Charles University
Faculty of Mathematics and Physics

DOCTORAL THESIS

Iveta Hnětynková

Krylov subspace approximations in linear algebraic problems

Department of Numerical Mathematics
Supervisor: *Doc. RNDr. Jan Zítko, CSc.*

Title: Krylov subspace approximations in linear algebraic problems
Author: Iveta Hnětynková
Department: Department of Numerical Mathematics
Supervisor: Doc. RNDr. Jan Zítko, CSc.
Author's e-mail address: iveta.hnetynkova@gmail.com
Supervisor's e-mail address: zitko@karlin.mff.cuni.cz

Abstract: The GMRES(m) method for solving linear systems $\mathbf{A}x = b$, i.e. restarted GMRES with restart parameter m, is attractive when a good preconditioner is available. The determining of some types of preconditioners is connected with construction of an $\mathbf{A}$–invariant subspace corresponding to eigenvalues closest to zero. One class of methods for computation of invariant subspaces is based on the construction of polynomial filters. In the first part of this thesis, we study using of Tchebychev polynomials for constructing suitable filters and compare them with classical ones proposed by D. C. Sorensen. Convergence of the presented algorithm is studied and also the case where geometrical multiplicity of a small eigenvalue of $\mathbf{A}$ is greater than one is analyzed. Numerical results assessing the quality of polynomial filters and preconditioners are presented.

In the second part of this thesis, an orthogonally invariant linear approximation problem $\mathbf{A}x \approx b$ is considered. C.C. Paige and Z. Strakoš proved that the (partial) upper bidiagonalization of the matrix $[b, \mathbf{A}]$ determines a core approximation problem $\mathbf{A}_{11}x_1 \approx b_1$, with all necessary and sufficient information for solving the original problem given by b_1 and $\mathbf{A}_{11}$. Here we derive the core problem formulation from the relationship between the Golub-Kahan bidiagonalization and the Lanczos tridiagonalization, and from the known properties of Jacobi matrices. We discussed how the presented relationship may be found useful in applications of the core problem formulation, especially in regularization methods for solving large ill-posed problems. Possible directions for further research are outlined and several open questions are formulated.

Keywords: Krylov methods, Arnoldi factorization, restarted GMRES, invariant subspaces, ill-posed problem, core problem, Golub-Kahan bidiagonalization.

Acknowledgments

I would like to thank all those who supported me in my doctoral study and the work on my thesis. I very appreciate the help and counsels received from my advisor Jan Zítko and I am grateful for numerous remarks, corrections and comments he gave me. I would like to thank particularly Zdeněk Strakoš for the help he provided me with the second part of this thesis. The former as well as Josef Málek and Miloslav Feistauer also enabled me to participate at various mathematical activities.

I am much obligated to Petr Tichý for useful remarks concerning the early version of this thesis. For the various help they provided me, I also thank my colleagues, particularly Martin Plešinger and Jurjen Duintjer Tebbens.

Last but not least, I am in debt to my husband and parents, whose support and patience made this work possible.

My thanks also go to institutions that provided financial support for my research work. Through my doctoral study, my work was partially supported by the Grant GAUK 6/2005/R and by the National Program of Research ”Information Society” under project 1ET400300415.

Contents

Part II

Notation

$\mathbb{N},\ \mathbb{R},\ \mathbb{C}$	the set of all natural, real, complex numbers
$\mathbf{A} = (a_{i,j})_{i,j=1}^{n,m}$	the square (if $n = m$) or rectangular (if $n \neq m$) matrix
$\mathbf{I}_k,\ \mathbf{I}$	the k by k identity matrix - if there is no doubt we omit the index
$\mathbf{0}_{k,l},\ \mathbf{0}$	the k by l zero matrix - if there is no doubt we omit the indexes
$e_i^{(k)},\ e_i$	the ith column of $\mathbf{I}_k$, $\mathbf{I}$
$(x)_i$	the ith component of the vector x
$\mathbf{A}^{-1}$	the inverse matrix
$\mathbf{A}^{+}$	the Moore-Penrose pseudoinverse matrix
$\mathbf{A}^T,\ x^T$	the transposed matrix, vector
$\mathbf{A}^H,\ x^H$	the transposed and complex conjugated matrix, vector
$\mathrm{diag}(\alpha, \beta, \ldots)$	the square diagonal matrix with $\alpha, \beta, \ldots$ on the diagonal
$\mathrm{diag}(\mathbf{B}_1, \mathbf{B}_2, \ldots)$	the square block diagonal matrix
$\bar{\alpha}$	the complex conjugated value for $\alpha \in \mathbb{C}$
$\mathcal{U}_\epsilon(\alpha)$	the ϵ-neighborhood of $\alpha \in \mathbb{C}$
$\|x\|$	the Euclidean vector norm $\|x\| = \sqrt{x^T x}$
$\|\mathbf{A}\|$	the spectral matrix norm $\|A\| = \sup_{x \neq 0} \frac{\|\mathbf{A}x\|}{\|x\|}$
$\|\mathbf{A}\|_F$	the Frobenius matrix norm $\|A\|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij}^2}$
$\mathcal{R}(A)$	the range of the matrix $\mathbf{A}$
$\sigma(\mathbf{A})$	the spectrum of the matrix $\mathbf{A}$
$\rho(\mathbf{A})$	the spectral radius of the matrix $\mathbf{A}$
P_k	set of all polynomials of degree at most k
MP_k	set of all monic polynomials of degree at most k

Introduction

Numerical methods for solving real-world problems often lead to large systems of linear algebraic equations and iterative methods, especially Krylov subspace methods, are typically used to solve them. In the first part of this thesis, we concentrate on accelerating convergence of one of the most widely used Krylov method for solving large, sparse, nonsymmetric linear systems - GMRES method [75]. In the second part, we analyze the core concept [64] in linear approximation problems that leads to better understanding of several well known iterative techniques used to solve these problems.

In Part I, we consider the linear system

$$\mathbf{A}x = b, \quad \mathbf{A} \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^n,$$

with a large, sparse and nonsingular matrix $\mathbf{A}$. The restarted GMRES method is often used to solve such system. It has been observed that restart slows down the convergence, stagnation may occur in many cases and it is advisable to use preconditioning to overcome these difficulties. If the matrix $\mathbf{A}$ is symmetric, convergence is strongly connected with the distribution of its eigenvalues. Eigenvalues close to zero may cause the troubles and thus many preconditioners are based on the idea to remove them from the spectrum. If $\mathbf{A}$ is nonsymmetric, the situation is more complicated and the convergence behavior is still not well understood. For matrices close to normal, i.e. matrices having well conditioned set of eigenvectors, ideas similar to the symmetric case can often be used. Though, note that range of applications can be found, where the matrix is strongly nonnormal and the convergence does not depend on the distribution of its eigenvalues, see, e.g., [50] where the convergence of the GMRES method for convection-diffusion model problem is analyzed, and the dependence of convergence on particular right-hand side and boundary condition is proved. Thus analysis of the spectrum of $\mathbf{A}$ is only one of possible approaches, that can offer several options for acceleration of convergence in some applications.

In this thesis, we concentrate on preconditioners of the restarted GMRES method that can be constructed by exploiting spectral information of $\mathbf{A}$ gathered in the previous restart. Usually a rectangular matrix $\mathbf{V}$ is constructed whose columns generate an invariant subspace of $\mathbf{A}$ corresponding to the smallest eigenvalues. There are several ways to use such matrix $\mathbf{V}$. The first possibility is to construct a preconditioner which is relatively inexpensively updated after each restart of GMRES, so called adaptive preconditioner, see [3], [12], [13], [21], [23]. Another possibility is to augment the Krylov subspace by the columns of the matrix $\mathbf{V}$, see [14], [23], [58], [57], [56], [55]. The quality of preconditioner depends strongly on the matrix $\mathbf{V}$ and thus the goal of the first part of this thesis is to analyze and improve procedures for computation of invariant subspaces.

Implicitly restarted Arnoldi process with shifts (IRA) proposed by D. C. Sorensen [81] is an attractive technique for computing a few eigenvalues and the associated eigenspace of a general matrix. This technique iteratively updates the starting vector v_1 of the Arnoldi process such that it produces an upper Hessenberg matrix of order $k \ll n$ with the property that its eigenvalues are good approximations to searched eigenvalues of the original matrix $\mathbf{A}$. Moreover, the Krylov subspace $\mathcal{K}_k(\mathbf{A}, v_1)$ approximates well the corresponding eigenspace. The vector v_1 is updated by a polynomial called polynomial filter which filters out selected part of the spectrum. Here we work out the idea to use Tchebychev polynomials for constructing polynomial filters. In the new approach the roots of transformed and scaled Tchebychev polynomials are used for the shifts in the IRA process. Then an analysis of convergence of constructed approximations to the searched invariant subspace is presented. This convergence was studied in [81], [49] for some special cases. We generalize some of these results.

In practice, usually an unreduced and therefore nonderogatory Hessenberg matrix is obtained. The question arises what happens if the smallest eigenvalue of $\mathbf{A}$, say λ, has geometric multiplicity greater than one. Here we prove that when the Jordan canonical form of the matrix $\mathbf{A}$ has $s > 1$ blocks corresponding to λ with maximal dimension of the block d, then an integer j exists such that the vectors $v_1^{(j)}, \mathbf{A}v_1^{(j)}, \ldots, \mathbf{A}^d v_1^{(j)}$ are almost linearly dependent, where $v_1^{(j)}$ is the jth update of the starting vector of the Arnoldi process. Hence the Arnoldi process stops after at most $(d+1)$ steps and thus it is important to modify the stopping criterion for determining an invariant subspace.

We compare the classical IRA process and the new technique numerically and show that the new approach leads to improvement of numerical accuracy and convergence properties.

In Part II, we consider the (possibly incompatible) linear approximation problem

$$\mathbf{A}x \approx b, \quad \mathbf{A} \in \mathbb{R}^{n \times m}, b \in \mathbb{R}^n,$$

with a nonzero matrix $\mathbf{A}$ and a nonzero vector b. Such systems arise in many scientific and technical areas and various techniques are used to solve them. When the matrix representing the model in the approximation problem is large, which is often the case, we need to consider iterative methods (e.g., Krylov subspace methods) with some appropriate stopping criteria.

A new contribution in theory and computation of linear approximation problems was published in a series of papers [64], [65], [67]. Here the authors define a core problem within an unitarily invariant linear approximation problem. It is proposed to orthogonally transform the original approximation problem to the block form that allows to separate the necessary and sufficient information present in the data $b, \mathbf{A}$ from the redundancies. It is shown that the core reduction represents a theoretical basis for several well known techniques as well as for new future developments. After reviewing the basic concept of core reduction, we present in this thesis alternative proofs of its fundamental characteristics based on the relationship between the Golub-Kahan bidiagonalization, the Lanczos tridiagonalization and the well known properties of Jacobi matrices. They were published in the paper [42]. Here the Golub-Kahan bidiagonalization and the Lanczos tridiagonalization are used not as computational methods, but as well suited mathematical tools for constructing proofs.

Then we discuss how the presented relationship may be found useful in applications of the core problem formulation, in particular in connection with regularization of ill-posed problems and investigation of efficient stopping criteria. Possible directions for further research are outlined and several open questions are formulated.

The structure of this thesis is as follows. Chapter 1 contains basic definitions and notation, followed by an overview of some adaptive preconditioning techniques. Finally, the properties of Arnoldi factorization are discussed. In Chapter 2, the IRA process is briefly reminded and a new technique based on the properties of Tchebychev polynomials is described. Algorithm for computation of invariant subspace is presented. Chapter 3 contains convergence analysis of discussed algorithms including the case that the matrix $\mathbf{A}$ is derogatory and/or defective. In Chapter 4, numerical results are presented comparing the algorithms for construction of invariant subspaces and preconditioning techniques. Chapter 5 summarizes main types of linear approximation problems and methods usually used for solving them, including iterative methods. Chapter 6 describes the concept of the core reduction, presents alternative proofs of its fundamental properties and outlines some directions for further research.

Part I

Approximation of invariant subspaces and convergence of GMRES

Chapter 1

Basic concepts

In this chapter we point out main advantages of iterative methods compared with other strategies in context of solving large, sparse, linear algebraic problems. We concentrate on the restarted GMRES method, especially on adaptive preconditioning techniques for this method that are based on the idea to remove smallest in magnitude eigenvalues from the spectrum of the matrix $\mathbf{A}$. *These preconditioners are constructed by exploiting spectral information of* $\mathbf{A}$ *gathered from the Arnoldi factorization obtained in the previous restart. Thus we remind two orthogonalization processes – Arnoldi process and Householder process, and summarize the fundamental properties of the Arnoldi factorization. The question when Ritz (respectively harmonic Ritz) values and vectors approximate the eigenvalues and the eigenvectors of* $\mathbf{A}$ *is discussed.*

1.1 Iterative methods

Consider the system of linear algebraic equations

$$\mathbf{A}x = b, \quad \mathbf{A} \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^n, \tag{1.1}$$

with a large, sparse, nonsingular and nonsymmetric matrix $\mathbf{A}$. We consider only real problems, but extension to complex ones is straightforward.

Systems of the form (1.1) arise in many scientific and technical areas. For example, a part of reality can be modeled by a system of differential or integral equations and their discretization leads to linear problem that can be characterized by the system of linear algebraic equations. Two types of methods can be used to solve these systems - direct and iterative. Most *direct methods* consists of performing some type of factorization of the matrix $\mathbf{A}$, as LU factorization, Cholesky factorization etc., followed by calculation of inverse of the matrix $\mathbf{A}$ which is applied to the vector b to obtain the solution. Direct method finds the solution at the very end of the computation and intermediate results do not represent any approximate solution. If the problem (1.1) arises from some discretized model, it is influenced by modeling and discretization errors. Errors on all stages of solving the real-word problems should be in balance and thus also the approximate solution of (1.1) suffices in accuracy which needn't be necessary large. Therefore *iterative methods* are often suitable for solving (1.1), because it is possible to stop the process at any iteration step and an approximation of the solution is obtained.

Having an initial approximation $x_0 \in \mathbb{R}^n$ of the exact solution $x^* \in \mathbb{R}^n$, iterative method computes in step k an approximation x_k of x^*. Natural requirement is that x_k converges to the solution x^* as k grows. Therefore we can define the residual of the approximation x_k by

$$r_k := b - \mathbf{A}x_k$$

and the iterative process is stopped as soon as the norm of the residual is smaller then the prescribed number, i.e. the required accuracy is reached. More frequently relative residual norm

$$\|r_k\|/\|r_0\|,$$

where r_0 is the initial residual, is tested because it gives better information about the improvement of the approximation. Recently also the relative backward error

$$\|r_k\|/(\|b\| + \|\mathbf{A}\|\|x_k\|),$$

has been preferred, see [66].

Another advantage of iterative methods is that the sparse structure of the matrix $\mathbf{A}$ allows to solve large system of millions of unknowns without transforming the whole system matrix or even without explicitly forming it, because usually only the matrix-vector product must be available.

In practice, direct and iterative methods are often combined in order to benefit from advantages of both approaches, e.g. incomplete factorizations can be used to precondition the system (1.1).

One of the most popular class of iterative methods form *Krylov subspace methods* - projective methods which seek an approximation

$$x_k \in x_0 + \mathcal{K}_k(\mathbf{A}, r_0) = x_0 + \text{ span}\{r_0, \mathbf{A}r_0, ..., \mathbf{A}^{k-1}r_0\}$$

satisfying Petroff-Galerkin condition

$$b - \mathbf{A}x_k \perp \mathcal{L}_k$$

for a subspace $\mathcal{L}_k$ of dimension k. Different methods can be obtained by choosing different $\mathcal{L}_k$, e.g., $\mathcal{L}_k = \mathcal{K}_k(\mathbf{A}, r_0)$ gives Full Orthogonalization Method (FOM),

$\mathcal{L}_k = \mathbf{A}\mathcal{K}_k(\mathbf{A}, r_0)$ gives Generalized Minimum Residual Method (GMRES) etc. Very nice overview of Krylov methods including detailed algorithms can be found in [73] Chapter 6,7, see also [89], [22], [31].

In this thesis, we concentrate on one of the most widely used Krylov method for solving large, sparse, nonsymmetric linear systems - the GMRES method. GMRES is a method with long recurrences, one more vector has to be saved in each iteration and memory requirements grow with number of iterations. Thus the classical GMRES algorithm is not suitable for practical computations and restart must be used. Unfortunately, convergence of the restarted GMRES method can be very slow, stagnation may occur in many cases and process can become inapplicable for practical use. These problems can often be removed by using preconditioning which might improve properties of the original system. Adaptive preconditioning techniques, that will be discussed in this thesis, are presented in Section 1.4.

Before we turn to the GMRES method, we remind two main algorithms for construction of appropriate basis of the Krylov subspace.

1.2 Building orthonormal basis

Denote by

$$\mathcal{K}_k(\mathbf{A}, v) \;=\; \text{span}\,\{v, \mathbf{A}v, ..., \mathbf{A}^{k-1}v\}$$

the kth Krylov subspace corresponding to the matrix $\mathbf{A}$ and the vector v, $k \in \mathbb{N}$. It is well known that Krylov basis $\{v, \mathbf{A}v, ..., \mathbf{A}^{k-1}v\}$ of this subspace is numerically unstable, because as k grows the vectors $\mathbf{A}^k v$ can become nearly linearly dependent. This can lead to loss of accuracy in Krylov subspace methods, e.g., GMRES, FOM, DIOM, QGMRES (see [73]). Thus orthonormal basis of the Krylov subspace is usually constructed by Arnoldi process or its modified version, or by Householder process. In this section, algorithms of the modified Arnoldi process and the Householder process are reminded. For comparison of these algorithms see, e.g., [19], [87].

Modified Arnoldi process

Let v be a starting vector and $\mathcal{K}_k(\mathbf{A}, v)$ required Krylov subspace. Then the algorithm of the modified Arnoldi process is the following:

ALGORITHM 1.1 (Modified Arnoldi process):
input: $v, \mathbf{A}$, k - dimension of Krylov subspace
output: $v_1, .., v_k$ - orthonormal basis of $\mathcal{K}_k(\mathbf{A}, v)$

```
$v_1 := v/\|v\|$
for $j = 1, .., k-1$
        $w_j := \mathbf{A}v_j$
        for $i = 1, .., j$
                $h_{ij} := v_i^T w_j$
                $w_j := w_j - h_{ij}v_i$
        end $i$
        $h_{j+1,j} := \|w_j\|$
        if $h_{j+1,j} = 0$ then STOP
        $v_{j+1} := w_j / h_{j+1,j}$
end $j$
```

At each iteration the vector v_j is multiplied by the matrix $\mathbf{A}$ and orthogonalized against the previous vectors using modified *Gramm-Schmidt orthogonalization* process. The algorithm stops as soon as k basis vectors of $\mathcal{K}_k(\mathbf{A}, v)$ are computed, or $h_{j+1,j} = \|w_j\| = 0$ for some j. The second situation appears if the vectors in Krylov basis are linearly dependent, i.e. the subspace does not have full dimension. Consider $h_{j+1,j} \neq 0$ for $j = 1, \dots, k$ and denote by $\mathbf{V}_k := (v_1, .., v_k)$,

$$\mathbf{H}_k := \begin{pmatrix} h_{11} & h_{12} & ... & h_{1,k-1} & h_{1k} \\ h_{21} & h_{22} & ... & h_{2,k-1} & h_{2k} \\ 0 & h_{32} & ... & h_{3,k-1} & h_{3k} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & h_{k,k-1} & h_{k,k} \end{pmatrix}, \ \mathbf{H}_{k+1,k} := \begin{pmatrix} \mathbf{H}_k \\ h_{k+1,k}e_k^T \end{pmatrix}.$$

Now we formulate a well known theorem.

Theorem 1.1: *Assume that Algorithm 1.1 does not stop before step k. Then it generates the vectors $v_1, v_2, \dots, v_k$ such that $v_i^T v_j = 0$ for $i \neq j$ and*

$$span\{v_1, ..., v_k\} = \mathcal{K}_k(\mathbf{A}, v).$$

Moreover

$$\begin{aligned} \mathbf{AV}_k &= \mathbf{V}_{k+1}\mathbf{H}_{k+1,k} && (1.2) \\ &= \mathbf{V}_k\mathbf{H}_k + w_k e_k^T, \quad \textit{where} \ \ w_k^T\mathbf{V}_k = 0. && (1.3) \end{aligned}$$

Proof: Follows immediately by induction, see [73] pp. 147. □

The basis $\{v_1, v_2, \ldots, v_k\}$ is called Arnoldi basis and the factorization (1.3) is called *Arnoldi factorization* with starting vector v_1. Its important properties are discussed in Section 1.5. Algorithm 1.1 is in exact arithmetics equivalent with classical Arnoldi algorithm (see [73] p.146), but modified version is numerically more stable. Nevertheless, in many cases loss of orthogonality in Arnoldi basis appears, see [8], [10], [27], and improvement can be obtained by using reorthogonalization which can be, on the other hand, computationally expensive. Another possibility is to use the following approach.

Householder process

Different approach to construction of orthonormal basis is used by Householder method, that is based on application of *reflection matrices* (sometimes also called elementary Hermitian matrices)

$$\mathbf{P} = \mathbf{I} - 2ww^T,$$

where $\|w\| = 1$. Such matrices are Hermitian, unitary and det($\mathbf{P}$)= - 1. From the properties of $\mathbf{P}$ it can be proved that a vector $\mathbf{P}v$ is a mirror reflection of the vector v with respect to span$\{w\}^\perp$, i.e. $\|\mathbf{P}v\| = \|v\|$ and $(v - \mathbf{P}v)$ is orthogonal to span$\{w\}^\perp$. The following algorithm can be found, e.g., in [73] p. 149.

ALGORITHM 1.2 (Householder process):
input: $v, \mathbf{A}$, k - dimension of Krylov subspace
output: $v_1, .., v_k$ - orthonormal basis of $\mathcal{K}_k(\mathbf{A}, v)$

```
$z_1 := v/\|v\|$
for $j = 1, .., k$
    $\hat{\beta} := -\ sign((z_j)_j)\ (\sum_{i=j}^{n} (z_j)_i^2)^{1/2}$
    for $i = 1, .., j-1$
        $(\hat{w}_j)_i := 0$
    end $i$
    $(\hat{w}_j)_j := (z_j)_j - \hat{\beta}$
    for $i = j+1, .., n$
        $(\hat{w}_j)_i := (z_j)_i$
    end $i$
    $w_j := \frac{\hat{w}_j}{||\hat{w}_j||}$
    $\mathbf{P}_j := \mathbf{I} - 2w_j w_j^T$
    $h_{j-1} := \mathbf{P}_j z_j$
    $v_j := \mathbf{P}_1..\mathbf{P}_j e_j$
    if $||v_j|| = 0$ then STOP
    if $j \leq k-1$ then $z_{j+1} := \mathbf{P}_j..\mathbf{P}_1 \mathbf{A} v_j$
end $j$
```

Remark: The Householder matrices $\mathbf{P}_j$ need not be formed explicitly. Multiplication of $\mathbf{P}_j$ by a vector x can be performed by

$$\mathbf{P}_j x = (\mathbf{I} - 2w_j w_j^T)x = x - 2\zeta w_j, \text{ where } \zeta := w_j^T x.$$

This multiplication needs only about $4 * (n - j + 1)$ operations if we use the fact that $(w_j)_i = 0$ for $i = 1, \dots, j-1$.

