1} - \left(I - M^{-*}L\right)^\nu L^{-1} \left(I - LM^{-1}\right)^\nu$. The symmetry and positive definiteness are justified in the same way as for the absolute value preconditioner in (3.39), constructed according to Algorithm 3.8; see Subsection 3.2.2.1.

Now let us assume that a hierarchy of $m + 1$ grids is available, and the grids are numbered by $l = m, m - 1, \ldots, 0$ with the corresponding mesh sizes $\{h_l\}$ in the decreasing order. To extend the two-grid polar factor preconditioner given by Algorithm 5.3 to the *multigrid*, we replace the inversion of the polar factor $(G_H^* G_H)^{\frac{1}{2}}$ in step 2 (formula (5.33)), by the recursive application of the algorithm to the restricted vector $R(r - Lw^{pre})$. This approach is then followed on all levels, with the exact inversion of the polar factor of the discrete gradient operator on the coarsest grid.

If started from the finest grid $l = m$, the following scheme gives the multilevel extension of the two-grid polar factor preconditioner defined by Algorithm 5.3. We note that the subscript l is introduced to match the occurring quantities to the corresponding grid.

Algorithm 5.4 (PFP-MG(r_l): MG polar factor preconditioner T_1)

Input r_l, output w_l.

1. *Pre-smoothing. Apply ν pre-smoothing steps (for the problem $L_l w_l = r_l$) with the zero initial guess ($w_l^{(0)} = 0$):*

$$w_l^{(i+1)} = w_l^{(i)} + M_l^{-1}(r_l - L_l w_l^{(i)}), \ \ i = 0, \ldots, \nu - 1,$$

   where the (nonsingular) matrix M_l defines the choice of a smoother on level l. This step results in the pre-smoothed vector $w_l^{pre} = w_l^{(\nu)}$, $\nu \geq 1$.

2. *Coarse grid correction. Restrict the vector $r_l - L_l w_l^{pre}$ to the grid $l - 1$. If $l = 1$, then multiply the restricted vector by the inverted coarse-level polar factor $(G_0^* G_0)^{\frac{1}{2}}$,*

$$w_0 = (G_0^* G_0)^{-\frac{1}{2}} R_0 \left(r_1 - L_1 w_1^{pre} \right), \ \ if \ l = 1. \tag{5.36}$$

   Otherwise, recursively apply PFP-MG to approximate the action of the inverted polar factor $\left(G_{l-1}^* G_{l-1} \right)^{\frac{1}{2}}$ on the restricted vector,

$$w_{l-1} = PFP\text{-}MG \left(R_{l-1} \left(r_l - L_l w_l^{pre} \right) \right), \ \ if \ l > 1. \tag{5.37}$$

   Prolongate the result back to the fine grid. This delivers the coarse-grid correction, which is added to w_l^{pre} to obtain the corrected vector w_l^{cgc}:

$$w_l^{cgc} = w_l^{pre} + P_l w_{l-1}, \tag{5.38}$$

   where w_{l-1} is given by (5.36)–(5.37). The operators R_{l-1} and P_l define the restriction from the level l to $l - 1$ and the prolongation from the level $l - 1$ to l, respectively.

3. *Post-smoothing. Apply ν post-smoothing steps (for the problem $L_l w_l = r_l$) with the initial guess $w_l^{(0)} = w_l^{cgc}$:*

$$w_l^{(i+1)} = w_l^{(i)} + M_l^{-*}(r_l - L_l w_l^{(i)}), \ \ i = 0, \ldots, \nu - 1.$$

This step results in the post-smoothed vector $w_l^{post} = w_l^{(\nu)}$. Return the vector $w_l = w_l^{post}$.

Similar to Algorithm 3.9 in Subsection 3.2.2.1, the multigrid polar factor preconditioner $T_1 = T_{1,mg}$, constructed according to Algorithm 5.4, has the following structure,

$$T_{1,mg} = \left(I - M^{-*}L\right)^{\nu} P T_{1,mg}^{(m-1)} R \left(I - LM^{-1}\right)^{\nu} + S, \tag{5.39}$$

with S as in (5.35) and $T_{1,mg}^{(m-1)}$ defined according to the recursion below,

$$T_{1,mg}^{(l)} = \left(I_l - M_l^{-*}L_l\right)^{\nu} P_l T_{1,mg}^{(l-1)} R_{l-1} \left(I_l - L_l M_l^{-1}\right)^{\nu} + S_l, \ l = 1, \ldots, m - 1,$$
$$T_{1,mg}^{(0)} = (G_0^* G_0)^{-\frac{1}{2}}, \tag{5.40}$$