Algorithms 1.1 and 1.2 are equivalent in exact arithmetics, but the Householder algorithm is more time consuming and retains orthogonality between the basis vectors better than the modified Arnoldi process. The following table presented in [73] pp. 151 shows estimations of memory requirements and number of flops needed for different orthogonalization processes under the assumption that all algorithms do not stop before step k. ARN denotes classical Arnoldi algorithm, ModARN and ReARN denotes modified Arnoldi process (see Algorithm 1.1) and modified Arnoldi process with reorthogonalization in each step, respectively. Finally, HHA is Householder process presented in Algorithm 1.2.

	ARN	**ModARN**	**ReARN**	**HHA**
Flops	$2k^2n$	$2k^2n$	$4k^2n$	$4k^2n - 4/3k^3$
Storage	$(k+1)n$	$(k+1)n$	$(k+1)n$	$(k+1)n - 1/2k^2$

1.3 GMRES method

Generalized Minimum Residual Method (GMRES) proposed in [75] is very popular technique for solving problems defined in (1.1). As we have mentioned in Section 1.1, GMRES is a projection method with $\mathcal{L}_k = \mathbf{A}\mathcal{K}_k(\mathbf{A}, r_0)$, i.e. it computes the kth approximate solution such that

$$x_k \in x_0 + \mathcal{K}_k(\mathbf{A}, r_0) = x_0 + \text{ span}\{r_0, \mathbf{A}r_0, ..., \mathbf{A}^{k-1}r_0\}$$

and

$$b - \mathbf{A}x_k \perp \mathbf{A}\mathcal{K}_k(\mathbf{A}, r_0).$$

This is equivalent to

$$x_k = \underset{x \in x_0 + \mathcal{K}_k(\mathbf{A}, r_0)}{\arg\min} \|b - \mathbf{A}x\| = x_0 + \underset{u \in \mathcal{K}_k(\mathbf{A}, r_0)}{\arg\min} \|r_0 - \mathbf{A}u\| = x_0 + \mathbf{V}_k y_k, \tag{1.4}$$

where the columns of $\mathbf{V}_k$ form an orthonormal basis of $\mathcal{K}_k(\mathbf{A}, r_0)$, that can be computed by Algorithm 1.1 or 1.2 with the starting vector r_0. Then $v_1 = r_0/\beta$, where $\beta = \|r_0\|$ and, according to (1.4),

$$y_k = \underset{y \in R^k}{\arg\min} \|r_0 - \mathbf{A}\mathbf{V}_k y\|. \tag{1.5}$$

It follows from Theorem 1.1 that

$$r_0 - \mathbf{A}\mathbf{V}_k y = \mathbf{V}_{k+1}(\beta e_1 - \mathbf{H}_{k+1,k} y)$$

and the relation (1.5) results in

$$y_k = \underset{y \in R^k}{\arg\min} \|\beta e_1 - \mathbf{H}_{k+1,k} y\|. \tag{1.6}$$

This minimalization problem is easy to compute and requires solving least-squares problem with upper Hessenberg matrix $\mathbf{H}_{k+1,k} \in \mathbb{R}^{(k+1)\times k}$, where k is usually small in comparison with the dimension of the original linear system.

Problem (1.6) is usually solved by using QR–decomposition. The matrix $\mathbf{H}_{k+1,k}$ has full column rank and thus there exists an orthogonal matrix $\mathbf{Q}_{k+1} \in \mathbb{R}^{(k+1)\times(k+1)}$ and an upper triangular matrix $\mathbf{R}_{k+1,k} \in R^{(k+1)\times k}$ such that

$$\mathbf{H}_{k+1,k} = \mathbf{Q}_{k+1}\mathbf{R}_{k+1,k}.$$

Let us rewrite the norm in (1.6) in the form

$$||\beta e_1 - \mathbf{H}_{k+1,k} y\|^2 = \|\beta e_1 - \mathbf{Q}_{k+1}\mathbf{R}_{k+1,k} y\|^2 = \|g_k - \mathbf{R}_{k+1,k} y\|^2,$$

where $g_{k+1} := \beta \mathbf{Q}_{k+1}^T e_1$, $g_{k+1} \in R^{k+1}$. Matrix $\mathbf{R}_{k+1,k}$ is upper triangular with zero last row. Let $\mathbf{R}_k$ be a matrix $\mathbf{R}_{k+1,k}$ without the last row and $g_{k+1} = (\hat{g}_k^T, \eta_k)^T$, where $\eta_k \in \mathbb{R}$. Then

$$\|g_{k+1} - \mathbf{R}_{k+1,k} y\|^2 = \|\hat{g}_k - \mathbf{R}_k y\|^2 + |\eta_k|^2.$$

Substituting this into (1.6) gives together with (1.4) equivalent formulation for the kth GMRES iteration

$$y_k = \underset{y \in R^k}{\arg\min} \|\hat{g}_k - \mathbf{R}_k y\|, \quad x_k = x_0 + \mathbf{V}_k y_k. \tag{1.7}$$

Matrix $\mathbf{H}_{k+1,k}$ has full column rank. Therefore $\mathbf{R}_k$ is square and nonsingular and y_k satisfies equation

$$\mathbf{R}_k y = \hat{g}_k. \tag{1.8}$$

Summarizing, the problem (1.4) reduces to solving small system (1.8) with nonsingular upper triangular matrix. Moreover,

$$\|r_k\| = \|b - \mathbf{A}x_k\| = |\eta_k|.$$

Remark: The QR–decomposition of the matrix $\mathbf{H}_{k+1,k}$ can be computed by Givens rotations. Define the rotation matrices $\mathbf{G}_i^{k+1} \in \mathbb{R}^{(k+1)\times(k+1)}$,

$$\mathbf{G}_i^{k+1} := \begin{pmatrix} \mathbf{I}_{i-1} & 0 & 0 & \mathbf{0}_{i-1,k-i} \\ 0 & c_i & -s_i & 0 \\ 0 & s_i & c_i & 0 \\ \mathbf{0}_{k-i,i-1} & 0 & 0 & \mathbf{I}_{k-i} \end{pmatrix},$$

where

$$c_i := \frac{h_{ii}}{\sqrt{h_{ii}^2 + h_{i,i+1}^2}}, \quad s_i := \frac{-h_{i+1,i}}{\sqrt{h_{ii}^2 + h_{i,i+1}^2}}.$$

Then the matrix $\mathbf{G}_i^{k+1}$ eliminates the element $h_{i+1,i}$ of $\mathbf{H}_{k+1,k}$. Therefore multiplying $\mathbf{H}_{k+1,k}$ successively by matrices $\mathbf{G}_i^{k+1}$ for $i = 1, \ldots, k$ transforms the matrix to upper triangular form. Moreover, $\mathbf{Q}_{k+1}^T := \mathbf{G}_k^{k+1}\mathbf{G}_{k-1}^{k+1}...\mathbf{G}_1^{k+1}$ is unitary and

$$\mathbf{H}_{k+1,k} = \mathbf{Q}_{k+1}\mathbf{R}_{k+1,k},$$

where $\mathbf{R}_{k+1,k} := \mathbf{Q}_{k+1}^T\mathbf{H}_{k+1,k}$. The rotation matrices need not be computed for each k separately, because the iteration scheme for computation of $\mathbf{G}_i^{k+1}$ can be derived. More computational cost can be saved if we use some interesting properties of the rotation matrices. For example the GMRES residuals satisfy

$$\|r_{i+1}\| = -s_i\|r_i\|,$$

because $\eta_{i+1} = -s_i\eta_i$, see [73] pp. 160–163.

Remark: Implementation of the GMRES method using Householder orthogonalization process is described in [90], [73] p. 159. Numerical stability of GMRES with the Arnoldi and Householder orthogonalization process is studied and compared in [19], [33].

Denote by

$$\delta := \min\{k | \dim \mathcal{K}_k(\mathbf{A}, r_0) = \dim \mathcal{K}_{k+1}(\mathbf{A}, r_0)\}$$

degree of the vector r_0 with respect to the matrix $\mathbf{A}$. The GMRES method has the following well known properties.

Lemma 1.2: *Let x^* be the exact solution of (1.1) and x_k the kth GMRES approximate solution, $k = 1, \ldots, n$. Then*
(i) $x_\delta = x^*$
(ii) $x_k \neq x^*$ for $k < \delta$
(iii) $x_k = x^*$ for $k > \delta$.
Moreover, dim $\mathcal{K}_k(\mathbf{A}, r_0)$ = dim $\mathcal{K}_\delta(\mathbf{A}, r_0)$ for $k > \delta$.

As a consequence of Lemma 1.2, in exact arithmetics the GMRES algorithm using orthogonalization process given by Algorithm 1.1 or 1.2 stops with $h_{k+1,k} = 0$ if and only if the exact solution is found, see [73] p. 164. Unfortunately, this fact is no longer true in finite precision arithmetics. Further, GMRES is a method with long recurrences, the memory requirements and computation time grow quickly with k and therefore the full version is not used in practice.

The number of operations per iteration grows generally as $O(k^2n)$ (the most costly operation is orthogonalization of the basis vector against all the previous Arnoldi vectors) and the storage requirements as $O(kn)$ (this represents the k Arnoldi vectors of the length n that need to be saved). Therefore the GMRES algorithm becomes impractical when n is large. There are two main remedies that can be used. In Quasi–GMRES or DQGMRES (see [73] pp.168-172) full orthogonalization process is replaced by incomplete one where the basis vector v_i is orthogonalized only against the last $l < i-1$ vectors. This technique accelerates the computation of one single iteration but leads to the loss of accuracy in the solution. Thus a *restarted version* of the GMRES algorithm is more popular, i.e. the algorithm is stopped after some number of iterations, say $m \ll n$, and the method is restarted with $x_0 := x_m$. Algorithm of the restarted GMRES method follows.

ALGORITHM 1.3 (GMRES(m)):

```
input: $x_0, b$, $\mathbf{A}$, $m$ - restart length,
        $TOL$ - tolerance for $\|r_m\|/\|r_0\|$
        $max$ – maximal number of iterations
output: $x_m$ - approximate solution, $r_m$ - corresponding residual

$\hat{r}_0 := b - \mathbf{A}x_0$
$x_m := x_0,\ r_m := r_0 = \hat{r}_0$
for $i = 0, 1, \dots, max$
        $x_0 := x_m,\ r_0 := r_m$
        perform $m$ steps of orthogonalization process with $\mathbf{A}$ and starting vector $r_0$
        denote by $\mathbf{V}_m := (v_1, .., v_m),\ \mathbf{V}_{m+1} := (\mathbf{V}_m, v_{m+1}),\ \mathbf{H}_{m+1,m} := (h_{ij})_{i,j=1}^{m+1,m}$
        compute QR–decomposition $\mathbf{Q}_{m+1}\mathbf{R}_{m+1,m} = \mathbf{H}_{m+1,m}$
        $g_{m+1} := \beta\mathbf{Q}_{m+1}^T e_1$
        compute $y_m$ solving (1.8)
        $x_m := x_m + \mathbf{V}_m y_m,\ r_m := b - \mathbf{A}x_m$
        if $\|r_m\|/\|\hat{r}_0\| < TOL$ then STOP
end $i$
```

It is well known, that restart slows down the convergence and stagnation may occur in many cases. This phenomenon was widely studied, e.g., in [84], [95], [2], but is still not well understood. It has been observed on many examples that the eigenvalues of the matrix $\mathbf{A}$ close to zero may cause the troubles for some types of linear systems. Hence several preconditioning techniques are based on the idea to remove them the spectrum. For preconditioning to be effective, the faster convergence have to overcome the costs of computing the preconditioner, so that the total cost of solving (1.1) is lower. Therefore these preconditioners are constructed by exploiting spectral information gathered in the previous restart from the Arnoldi factorization. First a rectangular matrix $\mathbf{V}$ is constructed whose columns generate an invariant subspace of $\mathbf{A}$ corresponding to the smallest eigenvalues and then the preconditioning matrix is relatively cheaply computed, see [3], [12], [13], [21], [23]. We will not discuss all of these techniques here, although they can give interesting results in some cases. We concentrate on preconditioners presented in [3] and [21]. Main ideas are briefly summarized in the following section.

1.4 Construction of preconditioners

The main goal of preconditioning is to decrease the computational effort needed to solve the linear system (1.1). Such system can be preconditioned by a nonsingular matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ from the left

$$\mathbf{M}^{-1}\mathbf{A}x = \mathbf{M}^{-1}b \tag{1.9}$$

or from the right

$$\mathbf{A}\mathbf{M}^{-1}\mathbf{M}x = b. \tag{1.10}$$

Both techniques can also be combined to obtain a split preconditioning

$$\mathbf{N}^{-1}\mathbf{A}\mathbf{M}^{-1}\mathbf{M}x = \mathbf{N}^{-1}b,$$

where $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{n\times n}$ are both nonsingular matrices. Matrices $\mathbf{M}$ from equation (1.9) and (1.10) are called left and right preconditioner, respectively, and can be constructed such that the matrix $\mathbf{M}^{-1}\mathbf{A}$ or $\mathbf{A}\mathbf{M}^{-1}$ does not have small eigenvalues of $\mathbf{A}$ in its spectrum. Now we present two preconditioning techniques with this property.

Adaptive preconditioning from left

Let the eigenvalues of $\mathbf{A}$ be ordered according to

$$0 < |\lambda_1| \leq |\lambda_2| \leq \ldots \leq |\lambda_k| < |\lambda_{k+1}| \leq \ldots \leq |\lambda_n|,$$

where usually $k \ll n$. In this part we assume that $|\lambda_n| = 1$ and, moreover, $\mathbf{A}$ is diagonalizable. Let $\mathbf{V} \in \mathbb{R}^{n\times k}$ be an orthogonal matrix whose columns generate an invariant subspace of $\mathbf{A}$ corresponding to the k smallest eigenvalues. The following theorem demonstrates the construction of preconditioner by using $\mathbf{V}$, see [3].

Theorem 1.3: *Let $\mathbf{V} \in \mathbb{R}^{n\times k}, \mathbf{Q} = [\mathbf{V}, \mathbf{W}] \in \mathbb{R}^{n\times n}$ be matrices with orthonormal columns, where span$\{\mathbf{V}\}$ is an invariant subspace of $\mathbf{A}$ corresponding to the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$. Let $\mathbf{H} = \mathbf{V}^T\mathbf{A}\mathbf{V}$ be nonsingular. Then the matrix*

$$\mathbf{M} = \mathbf{V}\mathbf{H}\mathbf{V}^T + \mathbf{W}\mathbf{W}^T \tag{1.11}$$

is nonsingular and $\mathbf{M}^{-1} = \mathbf{V}\mathbf{H}^{-1}\mathbf{V}^T + \mathbf{W}\mathbf{W}^T$*. Moreover, eigenvalues of the matrix* $\mathbf{M}^{-1}\mathbf{A}$ *are*

$$\lambda_{k+1}, \lambda_{k+2}, \ldots, \lambda_n, 1,$$

where 1 *has multiplicity at least* k*.*

Theorem 1.3 shows that a preconditioner removing the k smallest eigenvalues from the spectrum of $\mathbf{A}$ can be easily constructed, if we have an invariant subspace corresponding to these eigenvalues. The preconditioner replaces the small eigenvalues by multiple eigenvalue 1.

Remark: Theorem 1.2 can be proved with exactly the same statement without the assumption $|\lambda_n| = 1$. The theorem shows how a prescribed set of eigenvalues $\{\lambda_1, \ldots, \lambda_k\}$ can be replaced by the eigenvalue 1. Our aim is to replace small eigenvalues by a larger number, but we do not explicitly know the distribution of eigenvalues of $\mathbf{A}$. Therefore it seems to be a good idea to normalize (1.1) such that $|\lambda_n| = 1$ and thus the preconditioner (1.11) replaces the small eigenvalues of $\mathbf{A}$ by the eigenvalue with the magnitude equal to magnitude of λ_n.

Remark: Assumption $|\lambda_n| = 1$ can be approximately fulfilled after the first restart of GMRES(m) by dividing both sides of equation (1.1) by a number approximating $|\lambda_n|$. A norm of the largest in magnitude eigenvalue of the matrix $\mathbf{H}_m$, can be considered for such number. Details about when the eigenvalues of $\mathbf{H}_m$ approximate some eigenvalues of $\mathbf{A}$ are given in the following section.

This technique was proposed in [3] as an *adaptive preconditioning* technique, i.e. the preconditioner (1.11) is updated in each restart of the GMRES(m) method in the following sense. Let $\mathbf{V}^{(1)}$ be an approximation of the matrix $\mathbf{V}$ obtained in the first restart of GMRES(m). Then, using Theorem 1.2, the left preconditioner to the system (1.1) is set to

$$\mathbf{M}_1^{-1} := \mathbf{V}^{(1)}(\mathbf{H}^{(1)})^{-1}(\mathbf{V}^{(1)})^T + \mathbf{I} - \mathbf{V}^{(1)}(\mathbf{V}^{(1)})^T.$$

If $\mathbf{V}^{(1)}$ is a good approximation to $\mathbf{V}$, the preconditioned matrix $\mathbf{M}_1^{-1}\mathbf{A}$ does not have k smallest in magnitude eigenvalues of $\mathbf{A}$ in the spectrum. Now we perform second restart of GMRES(m) applied to the system

$$\mathbf{M}_1^{-1}\mathbf{A}x = \mathbf{M}_1^{-1}b$$

and compute a new approximation $\mathbf{V}^{(2)}$ of invariant subspace of $\mathbf{M}_1^{-1}\mathbf{A}$ corresponding to the k smallest eigenvalues. We form the second preconditioner $\mathbf{M}_2$ to remove these values from the spectrum of $\mathbf{M}_1^{-1}\mathbf{A}$. We update the preconditioner such that $\mathbf{M}^{-1} := \mathbf{M}_2^{-1}\mathbf{M}_1^{-1}$ and apply GMRES(m) method to the system

$$\mathbf{M}^{-1}\mathbf{A}x = \mathbf{M}^{-1}b,$$

etc. It is possible to continue removing small eigenvalues this way until a good preconditioner of (1.1) is found and fast convergence is obtained. In practice, the preconditioning matrix is usually updated only several times and then computation continues with fixed preconditioner. For detailed algorithm see [3] p. 217.

Remark: If only a coarse approximation $\mathbf{V}^{(1)}$ is used, which is often the case, some of the k smallest eigenvalues can remain in the spectrum of the preconditioned matrix $\mathbf{M}_1^{-1}\mathbf{A}$. This does not necessarily mean a problem in computation, because as we update the preconditioner these eigenvalues can be removed in some of the following restarts.

Adaptive preconditioning from right

Similar idea was used in [21] to construct a right preconditioner. The fundamental theorem from this paper follows.

Theorem 1.4: *Let* $\mathbf{V} \in R^{n\times k}$ *be a matrix with orthonormal columns, where span*$\{\mathbf{V}\}$ *is an invariant subspace of* $\mathbf{A}$ *corresponding to the eigenvalues* $\lambda_1, \dots, \lambda_k$. *Let this space have full dimension* k *and denote by* $\mathbf{T} = \mathbf{V}^T\mathbf{A}\mathbf{V}$. *Then the matrix*

$$\mathbf{M} = \mathbf{I}_n + \mathbf{V}(1/|\lambda_n|\mathbf{T} - \mathbf{I}_k)\mathbf{V}^T \tag{1.12}$$

is nonsingular and $\mathbf{M}^{-1} = \mathbf{I}_n + \mathbf{V}(|\lambda_n|\mathbf{T}^{-1} - \mathbf{I}_k)\mathbf{V}^T$. *Moreover, eigenvalues of the matrix* $\mathbf{A}\mathbf{M}^{-1}$ *are*

$$\lambda_{k+1}, \lambda_{k+2}, \dots, \lambda_n, |\lambda_n|,$$

where $|\lambda_n|$ *has multiplicity at least* k.

The preconditioner (1.12) has properties similar to the preconditioner (1.11), except that here the small eigenvalues are replaced by $|\lambda_n|$. Update of the matrix $\mathbf{M}$ can be performed similarly as in the previous case, but authors in [21] proposed slightly different method. After each restart, the matrix $\mathbf{V}$ is enriched by columns approximating a prescribed number (usually one or two) of basis vectors of invariant subspace of $\mathbf{A}$. Therefore the number of columns of the matrix $\mathbf{V}$ grows. The preconditioner is fixed after some number of restarts and computation continues with constant preconditioner. Number $|\lambda_n|$ can be approximated by the largest eigenvalue of $\mathbf{H}_m$. For detailed algorithm see [21] p. 309.

1.5 Properties of Arnoldi factorization

In this section, properties of the Arnoldi factorization are discussed. The following considerations lead to idea, how to construct a basis of an invariant subspace of $\mathbf{A}$ corresponding to a prescribed part of the spectrum.

We have presented in Section 1.2 that both the Arnoldi and the Householder algorithm with starting vector v_1, $\|v_1\| = 1$, yield after k steps the Arnoldi factorization of the form

$$\mathbf{A}\mathbf{V}_k = \mathbf{V}_k\mathbf{H}_k + w_k e_k^T, \tag{1.13}$$

where the columns of $\mathbf{V}_k$ form an orthonormal basis of the Krylov subspace $\mathcal{K}_k(\mathbf{A}, v_1)$, $w_k \perp \text{span}\{\mathbf{V}_k\}$ and $\mathbf{H}_k$ is an upper Hessenberg matrix. In the following we assume that the matrices $\mathbf{H}_k$ are *unreduced*, i.e. the Krylov subspace has full dimension. This is equivalent to

$$\|w_i\| = h_{i+1,i} \neq 0 \;\text{ for }\; i = 1, \dots, k-1.$$

In the opposite case, i.e. if $\dim(\mathcal{K}_k(\mathbf{A}, v_1)) < k$, it can be proved that the exact solution x^* of (1.1) satisfies

$$x^* \in x_0 + \mathcal{K}_k(\mathbf{A}, v_1)$$

and the GMRES algorithm stops yielding the exact solution. Assuming that $h_{k+1,k} \neq 0$, denote by $v_{k+1} := w_k / h_{k+1,k}$.

The matrix $\mathbf{H}_k$ is a Ritz-Galerkin approximation of the matrix $\mathbf{A}$ on the subspace $\mathcal{K}_k(\mathbf{A}, v_1)$. Therefore spectral information about $\mathbf{A}$ can be gathered through the factorization (1.13) using two types of values and vectors that characterize spectral properties of $\mathbf{H}_k$ - Ritz values and vectors and harmonic Ritz values and vectors, respectively.