where $S_l = L_l^{-1} - \left(I_l - M_l^{-*}L_l\right)^{\nu} L_l^{-1} \left(I_l - L_l M_l^{-1}\right)^{\nu}$. In (5.39) we skip the subscript in the notation for the quantities associated with the finest level $l = m$. The symmetry and positive definiteness of $T_1 = T_{1,mg}$, defined by (5.39)–(5.40), are justified in the same way as for the absolute value preconditioner in (3.45)–(3.46), constructed according to Algorithm 3.9; see Subsection 3.2.2.1.

In Figure 5.1 (left) we show the improved convergence behavior of the PLMR-SVD method with the preconditioner T_1, constructed according to the multilevel Algorithm 5.4, compared to the *unpreconditioned* idealized singular value solver, based on the (globally optimal) MINRES algorithm, applied to
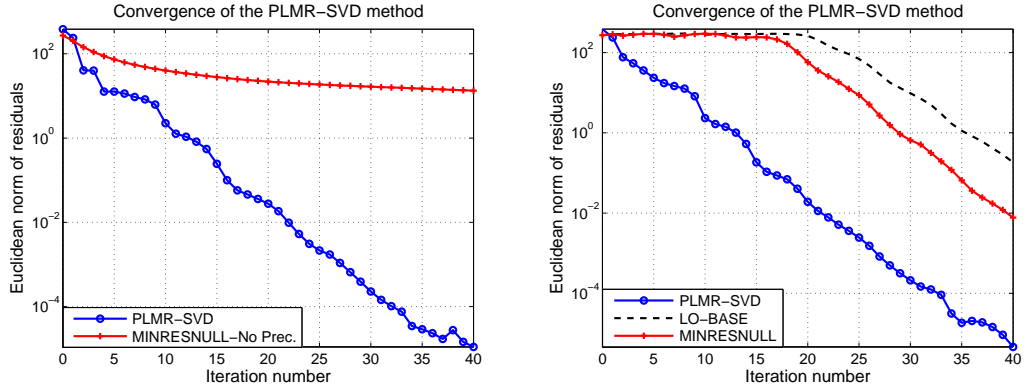
Figure 5.1: Comparison of the PLMR-SVD method with one MG precondi-
tioner versus the idealized singular value solvers, applied to find the smallest sin-
gular triplet of the m-by-n discrete gradient operator, $n = (2^7 - 1)^2 \approx 1.6 \times 10^4$,
$m \approx 2n$.

system (5.9), where the exact smallest singular value is known. As discussed in
Section 4.3 of the previous chapter, such idealized solver is obtained by modify-
ing the matlab funcion "minres.m", and is referred to as "MINRESNULL". We
discretize the gradient G on the grid of the mesh size $h = 2^{-7}$, initial singular
vector approximations are randomly chosen, with the initial left singular vector
in the range of G. The MG components for the preconditioner are defined simi-
larly to Subsection 3.2.2.2, with one step of the 4/5-damped Jacobi iteration as
a (pre- and post-) smoother, standard coarsening scheme with the coarsest grid
of the mesh size 2^{-4}, full weighting for the restriction, and piecewise multilinear
interpolation for the prolongation.

In Figure 5.1 (right) we compare the PLMR-SVD algorithm with the *pre-
conditioned* "control" methods introduced in Section 4.3, i.e., (preconditioned)
"MINRESNULL" and "LO-BASE", which is the base idealized scheme (3.21),
(3.24), applied to (5.9), with the known smallest singular value. As a precondi-

139

tioner for both methods we use an operator of the form

$$
\begin{bmatrix} T_1 & 0 \\ 0 & I_m \end{bmatrix},
$$

where the action of T_1 is constructed according to Algorithm 5.4. The test setting, including the definition of the MG components for constructing T_1, is the same as described in the previous paragraph. Figure 5.1 (right) demonstrates that the convergence rate of PLMR-SVD, at least at a significant number of the initial steps, is similar to that of the *idealized* methods.