Ritz values and vectors

Denote by $\theta_1, \dots, \theta_k$ eigenvalues of the matrix $\mathbf{H}_k$ and g_i the eigenvector corresponding to θ_i, i.e.

$$\mathbf{H}_k g_i = \theta_i g_i, \;\; g_i \in \mathbb{C}^n, \;\; \|g_i\| = 1. \tag{1.14}$$

The numbers θ_i are called *Ritz values* of $\mathbf{A}$ and the vectors $y_i := \mathbf{V}_k g_i$ are called the corresponding *Ritz vectors.* The pair (θ_i, y_i) is called Ritz pair.

It can be proved (see [49]) that if the Hessenberg matrix $\mathbf{H}_k$ is unreduced, each of its multiple eigenvalues has only one eigenvector associated with it. Hence multiple Ritz values have only one corresponding Ritz vector and the matrix $\mathbf{H}_k$ is *nonderogatory.* In case the Ritz value θ_i has algebraic multiplicity k_i we have

$$\mathbf{H}_k(g_i^{(1)}, g_i^{(2)}, \dots, g_i^{(k_i)}) = (\theta_i g_i^{(1)}, \theta_i g_i^{(2)} + g_i^{(1)}, \dots, \theta_i g_i^{(k_i)} + g_i^{(k_i-1)}).$$

Here $g_i^{(1)}$ is an eigenvector of $\mathbf{H}_k$ and $g_i^{(j)}$ is a principal vector of order j. Analogously we can define $y_i^{(j)} := \mathbf{V}_k g_i^{(j)}$ for $j = 1, 2, \dots, k_i$. Although all the following consequences remain true in this general case, we assume in the following that $\mathbf{H}_k$ has k distinct eigenvalues to avoid the notational difficulties. Opposite situation is discussed in the second part of Chapter 3.

Harmonic Ritz values and vectors

Consider the inverse problem, i.e. an orthogonal section of the matrix $\mathbf{A}^{-1}$ onto the Krylov subspace $\mathcal{K}_k(\mathbf{A}^{-1}, \mathbf{A}^k v_1) = \mathbf{A}\mathcal{K}_k(\mathbf{A}, v_1)$. The corresponding Arnoldi factorization has the form

$$\mathbf{A}^{-1}\mathbf{S}_k = \mathbf{S}_k\mathbf{T}_k + t_{k+1,k}s_{k+1}e_k^T, \tag{1.15}$$

where $\mathbf{S}_k e_1 = s_1 = \mathbf{A}^k v_1/\beta_k$ with $\beta_k = \|\mathbf{A}^k v_1\|$ and the columns of the matrix $(\mathbf{S}_k, s_{k+1})$ form an orthonormal basis of $\mathcal{K}_{k+1}(\mathbf{A}^{-1}, \mathbf{A}^k v_1) = \mathcal{K}_{k+1}(\mathbf{A}, v_1)$.

Denote by $\tilde{\theta}_i$ inverses of the eigenvalues of the matrix $\mathbf{T}_k$ and $\tilde{z}_i$ eigenvector corresponding to $\tilde{\theta}_i^{-1}$, i.e.

$$\mathbf{T}_k\tilde{z}_i = \tilde{\theta}_i^{-1}\tilde{z}_i, \;\; \tilde{z}_i \in \mathbb{C}^n, \;\; \|\tilde{z}_i\| = 1. \tag{1.16}$$

The numbers $\tilde{\theta}_i$ are called *harmonic Ritz values* of $\mathbf{A}$. To obtain these values we need not form the inverse of the matrix $\mathbf{A}$ because they satisfy the generalized eigenvalue problem

$$\mathbf{V}_k^T\mathbf{A}^T\mathbf{V}_k\tilde{g}_i = \tilde{\theta}_i^{-1}\mathbf{V}_k^T\mathbf{A}^T\mathbf{A}\mathbf{V}_k\tilde{g}_i, \tag{1.17}$$

for more details see [23], [32], [56], [58]. The vectors $\tilde{y}_i := \mathbf{V}_k\tilde{g}_i$ are called *harmonic Ritz vectors* of $\mathbf{A}$ with respect to $\mathcal{K}_k(\mathbf{A}, v_1)$. The pair $(\tilde{\theta}_i, \tilde{y}_i)$ is called harmonic Ritz pair.

Using (1.3), equation (1.17) becomes

$$\begin{aligned} \mathbf{H}_k^T\tilde{g}_i &= \tilde{\theta}_i^{-1}\mathbf{H}_{k+1,k}^T\mathbf{H}_{k+1,k}\tilde{g}_i \quad \text{or} \\ \mathbf{H}_k^T\tilde{g}_i &= \tilde{\theta}_i^{-1}\mathbf{R}_k^T\mathbf{R}_k\tilde{g}_i \quad \text{or} \\ (\mathbf{H}_k + h_{k+1,k}\mathbf{H}_k^{-T}e_ke_k^T)\tilde{g}_i &= \tilde{\theta}_i^{-1}\tilde{g}_i, \end{aligned}$$

where $\mathbf{R}_k$ is an upper triangular matrix obtained from QR–decomposition of the matrix $\mathbf{H}_{k+1,k}$. This gives several possibilities to compute the harmonic Ritz values and vectors by using only the matrix $\mathbf{H}_k$ of smaller dimension k.

Remark: The vectors $\mathbf{S}_k\tilde{z}_i$ are harmonic Ritz vectors of $\mathbf{A}^{-1}$ with respect to $\mathcal{K}_k(\mathbf{A}^{-1}, \mathbf{A}^k v_1)$. It is easy to show that $\mathbf{S}_k\tilde{z}_i = \mathbf{A}\tilde{y}_i$. Here we work only with harmonic Ritz vectors of $\mathbf{A}$ with respect to $\mathcal{K}_k(\mathbf{A}, v_1)$ defined above and call them shortly harmonic Ritz vectors.

Approximation of eigenpairs

Now the question arises when the Ritz value θ_i or the harmonic Ritz value $\tilde{\theta}_i$ can be considered a good approximation to an eigenvalue of $\mathbf{A}$. First, consider the situation that the orthogonalization process gives the decomposition (1.13) with $w_k = 0$, i.e. $\mathbf{A}\mathbf{V}_k = \mathbf{V}_k\mathbf{H}_k$. Let $\mathbf{G}_k^{-1}\mathbf{H}_k\mathbf{G}_k = \mathbf{J}$ be the Jordan canonical decomposition of the matrix $\mathbf{H}_k$, with the columns of the matrix $\mathbf{G}_k$ being the eigenvectors of $\mathbf{H}_k$. Then the Jordan matrix $\mathbf{J}$ is a block in the Jordan canonical form of the matrix $\mathbf{A}$ and the columns of the matrix $\mathbf{V}_k\mathbf{G}_k$ are eigenvectors of the matrix $\mathbf{A}$ corresponding to the eigenvalues included in the matrix $\mathbf{J}$. In this case the Ritz values are exact eigenvalues of the matrix $\mathbf{A}$ and the columns of $\mathbf{V}_k\mathbf{G}_k$ form a basis of the corresponding eigenspace. The following assertion indicates when the orthogonal process stops with $w_k = 0$, see [81]. This result is important for our further considerations.

Proposition 1.5: *Let the matrix* $\mathbf{H}_k$ *in the Arnoldi factorization (1.13) be unreduced and* $\mathbf{G}\mathbf{J} = \mathbf{A}\mathbf{G}$ *be a Jordan matrix of order* k*, where* $\mathbf{G}$ *has rank* k*. Then* $w_k = 0$ *if and only if* v_1 *lies in the space generated by the columns of the matrix* $\mathbf{G}$.

Now, consider a more usual situation $w_k = h_{k+1,k}v_{k+1} \neq 0$. Let (θ, g) be an arbitrary eigenpair of $\mathbf{H}_k$ and y the corresponding Ritz vector, i.e. $y = \mathbf{V}_k g$. Then

$$\|\mathbf{A}\mathbf{V}_k g - \mathbf{V}_k\mathbf{H}_k g\| = \|\mathbf{A}y - \theta y\| = h_{k+1,k}\,|e_k^T g| \tag{1.18}$$

indicates that if $h_{k+1,k}$ is small, then the Ritz pair (θ, y) will be in many cases a reasonable approximation to an eigenpair of $\mathbf{A}$ and, according to (1.13), the columns of $\mathbf{V}_k$ will nearly span an invariant subspace of $\mathbf{A}$, see [4]. Unfortunately, we must note that these considerations are not true for matrices that are highly non-normal. The relative departure from normality can be measured by *Henrici number* [41]

$$\|\mathbf{A}\mathbf{A}^T - \mathbf{A}^T\mathbf{A}\|_F / \|\mathbf{A}\|_F.$$

If this number is large, the matrix $\mathbf{A}$ is considered highly non-normal and the basis of its eigenvectors is ill-conditioned, see [16]. Thus small number $h_{k+1,k}$ or more precisely $h_{k+1,k}\,|e_k^T g|$ does not imply that the Ritz pair (or similarly the harmonic Ritz pair) is an accurate approximation to an eigenpair of $\mathbf{A}$. The influence of high nonnormality on convergence of iterative methods for solving system (1.1) and eigenvalue solvers is discussed, e.g., in [5], [11].

To make these facts more obvious, rewrite the Arnoldi factorization (1.13)

$$\mathbf{A}\mathbf{V}_k = \mathbf{V}_k\mathbf{H}_k + h_{k+1,k}v_{k+1}e_k^T = \mathbf{V}_k\mathbf{H}_k + h_{k+1,k}v_{k+1}v_k^T\mathbf{V}_k,$$

in the form

$$(\mathbf{A} - h_{k+1,k}\mathbf{E})\mathbf{V}_k = \mathbf{V}_k\mathbf{H}_k,$$

where $\mathbf{E} = v_{k+1}v_k^T$. Multiplying the last equation by g_i yields

$$(\mathbf{A} - h_{k+1,k}\mathbf{E})y_i = \theta_i y_i$$

and thus the Ritz pair (θ_i, y_i) is at the same time the eigenpair of the perturbed matrix $\mathbf{A} - h_{k+1,k}\mathbf{E}$. Analogously from (1.15) we have

$$(\mathbf{A}^{-1} - t_{k+1,k}\tilde{\mathbf{E}})\tilde{y}_i = \tilde{\theta}_i^{-1}\tilde{y}_i,$$

where $\tilde{\mathbf{E}} = \mathbf{A}^{-1}s_{k+1}s_k^T\mathbf{A}$. Hence the following considerations can be transformed for a harmonic Ritz pair. According to analysis given in [74], [82], we have the following assertion:

Proposition 1.6: *Let θ_i be a simple eigenvalue of the matrix $\mathbf{A} - h_{k+1,k}\mathbf{E}$, p_i the corresponding right eigenvector and q_i the corresponding left eigenvector. If $h_{k+1,k} \ll 1$ then there exists exactly one eigenvalue λ of the matrix $\mathbf{A}$ such that*

$$\lambda = \theta_i + h_{k+1,k}\frac{q_i^H\mathbf{E}p_i}{q_i^H p_i} + O(h_{k+1,k}^2).$$

Hence if $h_{k+1,k} \ll 1$ and if the angles between each pair p_i and q_i are small, then each Ritz value approximates some eigenvalue of $\mathbf{A}$ well. The last relation, together with (1.18), motivates the search for an algorithm that reduces the magnitude of $\|w_k\| = h_{k+1,k}$ in the Arnoldi factorization by an appropriate update of the starting vector v_1.

It was observed that the Arnoldi process estimates the large eigenvalues more accurately than the small ones and that the Ritz vectors corresponding to the largest eigenvalues have better significance. The harmonic Ritz values are computed by a process which moves the small in magnitude eigenvalues to the exterior of the spectrum, see [23], [32], [59]. Therefore it is often argued that the harmonic Ritz values can give better approximations to the small eigenvalues than the Ritz values. Discussion can be found in [23], [59], [56], [55]. The appropriateness of using harmonic Ritz values in connection with the GMRES method is described in [32]. In our theoretical investigations it is often not important to distinguish between the Ritz and the harmonic Ritz values and vectors. Hence we do not specify what sort is used, beyond some special cases. Difference will be seen in numerical experiments.

Arnoldi basis in polynomial form

Orthonormal basis of the Krylov subspace $\mathcal{K}_k(\mathbf{A}, v_1)$ can be constructed using a Frobenius matrix. Even thought this technique is not suitable for real computations, it shows several interesting properties of the Arnoldi factorization.

The difference between the Arnoldi process and the application of the Frobenius matrix is in construction. The Arnoldi process starting with v_1, constructs the vector v_2 in the following way. The vector $\mathbf{A}v_1$ is projected to the span$\{v_1\}$ using the projection $v_1 v_1^T$. Putting $w_1 := (\mathbf{I} - v_1 v_1^T)\mathbf{A}v_1$ immediately yields that $w_1 \perp v_1$ and v_2 is obtained by normalizing w_1. In the second step the vector $\mathbf{A}v_2$ is projected to the span$\{v_1, \mathbf{A}v_1\}$, denote this orthogonal projection by $\hat{w}$, and the vector v_3 is obtained by normalizing $w_2 := \mathbf{A}v_2 - \hat{w}$. If the Frobenius matrix is used, the first step is the same but in the second step the vector $\mathbf{A}^2 v_1$ is projected instead of $\mathbf{A}v_2$ and the projection is orthogonal to the span$\{v_1, \mathbf{A}v_1\}$. For completeness, we briefly describe this technique for arbitrary step k.

Denote by $\mathbf{K} := (v_1, \mathbf{A}v_1, \dots, \mathbf{A}^{k-1}v_1)$ the Krylov matrix and by $\mathbf{F} \in \mathbb{R}^{k \times k}$ the Frobenius matrix, i.e.

$$\mathbf{F} := \begin{pmatrix} 0 & f_1 \\ \mathbf{I} & \hat{f} \end{pmatrix},$$

where $f_1 \in \mathbb{R}, \hat{f} \in \mathbb{R}^{k-1}$ and put $f := (f_1, \hat{f}^T)^T$. The form of the matrices $\mathbf{K}$ and $\mathbf{F}$ immediately yields

$$\mathbf{AK} = \mathbf{KF} + \hat{r}e_k^T, \quad \hat{r} = \mathbf{A}^k v_1 - \mathbf{K}f \tag{1.19}$$

and a vector f is sought such that $\|\hat{r}\|$ is minimal. Denoting

$$\mathbf{S}_l := \mathbf{H}_{k-1,k-2}\ \mathbf{H}_{k-2,k-3}\ \dots\ \mathbf{H}_{k-l,k-l-1}$$

it can be proved by induction that

$$\mathbf{K} = \mathbf{V}_k \mathbf{R}_k, \tag{1.20}$$

where $\mathbf{R_k} := (e_1^{(k)}, \mathbf{S}_1 e_1^{(k-1)}, \mathbf{S}_2 e_1^{(k-2)}, \dots, \mathbf{S}_{k-1})$ is an upper triangular matrix with diagonal elements

$$r_{ii} := e_i^T \mathbf{R}_k e_i = h_{1,0} h_{2,1} \dots h_{i,i-1}, \tag{1.21}$$

where $h_{1,0} := 1$ and $h_{i+1,i} = \|w_i\|$. Apparently $r_{ii} > 0$ for all $i = 1, \dots, k$ and (1.20) represents the QR–decomposition of the matrix $\mathbf{K}$. Multiplying (1.19) by $\mathbf{R}_k^{-1}$ we obtain

$$\mathbf{AV}_k = \mathbf{V}_k \mathbf{R}_k \mathbf{F} \mathbf{R}_k^{-1} + \hat{r} e_k^T \mathbf{R}_k^{-1}, \tag{1.22}$$

where $\mathbf{T}_k := \mathbf{R}_k \mathbf{F} \mathbf{R}_k^{-1}$ is an upper Hessenberg matrix and

$$\hat{r} e_k^T \mathbf{R}_k^{-1} = \hat{r}(0, \dots, 0, 1/r_{kk}) = (1/r_{kk})\hat{r} e_k^T = \hat{w} e_k^T,$$

where $\hat{w} = (1/r_{kk})\,\hat{r}$. If f minimizes the norm $\|\mathbf{A}^k v_1 - \mathbf{K}f\|$, i.e.

$$f = \arg\min_{\tilde{f} \in \mathbb{R}^k} \|\mathbf{A}^k v_1 - \mathbf{K}\tilde{f}\|, \tag{1.23}$$

then $\hat{r}$ is orthogonal to $\mathcal{K}_k(\mathbf{A}, v_1)$ and thus $\mathbf{V}^T \hat{w} = 0$. Therefore (1.22) is the Arnoldi factorization identical with (1.13), i.e. $\mathbf{T}_k = \mathbf{H}_k$ and $\hat{w} = w_k$, and the Frobenius matrix formulation yields in exact arithmetics the Arnoldi basis $v_1, \dots, v_k$, for more details see, e.g., [81], [87].

The second equation in (1.19) implies that $\hat{r} = \hat{p}_k(\mathbf{A})v_1$, where $\hat{p}_k$ is the characteristic polynomial of $\mathbf{F}$. This fact together with (1.23) gives

$$\hat{p}_k(\mathbf{A}) = \arg \min_{p \in MP_k} \|p(\mathbf{A})v_1\|. \tag{1.24}$$

Combining the last result with the fact that $w_k = \hat{w} = (1/r_{kk})\,\hat{r}$ yields

$$\|w_k\| \ = \ 1/r_{kk}\|\hat{r}\| \ = \ \|\hat{p}_k(\mathbf{A})v_1\|/r_{kk}. \tag{1.25}$$

It is easy to see that the last equality remains true if any $i \in \{1, \dots, k\}$ is substituted for k. Moreover, it follows from (1.21) that

$$\frac{r_{i+1,i+1}}{r_{ii}} = h_{i+1,i} = \|w_i\|. \tag{1.26}$$

The equalities (1.25) and (1.26) yield $r_{i+1,i+1} = \|\hat{p}_i(\mathbf{A})v_1\|$. Finally, we have

$$w_i \ = \ \hat{p}_i(\mathbf{A})v_1 \ / \ \|\hat{p}_{i-1}(\mathbf{A})v_1\|, \tag{1.27}$$

$$h_{i+1,i} = \|w_i\| = \|\hat{p}_i(\mathbf{A})v_1\| \ / \ \|\hat{p}_{i-1}(\mathbf{A})v_1\|. \tag{1.28}$$

The important result is that the number $h_{k+1,k}$, that represents a part of the residual of the Ritz eigenpair, see (1.18), is a function of the starting vector v_1. Its form (1.28) will be used in Chapter 4 to derive some convergence results for algorithms for computation of invariant subspaces of $\mathbf{A}$.

Chapter 2

Polynomial filters

In the previous chapter, it was described how a matrix $\mathbf{V}$*, whose columns generate an invariant subspace of* $\mathbf{A}$ *corresponding to the smallest eigenvalues, can be used to construct the left and right preconditioner of the GMRES method. In this chapter we concentrate on techniques for computation of the matrix* $\mathbf{V}$*, by using an iterative procedure which updates the starting vector of the Arnoldi factorization by a polynomial in matrix* $\mathbf{A}$ *called polynomial filter. First the classical technique - implicitly restarted Arnoldi process with shifts is briefly summarized. Then a new technique based on the properties of Tchebychev polynomials is presented.*

2.1 Construction of invariant subspaces

In this chapter, we discuss techniques for construction of a matrix $\mathbf{V}$, whose columns generate an invariant subspace of $\mathbf{A}$ corresponding to the prescribed part of the spectrum. After one restart of GMRES(m), the Arnoldi factorization is obtained. In Section 1.5, we have discussed when the Ritz or the harmonic Ritz values and vectors, available through this factorization, are good approximations to the eigenvalues and the eigenvectors of $\mathbf{A}$. The idea steams from the fact that $h_{k+1,k} = \|w_k\| = 0$ in (1.13) if and only if the columns of $\mathbf{V}$ span an invariant subspace of $\mathbf{A}$ (see Proposition 1.5). When $\mathbf{V}$ nearly spans such subspace, $h_{k+1,k}$ will be small. On the other hand, if $\mathbf{A}$ is not too nonnormal and $h_{k+1,k}$ (or more precisely $h_{k+1,k}|e_k^T g|$, where g is an eigenvector of $\mathbf{H}_{k+1,k}$) becomes small, the Ritz values can be considered for good approximations to the eigenvalues of $\mathbf{A}$ (see Proposition 1.6). The same considerations holds for the Ritz vectors in many cases. These considerations motivates the search for an algorithm that reduces the magnitude of $h_{k+1,k}$ by an appropriate update of the starting vector $v_1 = \mathbf{V}_k e_1$, because $h_{k+1,k}$ is a function of v_1. Proposition 1.5 also suggests that one might find an invariant subspace by replacing v_1 by a linear combination of approximate eigenvectors corresponding to the smallest eigenvalues. Note that construction of invariant subspaces by modifying the vector v_1 was generally studied, e.g., in [4].

Let the eigenvalues of $\mathbf{A}$ be ordered according to

$$0 < |\lambda_1| \leq |\lambda_2| \leq \ldots \leq |\lambda_k| < |\lambda_{k+1}| \leq \ldots \leq |\lambda_n|.$$

and define the sets

$$\sigma_W = \{\lambda_1, \ldots, \lambda_k\}, \quad \sigma_U = \{\lambda_{k+1}, \ldots, \lambda_n\}$$

of *wanted* and *unwanted eigenvalues*, respectively. Then $\sigma(\mathbf{A}) = \sigma_W \cup \sigma_U$. Let l be a positive integer and $m := k + l \ll n$. Assume that the Arnoldi process with the starting vector v_1 does not stop before the $(k + l)$th step. Then we obtain the Arnoldi factorization

$$\mathbf{A}\mathbf{V}_{k+l} = \mathbf{V}_{k+l}\mathbf{H}_{k+l} + w_{k+l}e_{k+l}^T, \tag{2.1}$$

where $\mathbf{V}_{k+l} \in \mathbb{R}^{n\times(k+l)}, \mathbf{H}_{k+l} \in \mathbb{R}^{(k+l)\times(k+l)}, w_{k+l} \in \mathbb{R}^n$. Let us order the Ritz and the harmonic Ritz values as

$$0 < |\theta_1| \leq |\theta_2| \leq \ldots \leq |\theta_{k+l}| \quad \text{and} \quad 0 < |\tilde{\theta}_1| \leq |\tilde{\theta}_2| \leq \ldots \leq |\tilde{\theta}_{k+l}|.$$

We will assume in the rest of this chapter that $\mathbf{H}_{k+l}$ has $k+l$ distinct eigenvalues to avoid the notational difficulties, although all the following consequences remain true in the general case.

Our goal is to update the starting vector $v_1 = \mathbf{V}_{k+l}e_1$ of the Arnoldi process in order to obtain good approximations to the eigenvalues from σ_W and the corresponding eigenspace. Putting $v_1^{(0)} := v_1$, we find a new $v_1^{(1)} = \psi_1(\mathbf{A})v_1$ such that $\psi_1 \in P_l$ filters out the l largest eigenvalues from $\sigma(\mathbf{H}_{k+l})$. Such polynomials will be called *polynomial filters.* Repeated application of filters gives the following schema

$$\begin{aligned} v_1^{(1)} &= \psi_1(\mathbf{A})v_1 \\ v_1^{(2)} &= \psi_2(\mathbf{A})v_1^{(1)} \\ &\vdots \\ v_1^{(i)} &= \psi_i(\mathbf{A})v_1^{(i-1)} \end{aligned} \tag{2.2}$$

Denote by

$$\mathbf{A}\mathbf{V}_k^{(i)} = \mathbf{V}_k^{(i)}\mathbf{H}_k^{(i)} + w_k^{(i)}e_k^T, \quad \mathbf{V}_k^{(i)}e_1 = v_1^{(i)}, \tag{2.3}$$

the Arnoldi factorization obtained in the ith step. We want to construct the polynomials ψ_i such that $\lim_{i\to\infty} \|w_k^{(i)}\| = 0$. In the rest of this section we assume that the polynomials ψ_i are scaled so that $||v_1^{(i)}|| = 1$ for $i = 1, 2, \ldots$

The following algorithm presents the successive application of polynomial filters. Note that the modified Arnoldi or the Householder process (see Section 1.2) can be used as an orthogonalization process.

```
ALGORITHM 2.1 (Computation of invariant subspace):
input: $v_1, k, l, TOL$,
        $max$ – maximal number of iterations
        POL – procedure for construction of polynomial $\psi$
output: $\mathbf{V}_k$ such that $\|w_k\| < TOL$

perform k steps of orthogonalization process with $\mathbf{A}$ and starting vector $v_1$ (we
obtain (2.3) for $i = 0$)
for $i = 1, \ldots, max$
    1. perform l additional steps of orthogonalization process (we obtain (2.1))
    2. compute Ritz (resp. harmonic Ritz) values $\theta_1, \ldots, \theta_{k+l}$
    3. call POL($\psi(\lambda), \theta_1, \ldots, \theta_{k+l}$)
    4. perform k steps of orthogonalization process with starting vector
       $v_1 := \psi(\mathbf{A})v_1$ (we obtain (2.3))
    5. if $\|w_k\| < TOL$ then STOP
end i
```

Remark: If the Ritz values are used for the shifts and we want to calculate an approximation to the (say) jth eigenpair of $\mathbf{A}$, $j \leq k$, then the convergence condition in step *5.* is replaced by more accurate one $\|w_k\|\, |e_k^T g_j| < TOL$, where g_j is defined by (1.14). Accuracy of this condition is analyzed in [15].

Remark: Computational difficulties can arise from the fact that if $\|w_k^{(i)}\|$ is "small" (i.e. span$\{\mathbf{V}_k\}$ is near to an invariant subspace of $\mathbf{A}$), then the vector $v_{k+1} := w_k^{(i)}/\|w_k^{(i)}\|$ looses significant digits, see [49]. The same is true if $w_j^{(i)}$ is "small" for some $j < k$. This is demonstrated on Example 3 in Section 4.2.

Now we will discuss the procedure POL. First a few details about the implicitly restarted Arnoldi process with shifts (IRA) are given in Section 2.2. Then the idea to use Tchebychev polynomials for constructing suitable filters is described and worked out in Section 2.3.

2.2 Implicitly restarted Arnoldi process

A well-known technique for constructing suitable polynomial filters was described in [81], see also [49]. It is based on application of shifted QR–algorithm on the matrix $\mathbf{H}_{k+l}$. Let $\mu_1, \ldots, \mu_l$ be shifts and let $\mathbf{G}_1 := \mathbf{H}_{k+l}$, $\mathbf{S}_1 := \mathbf{V}_{k+l}$. Application of the ith shift consists of the following steps. First the QR–decomposition

$(\mathbf{G}_i - \mu_i \mathbf{I}) = \mathbf{Q}_i \mathbf{R}_i$ is computed, where $\mathbf{Q}_i \in R^{(k+l)\times(k+l)}$ is an unitary upper Hessenberg matrix and $\mathbf{R}_i \in R^{(k+l)\times(k+l)}$ is an upper triangular matrix. Then (2.1) multiplied by $\mathbf{Q}_i$ gives succesively

$$(\mathbf{A} - \mu_i \mathbf{I})\mathbf{S}_i - \mathbf{S}_i \mathbf{Q}_i \mathbf{R}_i = w_{k+l} e_{k+l}^T \mathbf{Q}_1 \dots \mathbf{Q}_{i-1}, \tag{2.4}$$

$$\mathbf{A}(\mathbf{S}_i \mathbf{Q}_i) - (\mathbf{S}_i \mathbf{Q}_i)(\mathbf{R}_i \mathbf{Q}_i + \mu_i \mathbf{I}) = w_{k+l} e_{k+l}^T \mathbf{Q}_1 \dots \mathbf{Q}_{i-1} \mathbf{Q}_i. \tag{2.5}$$

Denote by $\mathbf{S}_{i+1} := \mathbf{S}_i \mathbf{Q}_i$ and $\mathbf{G}_{i+1} := \mathbf{R}_i \mathbf{Q}_i + \mu_i \mathbf{I}$. It is easy to show that $\mathbf{G}_{i+1}$ is an upper Hessenberg matrix and, moreover, $\mathbf{G}_{i+1} = \mathbf{Q}_i^T \mathbf{G}_i \mathbf{Q}_i$. Consequent applications of the shifts $\mu_1, \dots, \mu_l$ yields

$$\mathbf{A}\mathbf{S}_{l+1} = \mathbf{S}_{l+1} \mathbf{G}_{l+1} + w_{k+l} e_{k+l}^T \hat{\mathbf{Q}}, \tag{2.6}$$

where $\mathbf{S}_{l+1} := \mathbf{V}_{k+l} \hat{\mathbf{Q}}$, $\mathbf{G}_{l+1} := \hat{\mathbf{Q}}^T \mathbf{H}_{k+l} \hat{\mathbf{Q}}$ and $\hat{\mathbf{Q}} := \mathbf{Q}_1 \mathbf{Q}_2 ... \mathbf{Q}_l$. Denote by $s^{(i)} := \mathbf{S}_i e_1$ for $i = 1, \dots, l$. Multiplying (2.4) by e_1 yields

$$s^{(i+1)} = \mathbf{S}_{i+1} e_1 = \frac{1}{e_1^T \mathbf{R}_i e_1} (\mathbf{A} - \mu_i \mathbf{I}) s^{(i)}$$

and thus

$$v_1^+ := \mathbf{S}_{l+1} e_1 = \frac{1}{\tau} \prod_{i=1}^{l} (\mathbf{A} - \mu_i \mathbf{I}) v_1, \tag{2.7}$$

where $\tau := \prod_{i=1}^{l} e_1^T \mathbf{R}_i e_1$.

Denote by $\mathbf{V}_k^+$ the first k columns of the matrix $\mathbf{S}_{l+1}$ and by $\mathbf{H}_k^+$ the main submatrix of order k of the matrix $\mathbf{G}_{l+1}$. Then (2.6) becomes

$$\mathbf{A}[\mathbf{V}_k^+, \hat{\mathbf{V}}_l] = [\mathbf{V}_k^+, \hat{\mathbf{V}}_l] \begin{bmatrix} \mathbf{H}_k^+ & \mathbf{B} \\ \beta e_1 e_k^T & \mathbf{C} \end{bmatrix} + v_{k+l+1} u^T, \tag{2.8}$$

where $\beta e_1 e_k^T \in R^{l \times k}$, $v_{k+l+1} := w_{k+l} / \|w_{k+l}\|$ and $u^T := \|w_{k+l}\| e_{k+l}^T \hat{\mathbf{Q}}$. From (2.8) and the fact that $(u)_i = 0$ for $i = 1, \dots, k-1$ it follows that

$$\mathbf{A}\mathbf{V}_k^+ = \mathbf{V}_k^+ \mathbf{H}_k^+ + w_k^+ e_k^T, \quad w_k^+ := \hat{\mathbf{V}}_l e_1 \beta + v_{k+l+1} (u)_k. \tag{2.9}$$

The matrix $\mathbf{S}_{l+1}$ has orthonormal columns and $\mathbf{S}_{l+1}^T v_{k+l+1} = \hat{\mathbf{Q}}^T \mathbf{V}_{k+l}^T v_{k+l+1} = 0$. Thus

$$(\mathbf{V}_k^+)^T w_k^+ = (\mathbf{V}_k^+)^T \hat{\mathbf{V}}_l e_1 \beta + (\mathbf{V}_k^+)^T v_{k+l+1} (u)_k = 0$$

and w_k^+ is orthogonal to $\mathbf{V}_k^+$. Summarizing, (2.9) is the Arnoldi factorization of the matrix $\mathbf{A}$ with the starting vector (2.7).

Complete IRA process with shifts is described by the following algorithm.

ALGORITHM 2.2 (IRA with shifts):

```
input: $v_1, k, l, TOL$,
        $max$ – maximal number of iterations
output: $\mathbf{V}_k$ such that $\|w_k\| < TOL$

perform k steps of orthogonalization process with $\mathbf{A}$ and starting vector $v_1$ (we obtain (2.3) for i=0)
for $i = 1, \ldots, max$
    1. perform l additional steps of orthogonalization process (we obtain (2.1))
    2. compute Ritz (resp. harmonic Ritz) values $\theta_1, \ldots, \theta_{k+l}$ and choose shifts $\mu_1, \ldots, \mu_l$
    3. $\mathbf{Q} := \mathbf{I}_{k+l}$
       for $j = 1, .., l$
           compute QR–decomposition $\mathbf{Q}_j\mathbf{R}_j = (\mathbf{H}_{k+l} - \mu_j\mathbf{I}_{k+l})$
           $\mathbf{H}_{k+l} := \mathbf{Q}_j^T\mathbf{H}_{k+l}\mathbf{Q}_j$
           $\mathbf{Q} := \mathbf{Q}\mathbf{Q}_j$
       end j
       $\mathbf{V}_k := [\mathbf{V}_{k+l}\mathbf{Q}] \begin{bmatrix} \mathbf{I}_k \\ \mathbf{0}_{l,k} \end{bmatrix}$
       $\mathbf{H}_k := [\mathbf{I}_k, \mathbf{0}_{l,k}]\,\mathbf{H}_{k+l} \begin{bmatrix} \mathbf{I}_k \\ \mathbf{0}_{l,k} \end{bmatrix}$
    4. compute $w_k$ according to (2.9)
    5. if $||w_k|| < TOL$ then STOP
end i
```

Now the question arises how to choose optimal shifts. The idea is to construct a starting vector v_1^+ being a linear combination of the eigenvectors corresponding to the eigenvalues from the set σ_W. These values are not available, but we can use numbers $\theta_1, \ldots, \theta_{k+l}$. The following theorem from [81] motivates the selection of the shifts.

Theorem 2.1: *Let (θ_i, g_i) for $i = 1, \ldots, k+l$ be the eigenpairs of the matrix $\mathbf{H}_{k+l}$. Then the IRA method with the shifts $\theta_{k+1}, \ldots, \theta_{k+l}$ yields*

$$\mathbf{G}_{l+1} = \begin{bmatrix} \mathbf{H}_k^+ & \mathbf{B} \\ \mathbf{0}_{l,k} & \mathbf{C} \end{bmatrix},$$

where $\mathbf{C}$ is an upper triangular matrix. Moreover, $\mathbf{H}_k^+$ has the eigenvalues $\{\theta_1, .., \theta_k\}$ and

$$v_1^+ = \mathbf{V}_{k+l}\hat{\mathbf{Q}}e_1 = \sum_{j=1}^{k} \alpha_j g_j, \qquad \alpha_j \in \mathbb{R},\ j = 1, \ldots, k.$$

Thus substituting the Ritz values $\theta_{k+1}, \ldots, \theta_{k+l}$ for the shifts, it can be proved that the smallest in magnitude eigenvalues of $\mathbf{H}_{k+l}$ form the spectrum of $\mathbf{H}_k^+$ and that the relation (2.3) holds. Therefore

$$\psi(\lambda) = \frac{1}{\tau} \prod_{j=1}^{l} (\lambda - \theta_{k+j})$$

is a polynomial filter of degree l. We refer to this filter as *classical filter*. Computation of the vector w_k^+ in (2.9) simplifies to

$$w_k^+ = v_{k+l+1}(u)_k = w_{k+l} e_{k+l}^T \hat{\mathbf{Q}} e_k,$$

because $\beta = e_{k+1}^T \mathbf{G}_{l+1} e_k = 0$.

Remark: Even if a Ritz value is complex, the computation described in Algorithm 2.2 can be done in the real arithmetics because eigenvalues of real matrices are complex conjugate. Two consequent applications of complex conjugate shifts $\mu_1, \mu_2 \equiv \bar{\mu}_1$ can be joined into one real step using the fact that the numbers $\mu_1 + \mu_2$ and $\mu_1 * \mu_2$ are real, see [49].

Application of the shifts μ_1, μ_2 on a matrix $\mathbf{H}$ according to Algorithm 2.2 consists of the steps

$$\mathbf{H} - \mu_1 \mathbf{I} = \mathbf{Q}_1 \mathbf{R}_1, \quad \mathbf{G}_2 = \mathbf{R}_1 \mathbf{Q}_1 + \mu_1 \mathbf{I},$$

$$\mathbf{G}_2 - \mu_2 \mathbf{I} = \mathbf{Q}_2 \mathbf{R}_2, \quad \mathbf{G}_3 = \mathbf{R}_2 \mathbf{Q}_2 + \mu_2 \mathbf{I}.$$

Moreover,

$$\mathbf{G}_3 = \mathbf{Q}_2^H \mathbf{G}_2 \mathbf{Q}_2 = \mathbf{Q}_2^H \mathbf{Q}_1^H \mathbf{H} \mathbf{Q}_1 \mathbf{Q}_2.$$

The matrix

$$\mathbf{M} := (\mathbf{H} - \mu_1 \mathbf{I})(\mathbf{H} - \mu_2 \mathbf{I}) = \mathbf{H}^2 - (\mu_1 + \mu_2)\mathbf{H} + \mu_1 \mu_2 \mathbf{I} \qquad (2.10)$$

is real and satisfies $\mathbf{M} = (\mathbf{Q}_1 \mathbf{Q}_2)(\mathbf{R}_2 \mathbf{R}_1)$, that is the QR–decomposition of the real matrix. Thus it is possible to construct the matrix $\mathbf{G}_3$ by a product of real matrices using the following technique. First the matrix $\mathbf{M}$ is computed according to (2.10), then the real QR–decomposition $\mathbf{M} = \mathbf{Z}\mathbf{R}$ is performed and finally it is set $\mathbf{G}_3 = \mathbf{Z}^T \mathbf{H} \mathbf{Z}$. Application of complex conjugate shifts on a general matrix pencil $(\mathbf{A} - \gamma \mathbf{B})$ was studied, e.g., in [70].

2.3 Filters based on Tchebychev polynomials

In this section we work out an alternative technique, which idea was mentioned in [81], for constructing a polynomial filter based on the properties of *Tchebychev polynomials*. We use some results from [52], [53] and [74]. Proposition 1.5

suggests that the invariant subspace can be constructed by replacing the starting vector v_1 by a linear combination of approximate eigenvectors corresponding to the wanted eigenvalues. Thus we try to modify v_1 to be, in the ideal case, a linear combination of the eigenvectors of the matrix $\mathbf{A}$ corresponding to the set σ_W, or more realistically to be a linear combination of the Ritz vectors corresponding to the Ritz values nearest to the eigenvalues from σ_W. The following considerations hold for the Ritz as well as the harmonic Ritz values. Therefore we work only with the Ritz values.

For simplicity we assume for a moment that $\mathbf{A}$ is diagonalizable. Denote by $u_1, \ldots, u_n$ eigenvectors of $\mathbf{A}$, where u_i corresponds to the eigenvalue λ_i. Let $v_1 = \sum_{i=1}^{n} \alpha_i u_i$. For any polynomial p of degree l it holds that

$$v_1^+ \equiv p(\mathbf{A})v_1 = \sum_{i=1}^{n} \alpha_i p(\lambda_i) u_i = \sum_{i=1}^{k} \alpha_i p(\lambda_i) u_i + \sum_{i=k+1}^{n} \alpha_i p(\lambda_i) u_i. \tag{2.11}$$

If $\mathbf{A}$ is defective, the form of $p(\mathbf{A})$ is more complicated and it was described by T. A. Manteuffel [53], [52]. From (2.11) it follows that the vector v_1^+ is a linear combination of the eigenvectors corresponding to σ_W if $p(\lambda) = 0$ for all $\lambda \in \sigma_U$. Unfortunately, we do not exactly know the eigenvalues and eigenvectors of the matrix $\mathbf{A}$, but only their approximations Ritz values $\theta_1, \ldots, \theta_{k+l}$ and Ritz vectors $y_1, \ldots, y_{k+l}$ (eventually the harmonic ones).

Let E be a domain in the complex plane such that

$$0, \theta_1, \ldots, \theta_k \notin E, \quad \theta_{k+1}, \ldots, \theta_{k+l} \in E, \tag{2.12}$$

and let p be a polynomial having the property $\max_{\lambda \in E} |p(\lambda)| < \varepsilon$, where ε is a "small" positive number. Assuming that $\sigma_U \subset E$ and $\sigma_W \not\subset E$, it follows from (2.11) that the vector v_1^+ is "close to" the subspace generated by the vectors $u_1, \ldots, u_k$. Assuming, moreover, that $|\theta_i - \lambda_i| < \varepsilon$ for $i = 1, \ldots, k$ it is reasonable to construct a polynomial p satisfying the condition

$$p \quad = \quad \arg \min_{\tilde{p} \in P_l,\ \tilde{p}(\theta_1)=1} \max_{\lambda \in E} |\tilde{p}(\lambda)|. \tag{2.13}$$

If the matrix $\mathbf{A}$ does not have strongly ill conditioned set of eigenvectors and if we will be successful in finding a polynomial p satisfying (2.13), then the contributions of the unwanted eigenvalues can be damped in the decomposition (2.11) of v_1^+.

First we describe the construction of polynomial filter in case that the matrix $\mathbf{H}_{k+l}$ has only real eigenvalues. Then the general case follows.

Real case

Let $\sigma(\mathbf{H}_{k+l}) \subset \mathbb{R}$. Denote by $C_l(\lambda)$ the real Tchebychev polynomial of the first kind of degree l defined for all $\lambda \in < -1, 1 >$, i.e.

$$C_l(\lambda) = \cos(l * \cos^{-1}(\lambda)).$$

It is well known that this polynomial satisfies the condition

$$C_l = \arg \min_{p \in MP_l} \max_{\lambda \in <-1,1>} |p(\lambda)|.$$

The function $f(\lambda) = 1 + 2\frac{\lambda - \beta}{\beta - \alpha}$ represents the mapping of the interval $< \alpha, \beta >$ on $< -1, 1 >$. Therefore it is easy to see that the transformed and scaled Tchebychev polynomial

$$\hat{C}_l(\lambda) \equiv \frac{C_l\left(1 + 2\frac{\lambda-\beta}{\beta-\alpha}\right)}{C_l\left(1 + 2\frac{\theta_1-\beta}{\beta-\alpha}\right)}, \tag{2.14}$$

satisfies (2.13) with $E =< \alpha, \beta >$, see [69]. The polynomial $\hat{C}_l(\lambda)$ will be taken for a polynomial filter, if the set E is defined as the smallest interval containing $\theta_{k+1}, \ldots, \theta_{k+l}$ such that $\theta_1, \ldots, \theta_k \notin < \alpha, \beta >$. It is easy to construct such interval if all the numbers $\theta_{k+1}, \ldots, \theta_{k+l}$ have the same sign. Then

$$\alpha = \min\{\theta_{k+1}, \theta_{k+l}\} \quad \text{and} \quad \beta = \max\{\theta_{k+1}, \theta_{k+l}\}.$$

In the opposite case, two intervals must be constructed – the interval $< \alpha_1, \beta_1 > \subset \mathbb{R}^-$ for negative and $< \alpha_2, \beta_2 > \subset \mathbb{R}^+$ for positive Ritz values, respectively. The values $\theta_{k+1}, \ldots, \theta_{k+l}$ are then eliminated from the spectrum in two steps. We alternately apply the polynomial filter (2.14) for the intervals $< \alpha_1, \beta_1 >$ and $< \alpha_2, \beta_2 >$. This technique will be called *alternation of intervals.*

The previous considerations lead to the following algorithm which is a special case of the Algorithm 2.1 for application of the Tchebychev filters in the real case. Details about realization of the update $v_1 := \psi(\mathbf{A})v_1$ in step *4.* are given in Section 2.4.

ALGORITHM 2.3 (Tchebychev filter):

```
input: $v_1, k, l, TOL$,
        $max$ – maximal number of iterations
output: $\mathbf{V}_k$ such that $\|w_k\| < TOL$

perform k steps of orthogonalization process with $\mathbf{A}$ and starting vector $v_1$ (we
obtain (2.3) for i=0)
dir = 1
```

```
for $i = 1, \ldots, max$
    1. perform $l$ additional steps of orthogonalization process (we obtain (2.1))
    2. compute Ritz (resp. harmonic Ritz) values $\theta_1, \ldots, \theta_{k+l}$
    3. $\mathcal{M}_1 := \mathcal{M}_{-1} := \emptyset, l_1 := l_{-1} := 0, dir := -dir$
       for $j = 1, \ldots, l$
           if ($\theta_{k+i} > 0$) then $\mathcal{M}_1 := \mathcal{M}_1 \cup \{\theta_{k+i}\}$, $l_1 := l_1 + 1$
           else $\mathcal{M}_{-1} := \mathcal{M}_{-1} \cup \{\theta_{k+i}\}$, $l_{-1} := l_{-1} + 1$
       end $j$
       if ($l_{dir} = 0$) then dir := - dir
       $\alpha := \min\{\theta | \theta \in \mathcal{M}_{dir}\}, \beta := \max\{\theta | \theta \in \mathcal{M}_{dir}\}$
       put $\psi(\lambda) := \hat{C}_{l_{dir}}(\lambda)$
    4. perform $k$ steps of orthogonalization process with $\mathbf{A}$ and starting vector
       $v_1 := \psi(\mathbf{A})v_1$ (we obtain (2.3))
    5. if $||w_k|| < TOL$ then STOP
end $i$
```

Remark: If the alternation of intervals in $\mathbb{R}^+$ and $\mathbb{R}^-$ is used, it may happen that a positive or a negative part of the set $\{\theta_{k+1}, \ldots, \theta_{k+l}\}$ is damped sooner then the other part. Then the iterations simply continue only in the remaining part of the set $\{\theta_{k+1}, \ldots, \theta_{k+l}\}$.

Remark: The intervals does not have to alternate regularly. In each iteration the interval can be chosen according to some prescribed rules - interval containing the largest in magnitude Ritz (or harmonic Ritz) value, interval containing more Ritz (or harmonic Ritz) values or the largest interval, etc.