We note that both figures compare the norms of the residual vectors for singular problem (5.1), with A replaced by G, produced by the PLMR-SVD algorithm, i.e.,

$$
\sqrt{\|G^* u^i - \sigma^{(i)} v^{(i)}\|^2 + \|G v^i - \sigma^{(i)} u^{(i)}\|^2},
$$

versus residual norms $\dfrac{\|(C - \sigma_n I) x^{(i)}\|}{\|x^{(i)}\|}$, given by "MINRESNULL" and "LO-BASE", evaluated at the (augmented) normalized iterates $\dfrac{x^{(i)}}{\|x^{(i)}\|}$; $C - \sigma_n I$ is the shifted augmented matrix in (5.9). We finally remark that, according to the discussion in Section 4.3, in order to meaningfully match the numbering of iterations of the three considered methods, we plot the values of the "MINRES-NULL" residual norms, which are measured after every other step. In other words, the "MINRESNUL" residual norm at step i, in Figure 5.1, corresponds to the norm of the MINRES (or, PMINRES) algorithm, applied to (5.9), at step $j = 2i$, evaluated at the normalized iterate. Small quadratically constrained quadratic problems (5.27)–(5.28), at each step of the PLMR-SVD algorithm, are solved using the "fmincon" function from the matlab optimization toolbox,

set up to use the interior point method with the exactly provided gradients and Hessians, with the tolerance level for the approximate solution equal to 10^{-6}, and random initial guess.

5.4 Conclusions

In this concluding chapter we have described a new technique, called the PLMR-SVD method, for computing the singular triplet corresponding to the smallest singular value of a general rectangular matrix. The method represents an iterative scheme, which is based on two linked four-term recurrent relations for approximating the right and left singular vectors, respectively. The iteration parameters at each step are determined as solutions of small quadratically constrained quadratic optimization problems. The method uses two SPD preconditioners. In particular, one of the preconditioners can be chosen to approximate an inverse of the symmetric positive (semi-) definite factor in the polar decomposition of the problem matrix. At the initial phase, the PLMR-SVD algorithm requires information about the dimensions of the input matrix (square or rectangular) and, possibly, about its rank (full rank or rank deficient).

In order to assess the performance of the PLMR-SVD algorithm, we have applied it to the model problem of finding the singular triplet corresponding to the smallest singular value of the two-dimensional discrete gradient operator. In our tests we have used only one of the two SPD preconditioners allowed by the method. This preconditioner has been constructed to approximate the inverse of the SPD polar factor of the discrete gradient using the (geometric) MG approach. In particular, we have shown that the use of only one preconditioner provides a significant improvement in the convergence rate as compared to *unpreconditioned*

idealized optimal singular value solvers. The construction of an example of the second preconditioner, in order to further accelerate the convergence, is one of the current goals of the related research. Other goals include the extension of the present version of the PLMR-SVD algorithm to the block (subspace) iteration, the theoretical study of the method, as well as the development and application of the relevant software.

REFERENCES

[1] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. *LAPACK Users' Guide.* Society for Industrial and Applied Mathematics, Philadelphia, PA, third edition, 1999.

[2] M. Arioli, , V. Pták, and Z. Strakoš. Krylov sequences of maximal length and convergence of GMRES. *BIT*, 38(4):636–643, 1998.

[3] O. Axelsson. *Iterative solution methods.* Cambridge University Press, New York, NY, 1994.

[4] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, editors. *Templates for the solution of algebraic eigenvalue problems.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000.

[5] A. H. Baker, E. R. Jessup, and Tz. V. Kolev. A simple strategy for varying the restart parameter in GMRES(m). *Journal of Computational and Applied Mathematics*, 230(2):751–761, 2009.

[6] A. H. Baker, E. R. Jessup, and T. Manteuffel. A Technique for Accelerating the Convergence of Restarted GMRES. *SIAM Journal on Matrix Analysis and Applications*, 26(4):962–984, 2005.

[7] A. Bayliss, C. I. Goldstein, and E. Turkel. An iterative method for the Helmholtz equation. *Journal of Computational Physics*, 49(3):443–457, 1983.

[8] B. Beckermann, S. A. Goreinov, and E. E. Tyrtyshnikov. Some remarks on the Elman estimate for GMRES. *SIAM Journal on Matrix Analysis and Applications*, 27(3):772–778, 2005.

[9] B. Beckermann and A. B. J. Kuijlaars. Superlinear convergence of conjugate gradients. *SIAM Journal on Numerical Analysis*, 39(1):300–329, 2001.

[10] M. Benzi, G. H. Golub, and J. Liesen. Numerical solution of saddle point problems. *Acta Numerica*, 14:1–137, 2005.

[11] M. W. Berry, D. Mezher, B. Philippe, and A. Sameh. *Parallel Algorithms for the Singular Value Decomposition, in: Handbook for Parallel Computing and Statistics, edited by Erricos John Kontoghiorghes*, pages 117–164. Chapman & Hall/CRC, 2006.

[12] A. Boriçi. A Lanczos Approach to the Inverse Square Root of a Large and Sparse Matrix. *Journal of Computational Physics*, 162(1):123–131, 2000.

[13] J. H. Bramble and X. Zhang. The analysis of multigrid methods. In P.G. Ciarlet and J.L. Lions, editors, *Solution of Equation in $\mathbb{R}^n$ (Part 3), Techniques of Scientific Computing (Part 3)*, volume 7 of *Handbook of Numerical Analysis*, pages 173–415. Elsevier, 2000.

[14] W. L. Briggs, V. E. Henson, and S. F. McCormick. *A Multigrid Tutorial*. Society for Industrial and Applied Mathematics, 2nd edition, 2000.

[15] J. R. Bunch and B. N. Parlett. Direct Methods for Solving Symmetric Indefinite Systems of Linear Equations. *SIAM Journal on Numerical Analysis*, 8(4):639–655, 1971.

[16] D. Calvetti and L. Reichel. An adaptive Richardson iteration method for indefinite linear systems. *Numerical Algorithms*, 12:125–149, 1996.

[17] S. Chandrasekaran and I. C. F. Ipsen. On the sensitivity of solution components in linear systems of equations. *SIAM Journal on Matrix Analysis and Applications*, 16(1):93–112, 1995.

[18] J. Demmel and W. Kahan. Accurate Singular Values of Bidiagonal Matrices. *SIAM Journal on Scientific and Statistical Computing*, 11(5):873–912, 1990.

[19] Z. Drmač and K. Veselić. New Fast and Accurate Jacobi SVD Algorithm. I. *SIAM Journal on Matrix Analysis and Applications*, 29(4):1322–1342, 2008.

[20] Z. Drmač and K. Veselić. New Fast and Accurate Jacobi SVD Algorithm. II. *SIAM Journal on Matrix Analysis and Applications*, 29(4):1343–1362, 2008.

[21] M. Eiermann. Fields of values and iterative methods. *Numerical Linear Algebra with Applications*, 180:167–197, 1993.

[22] S. C. Eisenstat, H. C. Elman, and M. H. Schultz. Variational iterative methods for nonsymmetric systems of linear equations. *SIAM Journal on Numerical Analysis*, 20:345–357, 1983.

[23] H. C. Elman. *Iterative methods for large sparse nonsymmetric systems of linear equations.* PhD thesis, Yale University: New Haven, CT, 1982.

[24] M. Embree. How descriptive are GMRES convergence bounds? Technical Report 99/08, Oxford University Computing Laboratory, 1999.

[25] M. Embree. The tortoise and the hare restart GMRES. *SIAM Review*, 45(2):259–266, 2003.

[26] J. Erhel, K. Burrage, and B. Pohl. Restarted GMRES preconditioned by deflation. *Journal of Computational and Applied Mathematics*, 69(2):303–318, 1996.

[27] Y. A. Erlangga, C. Vuik, and C. W. Oosterlee. On a class of preconditioners for solving the Helmholtz equation. *Applied Numerical Mathematics*, 50(3-4):409–425, 2004.

[28] K. V. Fernando and B. N. Parlett. Accurate singular values and differential qd algorithms. *Numerische Mathematik*, 67:191–229, 1994.

[29] P. E. Gill, W. Murray, D. B. Ponceleón, and M. A. Saunders. Preconditioners for indefinite systems arising in optimization. *SIAM Journal on Matrix Analysis and Applications*, 13(1):292–311, 1992.

[30] S.K. Godunov and V.S. Ryabenkii. *Difference Schemes: An Introduction to the Underlying Theory.* Elsevier, 1987.

[31] G. Golub and W. Kahan. Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial and Applied Mathematics*, 2(2):205–224, 1965.

[32] G. H. Golub and C. F. Van Loan. *Matrix Computations.* The Johns Hopkins University Press, 3d edition, 1996.

[33] A. Greenbaum. *Iterative Methods for Solving Linear Systems.* SIAM, 1997.

[34] A. Greenbaum, V. Pták, and Z. Strakoš. Any nonincreasing convergence curve is possible for GMRES. *SIAM Journal on Matrix Analysis and Applications*, 17(3):465–469, 1996.