Complex case

Let $\sigma(\mathbf{H}_{k+l}) \subset \mathbb{C}$. Generally, the Tchebychev polynomials of degree l are defined for all complex λ by the formula

$$C_l(\lambda) = \cosh(l * \cosh^{-1}(\lambda)),$$

where $\cosh(\zeta) = \frac{1}{2}(\mathrm{e}^{\zeta} + \mathrm{e}^{-\zeta})$, $\zeta \in \mathbb{C}$. For more details about the complex Tchebychev polynomials see, e.g., [52], [53], [72], [74]. The set E can be chosen as interior of an ellipse in $\mathbb{C}$ satisfying the condition (2.12). Because eigenvalues of a real matrix are complex conjugate, the center d of the constructed ellipse lies on the real axis and the semi-major axis is parallel with the real axis or the imaginary axis. Let a and e be the lengths of the semi-major axis and the eccentricity of an ellipse respectively. If the semi-major axis is parallel with the imaginary axis we define a and e as purely complex numbers. Denote by $E(d, a, e)$ the ellipse with the above defined parameters and for the set E consider the inner area of

$E(d, a, e)$ including its boundary. Then it can be proved that the polynomial

$$\hat{C}_l(\lambda) \equiv \frac{C_l\left(\frac{\lambda-d}{e}\right)}{C_l\left(\frac{\theta_1-d}{e}\right)} \tag{2.15}$$

is the near-best polynomial satisfying (2.13), see [74] pp. 144–148, and thus the polynomial filter in the complex case is also constructed. Similarly to the real case, (2.15) is transformed and scaled complex Tchebychev polynomial. The transformation represents the mapping of the ellipse with the center d and the eccentricity e on the circle with the center in zero and radius equal to 1.

If $\theta_{k+1}, \ldots, \theta_{k+l}$ do not lie in one complex half-plane, two ellipses are constructed - for the Ritz values with a real part in $\mathbb{R}^+$ and in $\mathbb{R}^-$, respectively. Then we can alternately apply polynomial filters for these two ellipses analogously as in the real case. Algorithm of this process is similar as Algorithm 2.3, where only construction of the interval $<\alpha, \beta>$ is replaced by the construction of the ellipse and filter (2.14) is replaced by (2.15).

Note that techniques for construction of the minimal ellipse containing prescribed complex numbers in the interior has been widely described in [53], [52] and can also be found in [74]. The main idea is to find a minimal positive hull of the prescribed numbers with vertexes in some of them. Using the symmetry of the hull with respect to the real axis, parameters d, a, e of the optimal ellipse containing this hull in the interior are computed.

Remark: There are still some open questions about the construction of an optimal ellipse satisfying (2.12) in some special cases, e.g., if a Ritz (or harmonic Ritz) value has a very big imaginary part and a small real part. The problem arises from combination of two facts. The first is that the minmax property of polynomial (2.15) is proved only for ellipses with the center on the real axis and the second is that zero must not lye in the inner area of the ellipse.

Remark: If the eccentricity e is a real number (i.e. the semi-major axis of the ellipse lies on the real axis) the real Tchebychev polynomial can be used for construction of the polynomial filter (2.15). Thus the polynomial filter is real and computation can be simplified.

2.4 Application of polynomial filters

Two methods for constructing a polynomial satisfying (2.13) were discussed in the previous text. Taking this polynomial as a filter (denoted by ψ) we have to decide how to apply a filter ψ to update the starting vector v_1 of the Arnoldi factorization. Direct calculation of $v_1^+ = \psi(\mathbf{A})v_1$ is very expensive because many matrix-vector multiplications must be computed. Multiplication of a vector by a matrix polynomial can also be performed by a Horner's scheme, see [41], [69]. But the most suitable and very cheap opportunity is to use the IRA process described in Section 2.2 in which the roots of the polynomial ψ are chosen for the shifts. Let $\psi(\lambda) \in P_l$ and denote by $\nu_1, \ldots, \nu_l$ it's roots. Then

$$\psi(\mathbf{A}) = \prod_{i=1}^{l} (\mathbf{A} - \nu_i \mathbf{I})$$

and (2.7) yields that Algorithm 2.2 with shifts $\nu_1, \ldots, \nu_l$ computes the factorization (2.3) with the starting vector $\psi(\mathbf{A})v_1$. This shows another advantage of using IRA with shifts compared to the Horner's scheme - we obtain the updated Arnoldi factorization of order k implicitly.

Now we will concentrate on application of the polynomial filters presented in the previous section. Consider the real Tchebychev polynomial filter (2.14). Denote by φ_i for $i = 1, \ldots, l$ the roots of the real Tchebychev polynomial in the interval $< -1, 1 >$, i.e.

$$\varphi_i = \cos\left(\frac{\pi}{l}\left(i - \frac{1}{2}\right)\right).$$

Using the inverse mapping $g(\lambda) = (\lambda - 1)\frac{\beta-\alpha}{2} + \beta$ of the interval $< -1, 1 >$ on $< \alpha, \beta >$, it is easy to verify that

$$\nu_i := (\varphi_i - 1) * \frac{\beta - \alpha}{2} + \beta, \quad i = 1, \ldots, l$$

are the roots of (2.14) lying in $< \alpha, \beta >$. Similarly, in the complex case the polynomial filter is determined by equation (2.15) and its roots are

$$\nu_i := \varphi_i * e + d, \quad i = 1, \ldots, l,$$

where φ_i for $i = 1, \ldots, l$ are roots of the Tchebychev polynomial constructed in the complex case. Note that here the numbers φ_i can be complex, but not necessary complex conjugate. Thus application of these shifts have to be carried out in complex arithmetic. This fact complicates the computation.

Remark: If eccentricity e of an ellipse is real, the roots of the real Tchebychev polynomial can be used for calculation of the roots of the corresponding Tchebychev filter (2.15), see [74]. Thus only real arithmetic is needed for application of such polynomial filter by the IRA process.

Chapter 3

Convergence behavior

In Chapter 2 we have described a technique for updating the Arnoldi factorization in order to obtain a matrix $\mathbf{V}$ *whose columns generate an invariant subspace of* $\mathbf{A}$*. The core of this method is in updating the starting vector of the Arnoldi process by a special polynomial - polynomial filter. In this chapter we concentrate on convergence behavior of the given algorithms. Putting assumptions on the polynomial filter, we show that* $\|w_k\|$ *converges to zero. Then we prove the convergence of the updated starting vector to the searched invariant subspace, assessed by the magnitude of the angle between the updated vector and the subspace. In practice, usually an unreduced and therefore nonderogatory Hessenberg matrix is obtained. The question arises what happens if the smallest eigenvalue of* $\mathbf{A}$ *has geometric multiplicity greater than one. Thus, in the last section, we present some convergence results for derogatory and/or defective matrix* $\mathbf{A}$*.*

3.1 Convergence to invariant subspace

The repeated application of polynomial filters according to (2.2) updates the matrices $\mathbf{V}_k, \mathbf{H}_k$ and the residual term $w_k e_k^T$ in the Arnoldi factorization. Convergence of this process was proved in [81] in case that $\mathbf{A}$ is symmetric or the polynomial filter is fixed, i.e. $\psi_i = \psi \quad \forall\ i$, see also [49]. The presented analysis does not apply to adaptive algorithms, but it gives an indication how these might behave near the final stages of the computation where polynomial filters tend to become stationary. This motivates us to assume polynomial filters that change in every iteration but converge to some polynomial.

We consider a general nonsymmetric matrix $\mathbf{A}$ and put assumptions only on polynomial filters. Assumptions of presented theorems are discussed at the end of this section.

Let the upper index (j) denote the jth repetition in (2.2), $v_1^{(0)} := v_1$. Denote by $\mu_1^{(j)}, \dots, \mu_l^{(j)}$ the shifts applied in the jth step of Algorithm 2.2 and let

$$\psi_j(\lambda) := \frac{1}{\tau_j} \prod_{m=1}^{l} (\lambda - \mu_m^{(j)}), \quad \Psi_j(\lambda) := \prod_{i=1}^{j} \tau_i \psi_i(\lambda).$$

Then

$$v_1^{(j)} \; = \; \psi_j(\mathbf{A}) v_1^{(j-1)} \; = \; \prod_{i=1}^{j} \frac{1}{\tau_i} \;\; \Psi_j(\mathbf{A}) v_1,$$

where the numbers $\tau_i > 0$ are chosen so that $\|v_1^{(i)}\| \; = 1$ for $i = 1, 2, \dots$

In this section we assume:

Assumption 1: Let $\epsilon > 0$ and complex numbers $\mu_1, \dots, \mu_l$ exist such that

$$\{\mu_1, \dots, \mu_l\} \; \cap \; \cup_{i=1}^{n} \mathcal{U}_\epsilon(\lambda_i) = \emptyset, \tag{3.1}$$

$$\sum_{j=1}^{\infty} |\mu_m - \mu_m^{(j)}| < \infty, \quad m = 1, \dots, l, \tag{3.2}$$

where $\{\mu_m^{(j)}\}_{j=1}^{\infty}$ for $m = 1, \dots, l$ are sequences of shifts. If $\mu_{m_1} \in \mathbb{C}$ then there exists an index $m_2 \in \{1, \dots, l\}$ such that $\mu_{m_1} = \bar{\mu}_{m_2}$ and $\mu_{m_1}^{(j)} = \bar{\mu}_{m_2}^{(j)}$ $\forall j$. □
Note that the last condition yields that the vectors $v_1^{(j)}$ are real.

Assumption 2: Let the polynomial $\psi(\lambda) = \prod_{i=1}^{l} (\lambda - \mu_i)$ satisfy the condition

$$M_1 := \min_{i=1,..,k} |\psi(\lambda_i)| \; > \max_{i=k+1,..,n} |\psi(\lambda_i)| =: M_2 \tag{3.3}$$

and denote by $\delta := M_1 - M_2$ and $\gamma := \frac{M_2 + \delta/3}{M_1 - \delta/3} < 1$. □

Let the columns of the matrix $\mathbf{X}_1 \in \mathbb{C}^{n \times k}$, $\mathbf{X}_1^H \mathbf{X}_1 = \mathbf{I}$ span the eigenspace of the matrix $\mathbf{A}$ corresponding to σ_W. Denote by $\mathcal{X}_1 = span\{\mathbf{X}_1\}$. The space $\mathcal{X}_1$ is simple due to $\sigma_U \cap \sigma_W = \emptyset$ (see [82] pp. 219–225). Hence a decomposition

$$\mathbf{A}[\mathbf{X}_1, \mathbf{X}_2] = [\mathbf{X}_1, \mathbf{X}_2] \begin{bmatrix} \mathbf{L}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_2 \end{bmatrix} \tag{3.4}$$

exists such that $\sigma(\mathbf{L}_1) = \sigma_W$ and $\sigma(\mathbf{L}_2) = \sigma_U$. Apparently, for any polynomial p the equality

$$p(\mathbf{A})[\mathbf{X}_1, \mathbf{X}_2] = [\mathbf{X}_1 p(\mathbf{L}_1), \mathbf{X}_2 p(\mathbf{L}_2)] \tag{3.5}$$

holds. The columns of the matrix $[\mathbf{X}_1, \mathbf{X}_2]$ form a basis in $\mathbb{C}^n$. Thus there exist vectors $t_1 \in \mathbb{C}^k, t_2 \in \mathbb{C}^{n-k}$ such that $v_1 = \mathbf{X}_1 t_1 + \mathbf{X}_2 t_2$. From (3.5) it follows that

$$\prod_{i=1}^{j} \tau_i \; v_1^{(j)} = \Psi_j(\mathbf{A}) v_1 = \mathbf{X}_1 \Psi_j(\mathbf{L}_1) t_1 + \mathbf{X}_2 \Psi_j(\mathbf{L}_2) t_2. \tag{3.6}$$

Assumption 3: Let $v_1 \notin \mathcal{X}_2 := span\{\mathbf{X}_2\}$. □

The following two lemmas will be used to proof the convergence theorems.

Lemma 3.1: *There exist a positive number $q < 1$ and a constant $C > 0$ such that*

$$\|\Psi_j(\mathbf{L}_2)\| \leq C\, q^j (M_1 - \delta/3)^j. \tag{3.7}$$

Proof. Since $\tau_j\psi_j$ converges to ψ as $j \to \infty$, there exists $j_0 \in N^+$ such that for each $j > j_0$ the inequalities

$$\min_{i=1,..,k} |\tau_j\psi_j(\lambda_i)| > M_1 - \frac{\delta}{3} > M_2 + \frac{\delta}{3} > \max_{i=k+1,..,n} |\tau_j\psi_j(\lambda_i)| \tag{3.8}$$

hold. According to Assumption 3 the vector $v_1^{(j_0)} \notin \mathcal{X}_2$. Therefore, without any loss of generality, we assume $j_0 = 0$. According to (3.1), $\psi(\mathbf{L}_2)$ is nonsingular and for each i we have

$$\begin{aligned}\tau_i\psi_i(\mathbf{L}_2) \quad &= \prod_{m=1}^{l} (\mathbf{L}_2 - \mu_m^{(i)}\mathbf{I}) = \prod_{m=1}^{l} (\mathbf{L}_2 - \mu_m\mathbf{I})\, [\mathbf{I} + (\mu_m - \mu_m^{(i)})(\mathbf{L}_2 - \mu_m\mathbf{I})^{-1}] \\ &= \psi(\mathbf{L}_2) \prod_{m=1}^{l} [\mathbf{I} + (\mu_m - \mu_m^{(i)})(\mathbf{L}_2 - \mu_m\mathbf{I})^{-1}].\end{aligned}$$

Using this fact we obtain

$$\frac{\|\Psi_j(\mathbf{L}_2)\|}{(M_1 - \delta/3)^j} = \Big\| \prod_{i=1}^{j} \frac{\tau_i\psi_i(\mathbf{L}_2)}{M_1 - \delta/3} \Big\| \leq$$

$$\leq \Big\| \left(\frac{\psi(\mathbf{L}_2)}{M_1 - \delta/3} \right)^j \Big\| \prod_{m=1}^{l} \prod_{i=1}^{j} \|\mathbf{I} + (\mu_m - \mu_m^{(i)})(\mathbf{L}_2 - \mu_m\mathbf{I})^{-1}\|.$$

It follows from the assumption (3.2) that for each $m = 1, \ldots, l$ there exists a constant $c_m > 0$ independent from j such that

$$\prod_{i=1}^{j} \|\mathbf{I} + (\mu_m - \mu_m^{(i)})(\mathbf{L}_2 - \mu_m\mathbf{I})^{-1}\| \leq \prod_{i=1}^{j} \big(1 + |\mu_m - \mu_m^{(i)}|\|(\mathbf{L}_2 - \mu_m\mathbf{I})^{-1}\|\big) \leq c_m.$$

Therefore

$$\frac{\|\Psi_j(\mathbf{L}_2)\|}{(M_1 - \delta/3)^j} \leq C_1 \Big\| \left(\frac{\psi(\mathbf{L}_2)}{M_1 - \delta/3} \right)^j \Big\|,$$

where $C_1 := \prod_{m=1}^{l} c_m$. The matrix $\mathbf{L}_2$ has the eigenvalues $\lambda_{k+1}, \ldots, \lambda_n$ and from (3.3) and (3.8) it follows that the spectral radius $\rho\left(\frac{\psi(\mathbf{L}_2)}{M_1 - \delta/3}\right) < \gamma$. Choose $\vartheta > 0$

such that $q := \gamma + \vartheta < 1$. There exists a multiplicative matrix norm $\|.\|_*$ such that $\|\frac{\psi(\mathbf{L}_2)}{M_1 - \delta/3}\|_* < \gamma + \vartheta = q$. The equivalence of both matrix norms implies the existence of a constant $C > 0$ such that

$$\frac{\|\Psi_j(\mathbf{L}_2)\|}{(M_1 - \delta/3)^j} \leq C_1 \Big\| \left(\frac{\psi(\mathbf{L}_2)}{M_1 - \delta/3}\right)^j \Big\| \leq C \Big\| \left(\frac{\psi(\mathbf{L}_2)}{M_1 - \delta/3}\right)^j \Big\|_*$$

$$\leq C \left\| \frac{\psi(\mathbf{L}_2)}{M_1 - \delta/3} \right\|_*^j \leq C\, q^j.$$

Multiplying the last inequality by $(M_1 - \delta/3)^j$ yields (3.7). □

Lemma 3.2: *The inequality*

$$\|\Psi_j(\mathbf{L}_1)\| > (M_1 - \delta/3)^j$$

holds.

Proof. It follows from (3.8) that $|\Psi_j(\lambda_i)| > (M_1 - \delta/3)^j \;\; \forall\, i = 1, \ldots, k$. Hence

$$\|\Psi_j(\mathbf{L}_1)\| \geq \rho(\Psi_j(\mathbf{L}_1)) = \max_{i=1,..,k} |\Psi_j(\lambda_i)| > (M_1 - \delta/3)^j.$$

□

Assumption 4: Let

$$\limsup_{j\to\infty} \frac{\|\Psi_j(\mathbf{L}_1)\|}{\|\Psi_j(\mathbf{A})v_1\|} < \infty. \tag{3.9}$$

□

Now we have prepared everything for proofs of fundamental convergence theorems.

Theorem 3.3: *Let Assumption 1-4 be valid. Then*

$$\lim_{j\to\infty} \sin \angle(v_1^{(j)}, \mathcal{X}_1) = 0. \tag{3.10}$$

Proof. By Lemma 3.1 and 3.2 we obtain the relation

$$\frac{\|\Psi_j(\mathbf{L}_2)\|}{\|\Psi_j(\mathbf{L}_1)\|} < C\, \frac{(M_1 - \delta/3)^j q^j}{(M_1 - \delta/3)^j} = C\, q^j \overset{j\to\infty}{\to} 0. \tag{3.11}$$

It follows from (3.6) that

$$\begin{aligned} v_1^{(j)} &= \frac{1}{\prod_{i=1}^j \tau_i}\, (\mathbf{X}_1 \Psi_j(\mathbf{L}_1) t_1 + \mathbf{X}_2 \Psi_j(\mathbf{L}_2) t_2) \\ &= \frac{\|\Psi_j(\mathbf{L}_1)\|}{\prod_{i=1}^j \tau_i} \Big[\mathbf{X}_1 \frac{\Psi_j(\mathbf{L}_1)}{\|\Psi_j(\mathbf{L}_1)\|} t_1 + \mathbf{X}_2 \frac{\Psi_j(\mathbf{L}_2)}{\|\Psi_j(\mathbf{L}_1)\|} t_2\Big]. \end{aligned} \tag{3.12}$$

For the second term in (3.12) we have from (3.11)

$$\left\| \mathbf{X}_2 \frac{\Psi_j(\mathbf{L}_2)}{\|\Psi_j(\mathbf{L}_1)\|} t_2 \right\| \le C \|\mathbf{X}_2\| \, \|t_2\| \, q^j. \tag{3.13}$$

From Assumption 4 it follows that a positive number D exists such that

$$\frac{\|\Psi_j(\mathbf{L}_1)\|}{\prod_{i=1}^{j} \tau_i} < D \;\; \forall j. \tag{3.14}$$

Hence we obtain from (3.12) and (3.13)

$$\|v_1^{(j)} - \mathbf{X}_1 d_j\| \le CD \|\mathbf{X}_2\| \, \|t_2\| \, q^j, \tag{3.15}$$

where

$$d_j := \frac{\|\Psi_j(\mathbf{L}_1)\|}{\prod_{i=1}^{j} \tau_i} \frac{\Psi_j(\mathbf{L}_1) t_1}{\|\Psi_j(\mathbf{L}_1)\|} = \frac{\Psi_j(\mathbf{L}_1) t_1}{\prod_{i=1}^{j} \tau_i}.$$

Let $P_{\mathcal{X}_1}$ be the orthogonal projection on $\mathcal{X}_1$. From inequality (3.15) we obtain

$$\|(\mathbf{I} - P_{\mathcal{X}_1}) v_1^{(j)}\| \le \|v_1^{(j)} - \mathbf{X}_1 d_j\| \le CD \, \|\mathbf{X}_2\| \, \|t_2\| \, q^j \xrightarrow{j \to \infty} 0.$$

The last inequality yields the assertion of the theorem due to the fact that $\|(\mathbf{I} - P_{\mathcal{X}_1}) v_1^{(j)}\| = \sin \angle(v_1^{(j)}, \mathcal{X}_1)$. □

Remark: Let us mention that the projection $P_{\mathcal{X}_1} = \mathbf{X}_1(\mathbf{X}_1^H \mathbf{X}_1)^{-1} \mathbf{X}_1^H = \mathbf{X}_1 \mathbf{X}_1^H$ is represented by a real matrix and hence $(\mathbf{I} - P_{\mathcal{X}_1}) v_1^{(j)}$ is a real vector. This statement can be easily proved using the fact that if the real matrix $\mathbf{A}$ has an eigenvector or principal vector with complex components belonging to $\mathcal{X}_1$, then the vector with complex conjugate components is also an eigenvector or principal vector of $\mathbf{A}$ corresponding to a wanted eigenvalue and thus it belongs also to $\mathcal{X}_1$.

Theorem 3.4: *Let Assumption 1-4 be valid. Let a positive number ω exist such that, in (2.3), $e_{i+1}^T \mathbf{H}_k^{(j)} e_i = h_{i+1,i}^{(j)} > \omega$ for $i = 1, \ldots, k-1$ and for all j. Then*

$$\lim_{j \to \infty} \|w_k^{(j)}\| = 0. \tag{3.16}$$

Proof. According to Assumption 3 the vector $t_1 \neq 0$. Define $t_1^{(j)} := \Psi_j(\mathbf{L}_1) t_1$, $t_2^{(j)} := \Psi_j(\mathbf{L}_2) t_2$. Combining (1.24) with (1.28) yields the well known formula

$$\pi_j := \Pi_{i=1}^{k} h_{i+1,i}^{(j)} = \min_{p \in MP_k} \|p(\mathbf{A}) v_1^{(j)}\|.$$

In the following, we prove the convergence of π_j to zero and this fact will immediately give the statement of the theorem.