[35] A. Greenbaum and Z. Strakoš. Matrices that generate the same Krylov residual spaces. In G. Golub, A. Greenbaum, and M. Luskin, editors, *Recent Advances in Iterative Methods*, pages 95–118. Springer, 1994.

[36] A. Greenbaum and L. N. Trefethen. GMRES/CR and Arnoldi/Lanczos as matrix approximation problems. *SIAM Journal on Scientific Computing*, 15(2):359–368, 1994.

[37] V. Hernández, J. E. Román, and A. Tomás. A Robust and Efficient Parallel SVD Solver Bsed on Restarted Lanczos Bidiagonalization. *Electronic Transactions on Numerical Analysis*, 31:68–85, 2008.

[38] U. Hetmaniuk and R. Lehoucq. Basis selection in LOBPCG. *Journal of Computational Physics*, 218(1):324–332, 2006.

[39] N. J. Higham. *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008.

[40] M. E. Hochstenbach. A Jacobi–Davidson Type SVD Method. *SIAM J. Sci. Comput.*, 23(2):606–628, 2001.

[41] M. E. Hochstenbach. Harmonic and Refined Extraction Methods for the Singular Value Problem, with Applications in Least Squares Problems. *BIT Numerical Mathematics*, 44:721–754, 2004.

[42] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

[43] I. C.F. Ipsen. Expressions and bounds for the GMRES residual. *BIT*, 40(3):524–535, 2000.

[44] Z. Jia. Refined iterative algorithms based on Arnoldi's process for large unsymmetric eigenproblems. *Linear Algebra and its Applications*, 259:1–23, 1997.

[45] Z. Jia and D. Niu. An Implicitly Restarted Refined Bidiagonalization Lanczos Method for Computing a Partial Singular Value Decomposition. *SIAM Journal on Matrix Analysis and Applications*, 25(1):246–265, 2003.

[46] W. Joubert. On the convergence behavior of the restarted GMRES algorithm for solving nonsymmetric linear systems. *Numerical Linear Algebra with Applications*, 1:427–447, 1994.

[47] A. V. Knyazev. *Computation of eigenvalues and eigenvectors for mesh problems: algorithms and error estimates.* Dept. Numerical Math. USSR Academy of Sciences, Moscow, 1986. (In Russian).

[48] A. V. Knyazev. Toward the Optimal Preconditioned Eigensolver: Locally Optimal Block Preconditioned Conjugate Gradient Method. *SIAM Journal on Scientific Computing*, 23(2):517–541, 2001.

[49] E. Kokiopoulou, C. Bekas, and E. Gallopoulos. Computing smallest singular triplets with implicitly restarted lanczos bidiagonalization. *Applied Numerical Mathematics*, 49(1):39–61, 2004.

[50] A. L. Laird and M. B. Giles. Preconditioned iterative solution of the 2D Helmholtz equation. Technical Report 02/12, Oxford University Computing Laboratory, Oxford, UK, 2002.

[51] V. I. Lebedev. Iterative methods for solving operator equations with spectrum contained in several intervals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 9(6):17–24, 1969. English transl. in USSR Comput. Math. Math. Phys. 9 (1972).

[52] N. M. Nachtigal, S. C. Reddy, and L. N. Trefethen. How fast are nonsymmetric matrix iterations? *SIAM Journal on Matrix Analysis and Applications*, 13(3):778–795, 1992.

[53] J. Nocedal and S. Wright. *Numerical Optimization.* Springer Series in Operations Research. Springer, 1999.

[54] E.E. Ovtchinnikov. Computing several eigenpairs of Hermitian problems by conjugate gradient iterations. *Journal of Computational Physics*, 227(22):9477–9497, 2008.

[55] C. C. Paige and M. A. Saunders. Solution of Sparse Indefinite Systems of Linear Equations. *SIAM Journal on Numerical Analysis*, 12(4):617–629, 1975.

[56] B. N. Parlett. *The symmetric eigenvalue problem*, volume 20 of *Classics in Applied Mathematics.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1998. Corrected reprint of the 1980 original.

[57] B. N. Parlett and O. A. Marques. An implementation of the dqds algorithm (positive case). *Linear Algebra and its Applications*, 309(1-3):217–259, 2000.

[58] Y. Saad. Iterative Solution of Indefinite Symmetric Linear Systems by Methods Using Orthogonal Polynomials over Two Disjoint Intervals. *SIAM Journal on Numerical Analysis*, 20(4):784–811, 1983.

[59] Y. Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, 2003.

[60] Y. Saad, J. R. Chelikowsky, and S. M. Shontz. Numerical methods for electronic structure calculations of materials. *SIAM Review*, 52(1):3–54, 2010.

[61] Y. Saad and M. H. Schultz. GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems. *SIAM Journal on Scientific and Statistical Computing*, 7(3):856–869, 1986.

[62] V. Simoncini and D. Szyld. On the occurrence of superlinear convergence of exact and inexact Krylov subspace methods. *SIAM Review*, 47:247–272, 2005.

[63] V. Simoncini and D. Szyld. New conditions for non-stagnation of minimal residual methods. *Numerische Mathematik*, 109(3):477–487, 2008.

[64] G. L. G. Sleijpen and H. A. Van der Vorst. A Jacobi–Davidson Iteration Method for Linear Eigenvalue Problems. *SIAM Journal on Matrix Analysis and Applications*, 17(2):401–425, 1996.

[65] G. W. Stewart. Collinearity and least squares regression. *Statistical Science*, 2(1):68–84, 1987.

[66] G. W. Stewart and J.-G. Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

[67] O. Tatebe. The multigrid preconditioned conjugate gradient method. In N. D. Melson, T. A. Manteuffel, and S. F. McCormick, editors, *Sixth Copper Mountain Conference on Multigrid Methods*, volume NASA Conference Publication 3224, pages 621–634, 1993.

[68] M. P. Teter, M. C. Payne, and D. C. Allan. Solution of Schrödinger's equation for large systems. *Physical Review B*, 40(18):12255–12263, 1989.

[69] A. N. Tikhonov and A. A. Samarskii. *Uravneniya matematicheskoi fiziki (Equations of Mathematical Physics)*. Nauka, Moscow, 1977. In Russian.

[70] K.-C. Toh. GMRES vs. Ideal GMRES. *SIAM Journal on Matrix Analysis and Applications*, 18(1):30–36, 1997.

[71] S. Tomov, J. Langou, J. Dongarra, A. Canning, and L.-W. Wang. Conjugate-gradient eigenvalue solvers in computing electronic properties of nanostructure architectures. *International Journal of Computational Science and Engineering*, 2(3/4):205–212, 2008.

[72] L. N. Trefethen. Approximation theory and numerical linear algebra. In J. Mason and M. Cox, editors, *Algorithms for Approximation II*. Chapman and Hall, London, U.K., 1990.

[73] U. Trottenberg, C. W. Oosterlee, and A. Schüller. *Multigrid*. Academic Press, 2001.

[74] H. A. van der Vorst and C.Vuik. The superlinear convergence behaviour of GMRES. *Journal of Computational and Applied Mathematics*, 48:327–341, 1993.

[75] M. B. van Gijzen, Y. A. Erlangga, and C. Vuik. Spectral Analysis of the Discrete Helmholtz Operator Preconditioned with a Shifted Laplacian. *SIAM Journal on Scientific Computing*, 29(5):1942–1958, 2007.

[76] E. Vecharynski and J. Langou. Any admissible cycle-convergence behavior is possible for restarted GMRES at its initial cycles. Numerical Linear Algebra with Applications, doi:10.1002/nla.739, 2010.

[77] E. Vecharynski and J. Langou. The Cycle-Convergence of Restarted GMRES for Normal Matrices Is Sublinear. *SIAM Journal on Scientific Computing*, 32(1):186–196, 2010.

[78] D. M. Young and K. C. Jea. Generalized conjugate-gradient acceleration of nonsymmetrizable iterative methods. *Linear Algebra and its Applications*, 34:159–194, 1980.

[79] I. Zavorin. Spectral factorization of the Krylov matrix and convergence of GMRES. Technical report, University of Maryland Computer Science Department, http://hdl.handle.net/1903/1168, 2002.

[80] I. Zavorin, D. P. O'Leary, and H. Elman. Complete stagnation of GMRES. *Linear Algebra and its Applications*, 367:165–183, 2003.

[81] B. Zhong and R. B. Morgan. Complementary cycles of restarted GMRES. *Numerical Linear Algebra with Applications*, 15(6):559–571, 2008.

[82] J. Zítko. Generalization of convergence conditions for a restarted GMRES. *Numerical Linear Algebra with Applications*, 7(3):117–131, 2000.