For the characteristic polynomial $\hat{p}_k(\lambda)$ of the matrix $\mathbf{L}_1$, we obtain from (3.6)

$$\begin{aligned} \pi_j \|\Psi_j(\mathbf{A})v_1\| \quad & \leq \ \|\hat{p}_k(\mathbf{A})\prod_{i=1}^{j}\tau_i \ v_1^{(j)}\| = \ \|\hat{p}_k(\mathbf{A})(\mathbf{X}_1 t_1^{(j)} + \mathbf{X}_2 t_2^{(j)})\| = \\ & = \ \|\hat{p}_k(\mathbf{A})\mathbf{X}_2 t_2^{(j)}\|, \end{aligned} \tag{3.17}$$

because $\hat{p}_k(\mathbf{A})\mathbf{X}_1 = \mathbf{X}_1\hat{p}_k(\mathbf{L}_1) = 0$ and $\|\Psi_j(\mathbf{A})v_1\| = \prod_{i=1}^{j}\tau_i$. Dividing both sides of (3.17) by $(M_1 - \delta/3)^j$ yields

$$\pi_j \frac{\|\Psi_j(\mathbf{A})v_1\|}{(M_1-\delta/3)^j} \leq \frac{\|\hat{p}_k(\mathbf{A})\mathbf{X}_2 t_2^{(j)}\|}{(M_1-\delta/3)^j} \leq \|\hat{p}_k(\mathbf{A})\mathbf{X}_2\| \ \|t_2\| \ \frac{\|\Psi_j(\mathbf{L}_2)\|}{(M_1-\delta/3)^j}.$$

According to Lemma 3.1 a positive constant C_2 exists such that

$$\pi_j \frac{\|\Psi_j(\mathbf{A})v_1\|}{(M_1-\delta/3)^j} \ \leq \ C_2\, q^j, \tag{3.18}$$

where $C_2 := C\|\bar{p}_k(\mathbf{A})\mathbf{X}_2\|\|t_2\|$. Moreover,

$$\frac{\|\Psi_j(\mathbf{A})v_1\|}{(M_1-\delta/3)^j} = \frac{\|\mathbf{X}_1\Psi_j(\mathbf{L}_1)t_1 + \mathbf{X}_2\Psi_j(\mathbf{L}_2)t_2\|}{(M_1-\delta/3)^j} \geq \frac{\|\mathbf{X}_1\Psi_j(\mathbf{L}_1)t_1\|}{(M_1-\delta/3)^j} - \frac{\|\mathbf{X}_2\Psi_j(\mathbf{L}_2)t_2\|}{(M_1-\delta/3)^j}.$$

The matrix $\mathbf{X}_1$ has orthonormal columns and therefore we can remove $\mathbf{X}_1$ on the right-hand side of the last inequality. Define $\mathbf{B}_j := \Psi_j(\mathbf{L}_1)/(M_1-\delta/3)^j$. It follows from (3.8) that all eigenvalues of $\mathbf{B}_j$ are greater than 1. Therefore $\rho(\mathbf{B}_j^{-1}) < 1$ and choosing $\zeta > 0$ such that $\rho(\mathbf{B}_j^{-1}) + \zeta < 1$, there exists a multiplicative matrix norm $\|.\|_+$ and a constant $C_3 > 0$ such that

$$\|\mathbf{B}_j^{-1}\| \leq C_3\|\mathbf{B}_j^{-1}\|_+ < C_3(\rho(\mathbf{B}_j^{-1}) + \zeta) < C_3.$$

Because $t_1 \neq 0$

$$\left\|\frac{\Psi_j(\mathbf{L}_1)t_1}{(M_1-\delta/3)^j}\right\| = \|\mathbf{B}_j t_1\| \geq \|t_1\| \ \frac{1}{\|\mathbf{B}_j^{-1}\|} > C_4,$$

where $C_4 := (1/C_3)\|t_1\|$. Using the fact that $q < 1$ there exists $j_1 \in N^+$ such that

$$(\|\mathbf{X}_2\| \ \|t_2\|C)q^j \ \leq \ C_4/2$$

for all $j > j_1$. According to Lemma 3.1 we have $\frac{\|\mathbf{X}_2\Psi_j(\mathbf{L}_2)t_2\|}{(M_1-\delta/3)^j} \leq C_4/2$ and thus

$$\frac{\|\Psi_j(\mathbf{A})v_1\|}{(M_1-\delta/3)^j} \geq C_4 - C_4/2 = C_4/2$$

for all $j > j_1$. From (3.18) we obtain

$$\pi_j \leq C_5\, q^j, \tag{3.19}$$

where $C_5 := 2C_2/C_4$. We have defined $\pi_j = \|w_k^{(j)}\| \, \Pi_{i=1}^{k-1} h_{i+1,i}^{(j)}$. According to the assumptions of the theorem the inequality $\pi_j > \omega^{k-1}\|w_k^{(j)}\|$ holds. Combining this result with (3.19) yields

$$\|w_k^{(j)}\| < \frac{C_5}{\omega^{k-1}}\, q^j$$

and hence (3.16) immediately follows. □

According to the above formulated assumption $h_{i+1,i}^{(j)} > \omega$ for $i = 1, \dots, k-1$ and for all j, the matrices $\mathbf{H}_k^{(j)}$ are unreduced and $\dim \mathcal{K}_k(\mathbf{A}, v_1^{(j)}) = k$ for all j. Theorem 3.3 states that $v_1^{(j)}$ lies arbitrary near to the space $\mathcal{X}_1$, which is the eigenspace of the matrix $\mathbf{A}$ corresponding to the k smallest eigenvalues and therefore an invariant subspace of $\mathbf{A}$. Consequently, the vectors $\mathbf{A}^q v_1^{(j)}$ for $q = 1, 2, \dots$ lie approximately in the space $\mathcal{X}_1$. Using the fact that $\mathcal{X}_1$ and $\mathcal{K}_k(\mathbf{A}, v_1^{(j)})$ have the same dimension k and $\mathcal{K}_k(\mathbf{A}, v_1^{(j)}) = \operatorname{span}\{v_1^{(j)}, \mathbf{A}v_1^{(j)}, \dots, \mathbf{A}^{k-1}v_1^{(j)}\}$ (if the set σ_W contains at least one complex eigenvalue, the span is considered over $\mathbb{C}^n$), we can see that the space $\mathcal{K}_k(\mathbf{A}, v_1^{(j)})$ is near to the searched eigenspace $\mathcal{X}_1$. The columns of the matrix $\mathbf{V}_k^{(j)}$ span the Krylov subspace $\mathcal{K}_k(\mathbf{A}, v_1^{(j)})$ and hence approximate the basis of the invariant subspace of $\mathbf{A}$ corresponding to the smallest in magnitude eigenvalues. The situation when a matrix $\mathbf{H}_k^{(j)}$ for some j is not unreduced, i.e. the assumption $h_{i+1,i}^{(j)} > \omega$ for $i = 1, \dots, k-1$ is not fulfilled, is discussed in Section 3.2.

Discussion on assumptions

Now we concentrate on polynomial filters described in Chapter 2, i.e. the classical filter (see Section 2.2) and the Tchebychev filter (see Section 2.3), and discuss verification of Assumptions 1-4 for these filters.

In Assumption 1, we have assumed that $\epsilon > 0$ exists such that

$$\{\mu_1, \dots, \mu_l\} \;\cap\; \cup_{i=1}^{n} \mathcal{U}_\epsilon(\lambda_i) = \emptyset.$$

Obviously, this condition is fulfilled for σ_W. The polynomials ψ_j are constructed such that a set of $k + l$ (harmonic) Ritz values is computed, the k smallest in magnitude values (hopefully approximating some small eigenvalues of $\mathbf{A}$) are separated, and the remaining l values are used for construction of the roots $\mu_i^{(j)}$ of the polynomial ψ_j. Therefore $\mu_i \neq \lambda_j$ for $j = 1, \dots, k$ and $i = 1, \dots, l$ is expected to be true. Fulfillment of this condition for $j = k+1, .., n$ is not so clear. In fact, we assume that the roots of constructed polynomials are uncorrelated with the eigenvalues of $\mathbf{A}$. If ψ_j is the classical filter, its roots are the largest (harmonic) Ritz values and these values can sometimes converge to some eigenvalues from σ_U. On the other hand, this assumption is usually fulfilled for the Tchebychev filter and this can be observed on many numerical examples. We document it on Example 1 in Chapter 4. Further, we think that it will be possible to remove this assumption (used to prove Lemma 3.1) in the future.

In Assumption 2, we have assumed that

$$\min_{i=1,..,k} |\psi(\lambda_i)| \; > \; \max_{i=k+1,..,n} |\psi(\lambda_i)|.$$

It follows directly from the construction of filters, that this condition is satisfied for the classical filter, and for the Tchebychev filter in case that alternation of intervals is not used. If the alternation is used, the minimality of Tchebychev filter is ensured only on large (harmonic) Ritz values in $\mathbb{R}^+$ or $\mathbb{R}^-$, respectively. Thus Assumption 2 is satisfied alternately for eigenvalues of $\mathbf{A}$ in $\mathbb{R}^+$ or $\mathbb{R}^-$ during the iteration process. Nevertheless, this fact does not usually induce problems with convergence in practice.

Assumption 3 is very natural and requires that the first starting vector v_1 has a nonzero component in eigenspace corresponding to σ_W.

Assumption 4 is expected to be fulfilled, because the polynomial Ψ_j is constructed such that it damps the components of v_1 corresponding to σ_U, i.e. the updated vector $\Psi_j(\mathbf{A})v_1$ has large components in eigenspaces corresponding to $\sigma_W = \sigma(\mathbf{L}_1)$. Numerical experiments also support this statement.

Remark: If the matrix $\mathbf{A}$ is strongly nonnormal problems with convergence of Algorithm 2.2 may appear. In this case the (harmonic) Ritz values do not have to approximate the eigenvalues of $\mathbf{A}$, actually they can lie far from the spectrum of the matrix $\mathbf{A}$. Moreover, even if the process converges in the sense of Theorem 3.4, the resulting matrix $\mathbf{V}_k$ sometimes does not have to span the desired invariant subspace, see the discussion in Section 1.5.

3.2 Generalization for derogatory matrices

Now we introduce some terminology, that will be used in this section.

Definition 3.1: *Let λ be an eigenvalue of a matrix $\mathbf{A}$. Then*

- *λ has algebraic multiplicity p, if it is a root of multiplicity p of the characteristic polynomial of $\mathbf{A}$.*
- *λ has geometric multiplicity p, if the maximal number of independent eigenvectors associated with it is equal to p.*

The eigenvalue λ is called

- *simple, if it has algebraic multiplicity one. Otherwise it is called multiple.*
- *semisimple, if its algebraic multiplicity is equal to its geometric multiplicity. Otherwise it is called defective.*

Definition 3.2: *A matrix $\mathbf{A}$ is called*

- *derogatory, if geometric multiplicity of at least one of its eigenvalues is larger than one.*
- *defective, if at least one of its eigenvalues is defective.*

We return to the assumption $h^{(j)}_{i+1,i} > \omega$ for $i = 1, \ldots, k-1$ and for all j of Theorem 3.4. It can be proved, see [49], that if $\mathbf{H}^{(j)}_k$ is unreduced then it is nonderogatory, i.e. geometric multiplicity of each of its eigenvalues is equal to one. The question arises what happens if the geometric multiplicity of some wanted eigenvalue of $\mathbf{A}$ is greater than one and we want to construct the whole invariant subspace corresponding to this eigenvalue. We ask, how the sequences $h^{(j)}_{2,1}$, $h^{(j)}_{3,2}$,..., $h^{(j)}_{k+1,k}$ for $j = 1, 2, ...$ behave in this case. In the following part, we prove that when the Jordan canonical form of $\mathbf{A}$ has $s > 1$ blocks corresponding to λ_1 with maximal dimension $d < k$, then an integer j exists such that the vectors $v^{(j)}_1, \mathbf{A}v^{(j)}_1, \ldots, \mathbf{A}^d v^{(j)}_1$ are almost linearly dependent. Hence the matrix $\mathbf{H}^{(j)}_k$ is not unreduced and the Arnoldi or Householder process stops after at most $(d+1)$ steps. The exact formulation is contained in Theorem 3.8 and 3.9. Note that the presented results can be reformulated for any small multiple eigenvalue. We consider the multiplicity of λ_1 to simplify the explanation only.

It is easy to estimate from (3.13)

$$\left\| \mathbf{X}_1 \frac{\Psi_j(\mathbf{L}_1)}{\|\Psi_j(\mathbf{L}_1)\|} t_1 \;+\; \mathbf{X}_2 \frac{\Psi_j(\mathbf{L}_2)}{\|\Psi_j(\mathbf{L}_1)\|} t_2 \right\| \;\leq\; \|\mathbf{X}_1\| \, \|t_1\| \;+\; C \, \|\mathbf{X}_2\| \, \|t_2\| \, q^j. \tag{3.20}$$

Denote by α the term on the right hand side of the previous inequality. Using (3.12) we obtain $1/\alpha \leq \|\Psi_j(\mathbf{L}_1)\|/(\prod_{i=1}^j \tau_i)$. Combining this with (3.14) immediately yields

$$\liminf_{j\to\infty} \frac{\|\Psi_j(\mathbf{L}_1)\|}{\prod_{i=1}^j \tau_i} > 0, \qquad \limsup_{j\to\infty} \frac{\|\Psi_j(\mathbf{L}_1)\|}{\prod_{i=1}^j \tau_i} < \infty. \tag{3.21}$$

Moreover, from(3.12) and (3.13) we have

$$v^{(j)}_1 = \mathbf{X}_1 \frac{\Psi_j(\mathbf{L}_1)}{\prod_{i=1}^j \tau_i} t_1 + O(q^j), \tag{3.22}$$

where the last term denotes a vector with all components $O(q^j)$.

To analyze the behavior of polynomial filters in this special case, the matrix $\mathbf{L}_1$ must be transformed to the Jordan form, i.e. $\mathbf{J}_1 := \mathbf{Z}^{-1}\mathbf{L}_1\mathbf{Z}$. From (3.4) we have

$$\mathbf{A}\mathbf{Z}_1 = \mathbf{Z}_1\mathbf{J}_1 \quad \text{and} \quad \Psi_j(\mathbf{A})\mathbf{Z}_1 = \mathbf{Z}_1\Psi_j(\mathbf{J}_1), \tag{3.23}$$

where $\mathbf{Z}_1 := \mathbf{X}_1\mathbf{Z}$. The columns of $\mathbf{Z}_1$ are eigenvectors and principle vectors of $\mathbf{A}$ corresponding to the wanted eigenvalues. Apparently, the inequalities (3.21) are valid if we substitute $\mathbf{J}_1$ instead of $\mathbf{L}_1$. From (3.22) we obtain

$$v_1^{(j)} = \mathbf{Z}_1 \frac{\Psi_j(\mathbf{J}_1)}{\prod_{i=1}^j \tau_i} z_1 + O(q^j), \quad \text{where} \;\; z_1 := \mathbf{Z}^{-1} t_1. \tag{3.24}$$

As will be seen later, we need to substitute $\psi(\mathbf{J}_1)^j$ instead of $\Psi_j(\mathbf{J}_1)$ in (3.24). For that reason we formulate some auxiliary assertions. According to the definition

$$\frac{\Psi_j(\lambda)}{\prod_{i=1}^j \tau_i} = \prod_{i=1}^j \frac{1}{\tau_i}\,(\tau_i \psi_i(\lambda)).$$

In analogy with the proof of Lemma 3.1 we write

$$\tau_i\psi_i(\mathbf{J}_1) = \prod_{m=1}^l (\mathbf{J}_1 - \mu_m^{(i)}\mathbf{I}) = \psi(\mathbf{J}_1) \prod_{m=1}^l [\mathbf{I} + (\mu_m - \mu_m^{(i)})(\mathbf{J}_1 - \mu_m\mathbf{I})^{-1}]. \tag{3.25}$$

Define $\mathbf{A}_m^{(i)} := (\mu_m - \mu_m^{(i)})(\mathbf{J}_1 - \mu_m\mathbf{I})^{-1}$ and $\alpha_m^{(i)} := \|\mathbf{A}_m^{(i)}\|$. It follows from Assumption 1 that $\lim_{j\to\infty} \sum_{i=j}^\infty \alpha_m^{(i)} = 0$ for $m = 1, \ldots, l$ and if we define the sequences $\{\zeta_j^{(m)}\}_{j=1}^\infty$ by the relation $\zeta_j^{(m)} := e^{\sum_{i=j}^\infty \alpha_m^{(i)}} - 1$ then

$$\lim_{j\to\infty} \zeta_j^{(m)} = 0 \;\; \text{for} \;\; m = 1, \ldots, l. \tag{3.26}$$

Now we seek to determine $\Psi_j(\mathbf{J}_1)/\prod_{i=1}^j \tau_i$. According to the relation (3.2) an integer j_1 exists such that $\alpha_m^{(i)} \le 1/2$ for all $m \in \{1, \ldots, l\}$ and for all $i \ge j_1$. **In the sequel let j_1 be fixed.** Let $j_1 \le s < j$. From (3.25) it follows that

$$\begin{aligned} \frac{\Psi_j(\mathbf{J}_1)}{\prod_{i=1}^j \tau_i} &= (\prod_{i=1}^j \tau_i)^{-1}\, \psi(\mathbf{J}_1)^j \prod_{m=1}^l \prod_{i=1}^j (\mathbf{I} + \mathbf{A}_m^{(i)}) \\ &= (\prod_{i=1}^j \tau_i)^{-1}\, \psi(\mathbf{J}_1)^j\, \mathbf{G}_s\, (\mathbf{I} + \mathbf{F}_{j,s}), \end{aligned} \tag{3.27}$$

where $\mathbf{G}_s := \prod_{m=1}^l \prod_{i=1}^s (\mathbf{I} + \mathbf{A}_m^{(i)})$ and $\mathbf{I} + \mathbf{F}_{j,s} := \prod_{m=1}^l \prod_{i=s+1}^j (\mathbf{I} + \mathbf{A}_m^{(i)})$.

In the following we show that the norms of the matrices $\mathbf{G}_s$ are uniformly bounded and the norm of $\mathbf{F}_{j,s}$ converges to zero. This auxiliary assertions will be used to prove theorems about derogatory and/or defective matrices.

Lemma 3.5: *Let $s \ge j_1$ be an arbitrary integer. Then*

$$\mathbf{G}_s\mathbf{J}_1 = \mathbf{J}_1\mathbf{G}_s \tag{3.28}$$

and the norms $\|\mathbf{G}_j\|$ *for* $j \in \{j_1, j_1+1, \ldots\}$ *are uniformly bounded, i.e. a constant* K *exists such that*

$$\|\mathbf{G}_j\| \leq K \quad \forall j \geq j_1. \tag{3.29}$$

Proof. The matrix $\mathbf{G}_s$ can be expressed in the form

$$\mathbf{G}_s = \prod_{m=1}^{l} p_{s,m}((\mathbf{J}_1 - \mu_m \mathbf{I})^{-1}),$$

where $p_{s,m}$ are polynomials of degree s such that $p_{s,m}(0) = 1$. This immediately yields (3.28). The integer j_1 is fixed and therefore a constant K_1 exists such that $\|\mathbf{G}_{j_1}\| \leq K_1$. Hence

$$\|\mathbf{G}_j\| \leq \|\mathbf{G}_{j_1}\| \prod_{m=1}^{l} \left\| \prod_{i=j_1+1}^{j} (\mathbf{I} + \mathbf{A}_m^{(i)}) \right\|.$$

But for the last norm we have

$$\left\| \prod_{i=j_1+1}^{j} (\mathbf{I} + \mathbf{A}_m^{(i)}) \right\| \leq \prod_{i=j_1+1}^{j} (1 + \alpha_m^{(i)}) = e^{\sum_{i=j_1+1}^{j} \ln(1+\alpha_m^{(i)})} \leq e^{\sum_{i=j_1+1}^{\infty} \alpha_m^{(i)}} =: K_2^{(m)}.$$

In the last inequality we have used the relation $\ln(1+x) \leq x$ for $x \in < 0, \frac{1}{2} >$. If we joint all previous estimates we obtain

$$\|\mathbf{G}_j\| \leq K_1 \prod_{m=1}^{l} K_2^{(m)} =: K. \tag{3.30}$$

□

Lemma 3.6: *Each matrix* $\mathbf{F}_{j,s}$ *commutes with* $\mathbf{J}_1$ *and*

$$\lim_{\substack{s\to\infty,\\ j>s}} \|\mathbf{F}_{j,s}\| = 0.$$

Proof. Let us rewrite $\mathbf{F}_{j,s}$ as a multi-polynomial function of the matrices $\mathbf{A}_m^{(i)}$ for $m = 1, \ldots, l$ and $i = s, s+1, \ldots$, i.e. $\mathbf{F}_{j,s} = q_{s+1,j}(\mathbf{A}_m^{(i)})$. For the norm we have the estimate

$$\begin{aligned} \|\mathbf{F}_{j,s}\| \quad &\leq 1 + q_{s+1,j}(\alpha_m^{(i)}) - 1 \leq \prod_{m=1}^{l} \prod_{i=s}^{j} (1 + \alpha_m^{(i)}) - 1 \\ &\leq \prod_{m=1}^{l} (e^{\sum_{i=s}^{\infty} \alpha_m^{(i)}}) - 1 = \prod_{m=1}^{l} (1 + \zeta_s^{(m)}) - 1 =: \eta_s \end{aligned}$$

and $\lim_{s\to\infty} \eta_s = 0$ according to (3.26). □

Lemma 3.7: *Let Assumption 1-4 be valid. Then the inequalities (3.21) are valid if we substitute $\psi(\mathbf{J}_1)^j$ instead of $\Psi_j(\mathbf{L}_1)$.*

Proof. The lemma immediately follows from (3.21) and previous assertions. □

According to previous lemmas we can now easy express the vector $v_1^{(j)}$ from (3.24) by using the limit polynomial ψ instead of the polynomial Ψ_j.

Theorem 3.8: *Let Assumption 1-4 be valid. Then sequences of vectors $\{\hat{v}^{(j)}\}_{j=1}^{\infty}$ and $\{t^{(j)}\}_{j=1}^{\infty}$ and a positive constant R exist such that*

$$v_1^{(j)} = \mathbf{Z}_1 \frac{\psi(\mathbf{J}_1)^j}{\prod_{i=1}^{j} \tau_i} t^{(j)} + \hat{v}^{(j)} \tag{3.31}$$

holds, where $\|t^{(j)}\| \leq R \ \forall j$ and $\lim_{j\to\infty} \|\hat{v}^{(j)}\| = 0$.

Proof. Let ξ be an arbitrary positive number. From (3.24), (3.27) and Lemma 3.5 we have

$$\begin{aligned} v_1^{(j)} &= \mathbf{Z}_1 \frac{\psi(\mathbf{J}_1)^j}{\prod_{i=1}^{j} \tau_i} \mathbf{G}_s(\mathbf{I} + \mathbf{F}_{j,s}) z_1 + O(q^j) \\ &= \mathbf{Z}_1 \frac{\psi(\mathbf{J}_1)^j}{\prod_{i=1}^{j} \tau_i} \mathbf{G}_s z_1 + \Big(\mathbf{Z}_1 \frac{\psi(\mathbf{J}_1)^j}{\prod_{i=1}^{j} \tau_i} \mathbf{F}_{j,s} \mathbf{G}_s z_1 + O(q^j)\Big). \end{aligned}$$

Using the previous lemmas an index $s(\xi)$ exists such that $\|\hat{v}^{(j)}(\xi)\| \leq \xi \ \forall j \geq s(\xi)$, where

$$\hat{v}^{(j)}(\xi) \ := \ \mathbf{Z}_1 \frac{\psi(\mathbf{J}_1)^j}{\prod_{i=1}^{j} \tau_i} \mathbf{F}_{j,s(\xi)} \mathbf{G}_{s(\xi)} z_1 \ + \ O(q^j). \tag{3.32}$$

For $t(\xi) := \mathbf{G}_{s(\xi)} z_1$ it is $\|t(\xi)\| \leq K\|z_1\| =: R$ according to (3.30). Hence for each $\xi > 0$ an integer $s(\xi)$, a vector $t(\xi)$ and a sequence of vectors $\{\hat{v}^{(j)}(\xi)\}_{j=s(\xi)}^{\infty}$ exist such that

$$v_1^{(j)} \ = \ \mathbf{Z}_1 \frac{\psi(\mathbf{J}_1)^j}{\prod_{i=1}^{j} \tau_i} t(\xi) \ + \ \hat{v}^{(j)}(\xi) \tag{3.33}$$

for all $j \geq s(\xi)$, where $\|t(\xi)\| \leq R$ and $\|\hat{v}^{(j)}(\xi)\| \leq \xi$ for all $j \geq s(\xi)$. Let $\{\xi_i\}_{i=1}^{\infty}$ be a decreasing sequence of positive numbers such that $\lim_{i\to\infty} \xi_i = 0$ and construct the increasing sequence of corresponding indexes $\{s(\xi_i)\}_{i=1}^{\infty}$. Define

$$t^{(j)} := t(\xi_i), \quad \hat{v}^{(j)} := \hat{v}^{(j)}(\xi_i) \quad \text{for} \quad j \in < s(\xi_i), s(\xi_{i+1}) - 1 > .$$

For $j < s(\xi_1)$ put $t^{(j)} := t(\xi_1)$ and define $\hat{v}^{(j)}$ such that the equality (3.31) is fulfilled. Then the vector $v_1^{(j)}$ has the form (3.31). □

To simplify the following explanation, consider for a moment a special form of the Jordan matrix $\mathbf{J}_1 := \operatorname{diag}(\mathbf{J}_1^{(1)}, \mathbf{J}_1^{(2)}, \mathbf{J}_1^{(3)})$, where

$$\mathbf{J}_1^{(1)} = \begin{pmatrix} \lambda_1 & 1 & 0 \\ 0 & \lambda_1 & 1 \\ 0 & 0 & \lambda_1 \end{pmatrix}, \quad \mathbf{J}_1^{(2)} = \begin{pmatrix} \lambda_1 & 1 \\ 0 & \lambda_1 \end{pmatrix}, \quad \mathbf{J}_1^{(3)} = \begin{pmatrix} \lambda_2 & 1 \\ 0 & \lambda_2 \end{pmatrix}.$$

The general case can then be formulated immediately.

It is easy to calculate that

$$\psi(\mathbf{J}_1)^j = \psi(\lambda_1)^j \operatorname{diag}(\mathbf{B}_j^{(1)}, \mathbf{B}_j^{(2)}, \mathbf{B}_j^{(3)}),$$

where

$$\mathbf{B}_j^{(1)} := \begin{pmatrix} 1, & \frac{(\psi(\lambda)^j)'_{\lambda=\lambda_1}}{\psi(\lambda_1)^j}, & \frac{(\psi(\lambda)^j)''_{\lambda=\lambda_1}}{2!\,\psi(\lambda_1)^j} \\ & 1, & \frac{(\psi(\lambda)^j)'_{\lambda=\lambda_1}}{\psi(\lambda_1)^j} \\ & & 1 \end{pmatrix}, \quad \mathbf{B}_j^{(3)} := \begin{pmatrix} (\frac{\psi(\lambda_2)}{\psi(\lambda_1)})^j, & \frac{(\psi(\lambda)^j)'_{\lambda=\lambda_2}}{\psi(\lambda_1)^j} \\ & (\frac{\psi(\lambda_2)}{\psi(\lambda_1)})^j \end{pmatrix}$$

and $\mathbf{B}_j^{(2)}$ is the leading 2×2 submatrix of $\mathbf{B}_j^{(1)}$. Let $\mathbf{Z}_1 = (z_1^{(1)}, \ldots, z_1^{(7)})$; for $\xi > 0$ let $t(\xi) = (t_1(\xi), \ldots, t_7(\xi))^T$ and $\hat{v}^{(j)}(\xi)$ be the vectors defined in (3.33). With this notation we have

$$\mathbf{Z}_1 \frac{\psi(\mathbf{J}_1)^j}{\prod_{i=1}^j \tau_i} t(\xi) = \frac{\psi(\lambda_1)^j}{\prod_{i=1}^j \tau_i} \sum_{r=1}^{7} \alpha_r(j, \xi) z_1^{(r)},$$

where

$$\begin{pmatrix} \alpha_1(j,\xi) \\ \vdots \\ \alpha_7(j,\xi) \end{pmatrix} = \operatorname{diag}(\mathbf{B}_j^{(1)}, \mathbf{B}_j^{(2)}, \mathbf{B}_j^{(3)}) \begin{pmatrix} t_1(\xi) \\ \vdots \\ t_7(\xi) \end{pmatrix}.$$

Assumption 5: Let $|\psi(\lambda_1)| > |\psi(\lambda_j)|$ for $j \in \{2, \ldots, k\}$. □

From Assumption 5 (that reduces to $|\psi(\lambda_1)| > |\psi(\lambda_2)|$ in our special case), it follows that

$$\lim_{j \to \infty} \mathbf{B}_j^{(3)} = 0. \tag{3.34}$$

Let $\varepsilon > 0$ be arbitrary and choose $\xi > 0$ and $j \geq s(\xi)$ such that

$$\xi \,+\, L \,\|\mathbf{B}_j^{(3)}\|\, K \,\|z_1\| \,<\, \varepsilon,$$

where $L := \lim_{j\to\infty} \sup \frac{\|\psi(\mathbf{J}_1)^j\|}{\prod_{i=1}^j \tau_i}$ and K is defined by (3.30). Such j exists due to (3.34). The vectors $z_1^{(1)}, z_1^{(2)}, z_1^{(3)}$ and $z_1^{(4)}, z_1^{(5)}$ form an invariant subspace respectively. Hence putting $\alpha_r := \alpha_r(j, \xi)$ (j and ξ are now fixed) and

$$c_1 \,:=\, \frac{\psi(\lambda_1)^j}{\prod_{i=1}^j \tau_i} \sum_{r=1}^{3} \alpha_r z_1^{(r)}, \quad c_2 \,:=\, \frac{\psi(\lambda_1)^j}{\prod_{i=1}^j \tau_i} \sum_{r=4}^{5} \alpha_r z_1^{(r)},$$

$$\hat{v}_j := \hat{v}^{(j)}(\xi) + \frac{\psi(\lambda_1)^j}{\prod_{i=1}^j \tau_i} \sum_{r=6}^{7} \alpha_r z_1^{(r)},$$

where $\hat{v}^{(j)}(\xi)$ is defined by (3.32), we have that $v_1^{(j)} = c_1 + c_2 + \hat{v}_j$ with $\|\hat{v}_j\| < \varepsilon$. From the last considerations we conclude that for every $\varepsilon > 0$ an integer $j > 0$ and vector $\hat{v}_j$ exist such that $\|\hat{v}_j\| < \varepsilon$ and the vectors

$$(v_1^{(j)} - \hat{v}_j),\ \mathbf{A}(v_1^{(j)} - \hat{v}_j),\ \mathbf{A}^2(v_1^{(j)} - \hat{v}_j),\ \mathbf{A}^3(v_1^{(j)} - \hat{v}_j)$$

are linearly dependent. From the above presented analysis the following general theorem can be formulated.

Theorem 3.10: *Let Assumption 1-5 be valid. Let the Jordan canonical form of the matrix* $\mathbf{A}$ *have* $s > 1$ *blocks corresponding to the eigenvalue* λ_1 *with the maximal dimension* $d < k$*. Then for every* $\varepsilon > 0$ *a positive integer* j *and a vector* $\hat{v}_j$ *exist such that* $\|\hat{v}_j\| < \varepsilon$ *and the vectors*

$$(v_1^{(j)} - \hat{v}_j),\ \mathbf{A}(v_1^{(j)} - \hat{v}_j),\ \ldots,\ \mathbf{A}^d(v_1^{(j)} - \hat{v}_j)$$

are linearly dependent.

Theorem 3.10 implies that the implicitly restarted Arnoldi process with shifts yields after j steps the Hessenberg matrix with the element $h_{i+1,i}^{(j)} \approx 0$ for some $i \in \{1, 2, \ldots, d\}$. Hence the iterative process described in Algorithm 2.1 cannot construct an approximation to the whole eigenspace corresponding to a multiple eigenvalue (to which corresponds more than one Jordan block), but it constructs an approximation to an eigenspace of the dimension less or equal to d. Therefore it is important to replace the condition in step *5.* of Algorithm 2.1

$$\text{5. if } \|w_k\| < TOL \text{ then STOP} \tag{3.35}$$

by a more general condition

$$\text{5. if } \|w_j\| < TOL \text{ for some } j \in \{1, 2, \ldots, k\} \text{ then STOP} \tag{3.36}$$

and one must be careful about the dimension of the constructed space. Let us remind that $\|w_j\|$ for $j = 1, \ldots, k-1$ need not be computed explicitly because $\|w_j\| = e_{j+1}^T \mathbf{H}_k e_j$.

We remark, moreover, that in numerical experiments the Arnoldi process starting with $v_1^{(j)}$ breaks down usually in the dth step. The above presented results do not imply a problem in computations with a derogatory and/or defective matrix $\mathbf{A}$, because preconditioning techniques presented in Section 1.4 remove small eigenvalues from the spectrum iteratively. Therefore a multiple eigenvalue is removed successively starting usually from the largest to the smallest Jordan block. The above statements are documented on Examples 2 in Chapter 4.

Chapter 4

Numerical experiments

In this chapter, we present numerical experiments to illustrate the behavior of the methods for computation of invariant subspaces described in Chapter 2. First we compare the widely used implicitly restarted Arnoldi process with shifts being the Ritz or the harmonic Ritz values, and the method using Tchebychev filters. Then we concentrate on comparison of the restarted GMRES method with its preconditioned versions described in Chapter 1. We exemplify that these preconditioners can accelerate the convergence of the GMRES method.

4.1 Description

In this chapter, we present numerical experiments comparing the method described in Section 2.2 (classical method) and the method based on Tchebychev polynomials described in Section 2.3 (Tchebychev method). Tchebychev filter is applied by IRA with shifts as it was described in Section 2.4. We compare the number of iterations needed to compute a good approximation to k smallest eigenvalues and the corresponding invariant subspace of a given matrix $\mathbf{A}$ for different choices of k and l. Arnoldi factorization is constructed by the modified Arnoldi process, see Algorithm 1.1. It is well known that the Householder process, see Algorithm 1.2, retains orthogonality of the computed matrix $\mathbf{V}_k$ better, but it is more time consuming. The effect of using the Householder process is documented on Example 1.

Further, in the last part of this chapter we exemplify that preconditioners based on invariant subspaces described in Section 1.4 can remove stagnation of the GMRES method and accelerate the convergence. Tables and graphs comparing the number of iterations and computation time are given. As it was mentioned in the previous, Ritz or harmonic Ritz values can be used in both the classical and the Tchebychev method. It can be observed that usage of harmonic Ritz values gives better results for larger k and l. This effect is also demonstrated on numerical examples in the last part of this chapter.

Numerical experiments were carried out on AMD Duron, 700 MHz/128 MB RAM computer. Example 6 was, by reason of its larger dimension, carried out on AMD Opteron 246, 2 GHz/2 GB RAM computer. We use the initial approximation $x_0 = (0, \dots, 0)^T$ and the right hand side $b = (1, 1, \dots, 1)^T$. Denote by

$$err = \log_{10}(||r_s||/||r_0||),$$

where r_s is the residual after s iterations of the GMRES(m) method, $s = j * m$, j is the number of restarts, and $r_0 = b - \mathbf{A}x_0$. We stopped the GMRES(m) process as soon as $err < -10$. Computation time is measured in seconds. If the invariant subspace is computed, one iteration of the classical method and the Tchebychev method for the same k, l and orthogonalization process takes the same time. Therefore we compare the number of iterations while computing invariant subspaces.

4.2 Computation of invariant subspace

In this section, we concentrate on computation of invariant subspaces. Presented experiments are theoretical and complete the convergence results from the previous chapter.

In the first example, we verify the assumptions of Theorems 3.3 and 3.4 and demonstrate the relation between the number of iterations in which condition (3.8) is satisfied and the convergence of the process. Moreover, the difference between using the modified Arnoldi process (ARN) and the Householder process (HHA) is analyzed.

To understand the complicated behavior of discussed preconditioners of the GMRES method, one needs first to understand the relevant simple cases. Convergence of GMRES for special matrices - (perturbed) Jordan blocks and Toeplitz matrices, is widely studied, e.g., in [45], [51], [85]. Thus Example 2 and 3 present results for special forms of the matrix $\mathbf{A}$. In Example 2 a block diagonal matrix with Jordan blocks on the diagonal is considered and the behavior of Algorithm 2.2 is studied. The matrix is derogatory and the statements of Theorem 3.10 are confirmed. Example 3 demonstrates the typical behavior of Algorithm 2.2 on the bidiagonal matrix. The lost of orthogonality between the columns of the matrix $\mathbf{V}_k$ in connection with convergence of subdiagonal elements of the matrix $\mathbf{H}_k$ to zero is analyzed.

Example 1: Consider the matrix $\mathbf{A} \in \mathbb{R}^{100 \times 100}$, where $\mathbf{A} = \mathbf{S}\,\mathbf{D}\,\mathbf{S}^{-1}$, $\mathbf{D} = \text{diag}(1, 2, .., 100)$ and $\mathbf{S}$ is bidiagonal matrix with 1 on the diagonal and 1.1 on the superdiagonal. This test matrix is taken from [21] and it is highly nonnormal with the Henrici number $\|\mathbf{A}\mathbf{A}^T - \mathbf{A}^T\mathbf{A}\|_F / \|\mathbf{A}\|_F \doteq 102\,080$.

method	iteration	$\|w_k\|$	Ass. 2	Ass. 1
classical filter	15	10^{-2}	1.-4. iter.	NO
Tchebychev + ARN	35	10^{-3}	1.-4. iter.	NO
Tchebychev + HHA	38	10^{-6}	1.-42. iter.	YES

Table 4.1: Convergence of the classical method and the method using Tchebychev polynomials for $k = 5, l = 5$ for Example 1.

In this example, our attention is concentrated on verification of the assumptions of the fundamental Theorem 3.3 and Theorem 3.4. Then the behavior of Algorithm 2.2 is observed, in case that the Arnoldi factorization is computed by ARN and HHA, respectively. We set $k := 5$, $l := 5$, i.e. an invariant subspace of dimension 5 is computed using the Arnoldi factorization (2.1) of order $k + l = 10$. The second column of Table 4.1 reflects the number of iterations in which the smallest $\|w_k\|$ was reached and the third column contains the smallest $\|w_k\|$. The fulfillment of the Assumptions 1 and 2 is displayed in the last two columns. The results in the first two lines of Table 4.1 are not quite as satisfying as we may expect. This is probably due to the fact that $\mathbf{A}$ is highly nonnormal. The HHA process, on the other hand, does not suffer from this. Figure 4.1 shows the progress of the shifts in the last five iterations for the process using HHA. Bold numbers represent the nearest eigenvalue of $\mathbf{A}$. The graph demonstrates that the part (3.1) of Assumption 1 is fulfilled, even though the shift μ_3 is very close to eigenvalue 53.

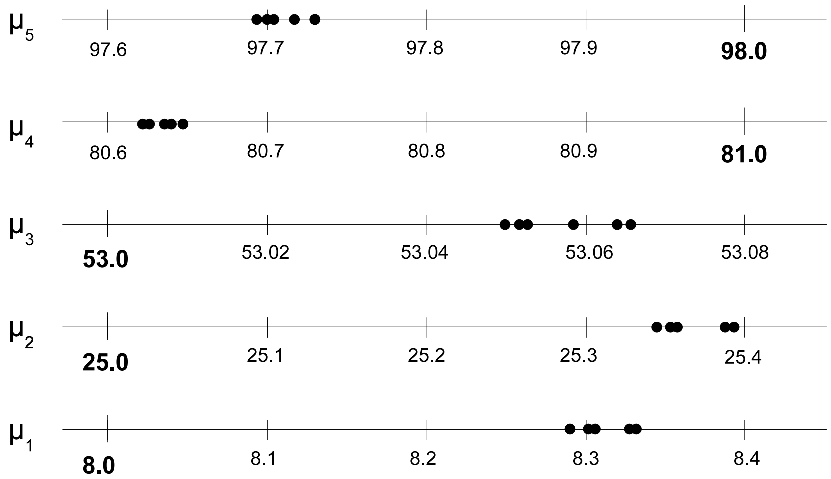

Figure 4.1: Shifts obtained by using the Tchebychev filter in the last five iterations for Example 1.

Example 2: Let $\mathbf{A} = \mathrm{diag}(\mathbf{J}_1, \mathbf{J}_2, \mathbf{J}_3, \mathbf{J}_4) \in R^{100\times 100}$ be a block diagonal matrix with Jordan blocks on the diagonal, where $\mathbf{J}_1, \mathbf{J}_2$ and $\mathbf{J}_3$ are blocks of dimensions $2, 3$ and 5 corresponding to $\lambda_1 = \ldots = \lambda_{10} = 0.1$ and $\mathbf{J}_4$ is a Jordan block corresponding to $\lambda_{11} = \ldots = \lambda_{100} = 100$. We demonstrate the behavior of the presented algorithms for computation of invariant subspaces on the classical method using ARN and the harmonic Ritz values. (The behavior is similar, if the Tchebychev filter, the Ritz values, HHA is used.) In this example, a large gap between λ_1 and λ_{11} was taken in order to obtain quickly $h_{d+1,d} \approx 0$.

Let $k := 8$, $l := 2$, i.e. an approximation to the eigenspace of dimension 8 is searched, corresponding to the eigenvalue $\lambda_1 = 0.1$ using the Arnoldi factorization of order $k + l = 10$. As proved in Theorem 3.10, the iterative process indeed constructs approximation to the eigenspace of dimension equal to the dimension of the largest Jordan block corresponding to λ_1, i.e. $d = 5$, and the element $h_{d+1,d} = h_{6,5}$ converges to zero. The only fast decreasing curve in Figure 4.2 corresponds to the element $h_{6,5}$. The number $h_{9,8}$ denotes $\|w_k\| = \|w_8\|$.

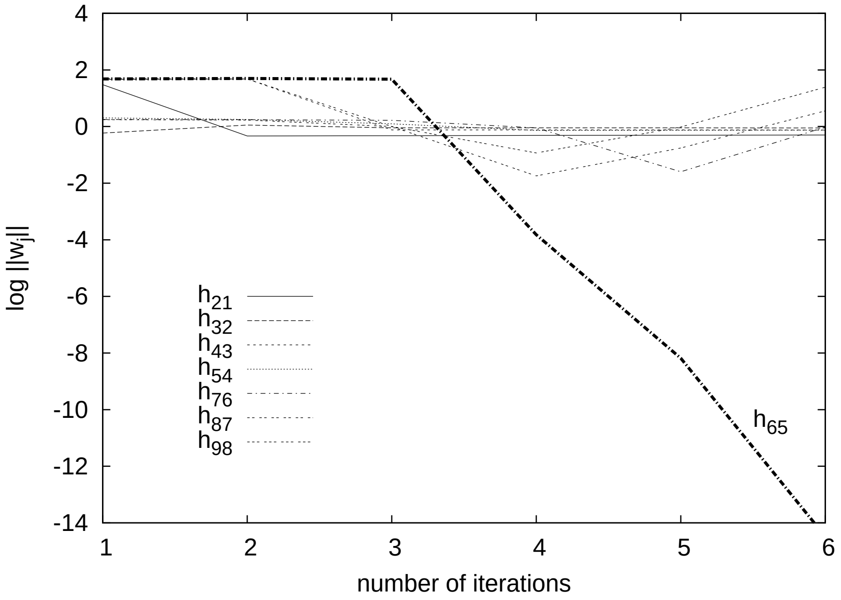

Figure 4.2: Progress of subdiagonal elements of the matrix $\mathbf{H}_k$ for Example 2.

Example 3: The matrix $\mathbf{A}$ is bidiagonal with entries $1, 2, \ldots, 20$ on the main diagonal and 0.1's on the superdiagonal. The following tests illustrate what happens if the required tolerance (TOL) of the approximation to the invariant subspace is too small. Let $k := 3$, $l := 2$. The results were obtained by the classical method using ARN and the Ritz values. (The behavior is similar, if the Tchebychev filter, the harmonic Ritz values, HHA is used.)

k	22	25	33	39	40	54	63
h_{21}	4.66E−3	6.06E−4	6.15E−5	7.00E−6	1.12E−7	9.90E−6	3.30E−6
h_{32}	3.03E−3	5.36E−6	3.15E−4	1.50E−2	2.24E−6	2.22E−6	2.22E−6
h_{43}	1.39E−3	**1.76E0**	3.84E−4	7.58E−7	**9.32E−1**	9.073E−6	8.23E−9

Table 4.2: Subdiagonal elements of the Hessenberg matrices for Example 3.

Figure 4.3 curve "full iteration" shows that the value $\|w_k\|$ decreases until the iteration 25 and then it jumps upwards. Then the process converges again until the same situation occurs in iteration 40 and again in iteration 64. From Figure 4.5 it is obvious that these jumps correspond to the loss of orthogonality in the last columns of the matrix $\mathbf{V}_k$. The behavior of functions $\log_{10} \|\mathbf{I} - \mathbf{V}_j^T \mathbf{V}_j\|$ for $j = k, k-1, k-2$ is drawn. For completeness the subdiagonal elements of the Hessenberg matrices $\mathbf{H}_3$ are presented in Table 4.2 for some choices of k. We observed that after each jump one more very small eigenvalue appears in $\sigma(\mathbf{H}_3)$, the matrix is more and more rank deficient, and thus we continue in estimating an invariant subspace of smaller dimension. Therefore replacing (3.35) by (3.36) in Algorithm 2.1 is important even if the matrix $\mathbf{A}$ has distinct eigenvalues.

Loss of orthogonality can be caused by implicit updating of the Arnoldi factorization by using shifts, see [49]. To emphasize this hypothesis we substitute the implicitly restarted Arnoldi process after some number of iterations by Arnoldi factorization starting from the beginning, i.e. from the updated vector v_1. We say that the restart has been performed. The process without restart is called "full iterations". In Figure 4.4 the function $\log_{10} \|\mathbf{I} - \mathbf{V}_k^T \mathbf{V}_k\|$ is drawn for the restarts 5, 10 and "full iterations". The iterative process can achieve smaller TOL if we use restart, see Figure 4.3 for the restarts 5 and 10. The error analysis is beyond the scope of this thesis. Let us remark that if the Tchebychev filter is used, all jumps appear a few iterations later.

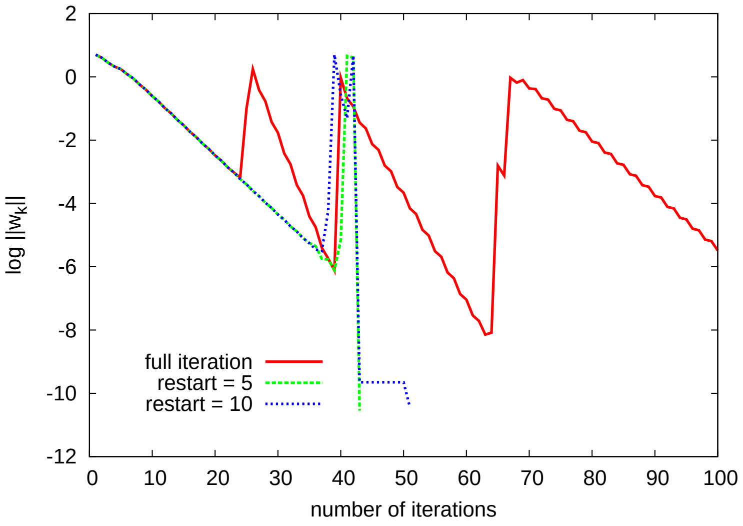

Figure 4.3: Convergence of $\|w_k\|$ for Example 3.

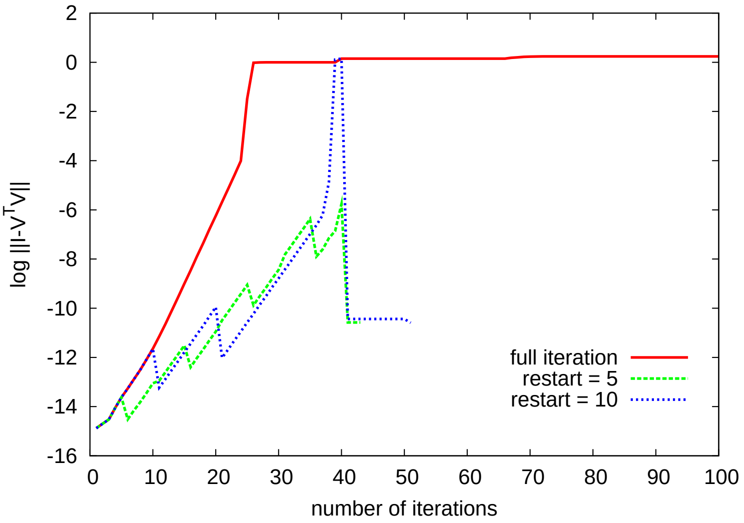

Figure 4.4: Convergence of $\|\mathbf{I} - \mathbf{V}_k^T \mathbf{V}_k\|$ for Example 3.

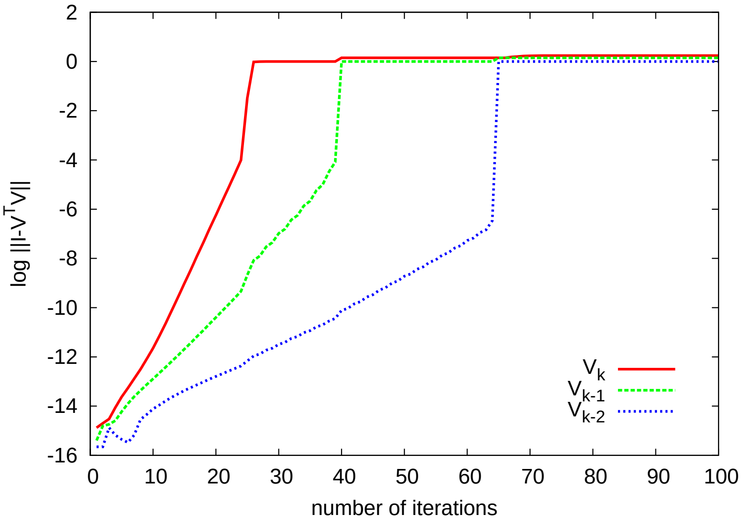

Figure 4.5: Behavior of $\|\mathbf{I} - \mathbf{V}^T\mathbf{V}\|$ for $\mathbf{V} = \mathbf{V}_k, \mathbf{V}_{k-1}, \mathbf{V}_{k-2}$ for Example 3.

4.3 Acceleration of GMRES method

Now we will use the results from the previous chapters to accelerate the convergence of the GMRES method by using preconditioners described in Section 1.4. We compare GMRES(m) with GMRESQR (method preconditioned from the left according to Theorem 1.3) and DEFLGMRES (method preconditioned from the right according to Theorem 1.4). Both algorithms change the preconditioning matrix adaptively after each restart, but the update is performed differently, see Section 1.4. This is described in details in Example 2. Note, that numerical experiments show that it is usually sufficient to take $TOL \in < 10^{-4}, 10^{-6} >$ in Algorithm 2.2 or Algorithm 2.3. Thus this TOL is used in the presented experiments.

The first two matrices are theoretical and the same as in Example 1 and 2 in the previous section. Example 1 shows that the preconditioners based on invariant subspaces can accelerate the convergence, even if the matrix is highly nonnormal. Example 2 describes the behavior of preconditioners for derogatory matrix $\mathbf{A}$. The following linear systems arise from practical problems. Example 3 considers strongly ill-conditioned, but close to normal matrix and shows that invariant subspaces of this matrix are computed very quickly and accurately. Thus also the preconditioners work very good. Example 4 gives the results for matrix that is not normal, but also not too nonnormal. Finally, Examples 5 and 6 consider linear system arising from discretization of partial differential equation and show that the behavior of all discussed GMRES methods is similar if we proceed to larger dimensions, i.e. the grid is refined.

Example 1: Let $\mathbf{A}$ be the matrix from Example 1 in Section 4.2. Though there is no extremely small eigenvalue, the high nonnormality of the matrix $\mathbf{A}$ may cause stagnation of the GMRES method.

Let the restart length be $m := 10$ for all considered solvers. Figure 4.6 shows the convergence behavior of GMRES(10), DEFLGMRES (where in each of the first twenty restarts one vector is added for constructing the adaptive preconditioner) and GMRESQR (where an invariant subspace of dimension $k := 5$ is constructed). We can observe that GMRES(10) stagnates; both preconditioned versions have an "initial phase" of stagnation (that corresponds to a few restarts in which the preconditioning matrix is refined and small eigenvalues are gradually captured) followed by a phase of fast convergence. This is a typical behavior and can be observed for almost all linear systems. The length of the initial phase depends on the quality of approximation of invariant subspace, i.e. on the matrix $\mathbf{V}_k$. Table 4.3 compares the residual norms and computation times in case that one of the studied methods attained $err < -10$.

Example 2: Consider the matrix $\mathbf{A}$ from Example 2 in Section 4.2. We have seen that setting $k := 8, l := 2$, the classical and the Tchebychev method

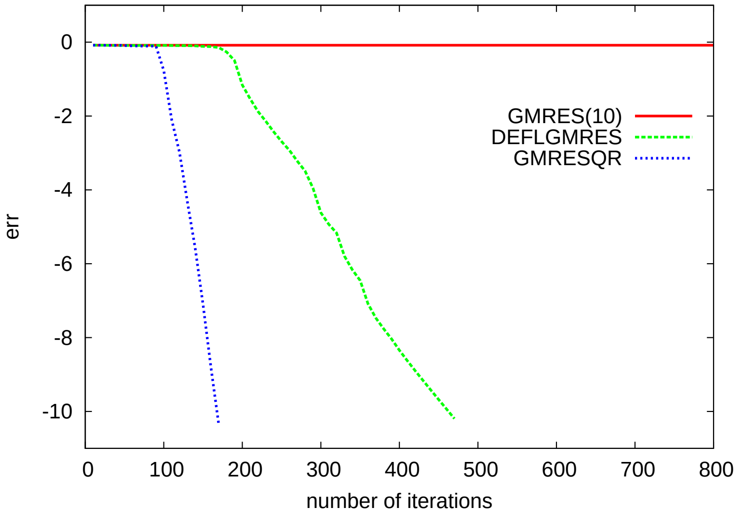

Figure 4.6: Convergence of GMRES, DEFLGMRES, GMRESQR for Example 1.

number of iterations	GMRES(10)		DEFLGMRES		GMRESQR	
	$\|r_s\|$	time	$\|r_s\|$	time	$\|r_s\|$	time
170	5,65E$-$1		7,14E$-$1		2,56E$-$11	1,76
470	5,65E$-$1		6,42E$-$11	10,02		

Table 4.3: Convergence times of GMRES, DEFLGMRES, GMRESQR for Example 1.

computes invariant subspace of dimension 5.

Let $m := k + l = 10$. Figure 4.7 and Table 4.4 display the convergence of GMRES(10) and its preconditioned versions DEFLGMRES and GMRESQR. Classical GMRES(10) converges very slowly, but both preconditioners accelerate the convergence. Let us have a look at the GMRESQR method. After the first restart the linear system is divided by the largest Ritz value $\theta_{10} = 100.14$ and then the eigenspace of the dimension 5 corresponding to λ_1 is computed. (Five iterations of the implicitly restarted Arnoldi process are performed according to Figure 4.2, see Example 2 in Section 4.2.) Having constructed the preconditioner $\mathbf{M}_1$ based on this eigenspace according to Theorem 1.3, the matrix $\mathbf{A}_1 := \mathbf{M}_1^{-1}\mathbf{A}$ is obtained that has λ_1 with multiplicity only $3 + 2 = 5$. We observed that the same invariant subspace is computed for arbitrary $k \geq 5$. In the second restart with the matrix $\mathbf{A}_1$ we can choose arbitrary $k \geq 3$ to obtain the eigenspace of dimension 3. The multiplicity of λ_1 is only 2 in the spectrum of $\mathbf{A}_2 := \mathbf{M}_2^{-1}\mathbf{M}_1^{-1}\mathbf{A}$. The following restart removes the eigenspace corresponding to λ_1 totally and the solution of the linear system is computed very quickly. Eigenvalues of precon-

number of iterations	GMRES(10)		DEFLGMRES		GMRESQR	
	$\|r_s\|$	time	$\|r_s\|$	time	$\|r_s\|$	time
30	1,12E−1		5,71E−24	0,21	1,05E−5	
40	9,56E−1				4,71E−11	0,10
350	1,80E−11	0,73				

Table 4.4: Convergence times of GMRES, DEFLGMRES, GMRESQR for Example 2.

ditioned matrices are reported in Table 4.5. In DEFLGMRES, the matrix $\mathbf{V}$ is gradually enriched by the vectors approximating required invariant subspaces and thus the convergence curve is more smooth.

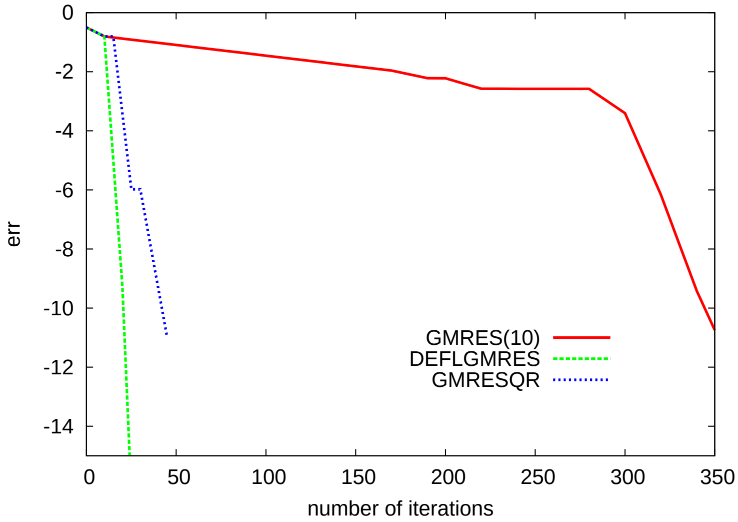

Figure 4.7: Convergence of GMRES, DEFLGMRES, GMRESQR for Example 2.

Example 3: Let $\mathbf{A}$ = Watt1, where Watt1 is the matrix from the Harwell-Boeing collection [40], which originates from problems in petroleum engineerings. The matrix is nonsymmetric, of order $n = 1\,856$ and has 11 360 nonzero entries. Its condition number is very large, it is estimated at 5,38E+9, but the matrix is close to normal with the Henrici number 9,55E−8. Thus we expect that the discussed preconditioners accelerate the convergence of GMRES strongly. This matrix has large eigenvalues with both positive and negative real parts. Therefore, alternation of intervals described in Section 2.3 must be used. The calculation of Tchebychev filter proceeds according to Algorithm 2.3.

We compute an invariant subspace of dimension $k = 2$ and $k = 15$ from the Arnoldi factorization of order 5 and 20, respectively. In Table 4.6, the numbers of

eigenvalues of	$\mathbf{A}$	$\mathbf{A}/\theta_{10}$	$\mathbf{A}_1$	$\mathbf{A}_2$	$\mathbf{A}_3$
λ_1	1E−1	9,97E−4	9,82E−4	1,35E−3	9,14E−1
λ_2	1E−1	9,97E−4	9,82E−4	1,35E−3	9,14E−1
λ_3	1E−1	9,97E−4	9,82E−4	9,50E−1	9,26E−1
λ_4	1E−1	9,97E−4	9,83E−4	9,50E−1	9,49E−1
λ_5	1E−1	9,97E−4	9,83E−4	9,50E−1	9,53E−1
λ_6	1E−1	9,97E−4	9,44E−1	9,74E−1	9,78E−1
λ_7	1E−1	9,97E−4	9,44E−1	9,74E−1	9,84E−1
λ_8	1E−1	9,97E−4	9,44E−1	9,74E−1	9,86E−1
λ_9	1E−1	9,97E−4	9,44E−1	9,75E−1	9,93E−1
λ_{10}	1E−1	9,97E−4	9,44E−1	9,75E−1	9,97E−1
λ_{11}	1E+2	9,97E−1	1,12E0	9,93E−1	1,07E0
λ_{100}	1E+2	9,97E−1	1,30E0	1,01E0	1,11E0

Table 4.5: Eigenvalues of preconditioned matrices for Example 2.

	k=2, l=3				**k=15, l=5**			
$\|w_k\|$	classical		Tchebychev		classical		Tchebychev	
10^{-7}	1	1	1	1	1	1	1	1
10^{-8}	43	4	3	1	7	7	2	2

Table 4.6: Convergence of the classical method and the method using Tchebychev polynomials for Example 3.

iterations are reported in which various precisions of approximations measured by $\|w_k\|$ are reached. The first and the second column show the results obtained by using the Ritz and the harmonic Ritz values, respectively. All methods compute invariant subspaces very quickly and with high accuracy.

We have used restart $m := 30$ for all considered GMRES methods. The preconditioner based on the invariant subspace of dimension only two suffices for fast convergence in the GMRESQR method. In the DEFLGMRES, one vector was added to construct the preconditioner in each of the 3 first restarts. Rates of convergence of all considered methods are presented in Table 4.7 and Figure 4.8. Both preconditioning techniques accelerate the convergence and, moreover, Figure 4.8 illustrates that there is no initial phase of stagnation or slow convergence of GMRESQR and DEFLGMRES, that can be observed in other examples. This is a consequence of the fact that the matrix $\mathbf{V}_2$ is computed in high accuracy already after the first restart.

Example 4: Consider the matrix Sherman4 from the Harwell-Boeing collection [40], which originates from oil reservoir modeling. The matrix is nonsym-

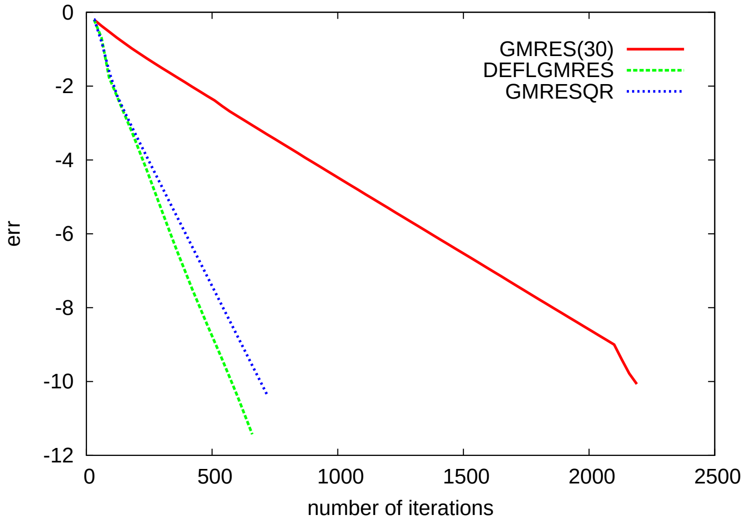

Figure 4.8: Convergence of GMRES, DEFLGMRES, GMRESQR for Example 3.

number of iterations	GMRES(30)		DEFLGMRES		GMRESQR	
	$\|r_s\|$	time	$\|r_s\|$	time	$\|r_s\|$	time
660	8,70E−4		1,61E−10	3,51	2,55E−9	
722	4,80E−4				1,72E−10	7,06
2190	8,55E−11	9,83				

Table 4.7: Convergence times of GMRES, DEFLGMRES, GMRESQR for Example 3.

metric, of order $n = 1\,104$ and has 3 786 nonzero entries. Its condition number is estimated at 7,20E+3 and the Henrici number is 1,3813. Though the small eigenvalues of this matrix are not sharply separated from the rest of the spectrum, the adaptive preconditioning accelerates the convergence of GMRES.

Table 4.8 is analogous to Table 4.6 and presents results for several choices of k and l. Again, the first and the second column display the results obtained by using the Ritz and the harmonic Ritz values, respectively, and a free space denotes that the method did not reach required precision. The case $k = 5$, $l = 5$ illustrates that in some cases usage of harmonic Ritz values can help to attain smaller $\|w_k\|$.

Let $m := 15$. Though GMRES(15) converges, both preconditioners accelerate the convergence, see Figure 4.9 and Table 4.9. The preconditioner in the GMRESQR method was constructed using the invariant subspace of order 5. In DEFLGMRES, one vector was added to construct the preconditioner in each of the 5 first restarts.

	k=5, l=5				**k=5, l=10**				**k=10, l=10**			
$\|w_k\|$	classical		Tchebychev		classical		Tchebychev		classical		Tchebychev	
10^{-2}	37	37	17	17	16	16	14	8	30	26	29	24
10^{-3}		60		53	20	20	20	20	32	32	30	30

Table 4.8: Convergence of the classical method and the method using Tchebychev polynomials for Example 4.

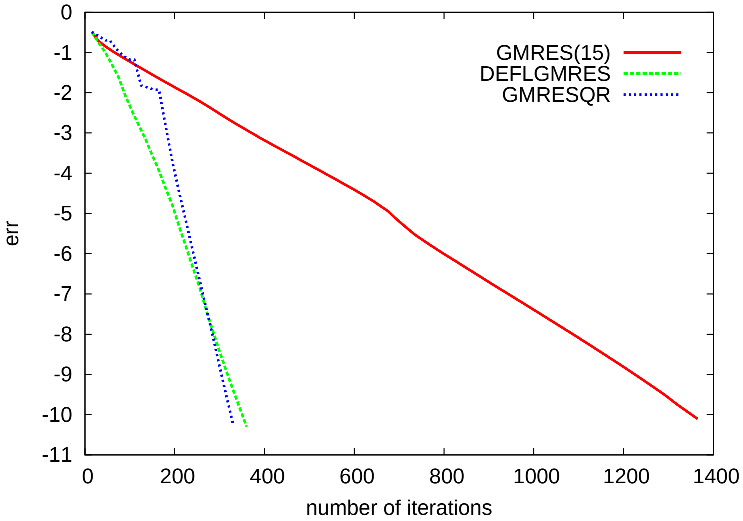

Figure 4.9: Convergence of GMRES, DEFLGMRES, GMRESQR for Example 4.

Example 5: In this example we consider a linear system that arises from discretization of the partial differential equation

$$-e^{-xy}u_{xx} - e^{-xy}u_{yy} + (10 + ye^{-xy})u_x + (10 + xe^{-xy})u_y - 60u = 1 \qquad (4.1)$$

in the domain $\Omega = (0,1) \times (0,1)$ with Dirichlet boundary condition (see [80]). Discretization was performed by standard finite differences on a 100×100 uniform grid and yields the stiffness matrix $\mathbf{A} \in \mathbb{R}^{10\,000 \times 10\,000}$ with 49 600 nonzero elements. This matrix is close to normal with the Henrici number 0,0097.

We compute an invariant subspace of dimension $k = 5$ and $k = 15$. Table 4.10 is analogous to Tables 4.6 and 4.8. The method using Tchebychev filter is faster than the classical one in all cases. Table 4.10 also shows that harmonic Ritz values give slightly better results for larger k and l.

Figure 4.10 and Table 4.11 illustrate convergence results for restart parameter $m := 35$. In the GMRESQR method an invariant subspace of dimension 5

number of iterations	GMRES(15)		DEFLGMRES		GMRESQR	
	$\|r_s\|$	time	$\|r_s\|$	time	$\|r_s\|$	time
330	$1,85E{-}3$		$1,92E{-}9$		$1,88E{-}10$	$2,34$
360	$1,48E{-}3$		$4,45E{-}11$	$2,10$		
1365	$7,87E{-}11$	$5,23$				

Table 4.9: Convergence times of GMRES, DEFLGMRES, GMRESQR for Example 4.

$\|w_k\|$	**k=5, l=5**				**k=5, l=15**				**k=10, l=20**			
	classical		Tchebychev		classical		Tchebychev		classical		Tchebychev	
10^{-2}	21	21	9	9	9	9	4	4	11	9	7	6
10^{-3}	87	87	55	55	25	18	15	15	19	15	16	16

Table 4.10: Convergence of the classical method and the method using Tchebychev polynomials for Example 5.

was computed and in DEFLGMRES one vector was added in each of the 10 first restarts. In this example restarted GMRES(35) fully stagnates.

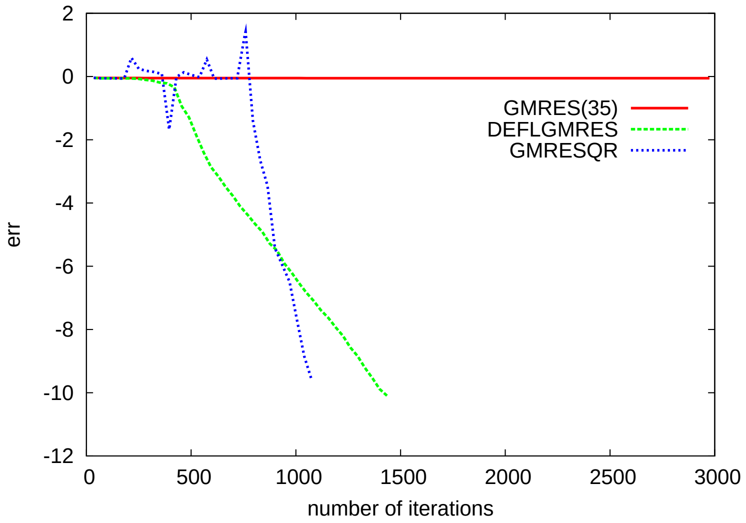

Figure 4.10: Convergence of GMRES, DEFLGMRES and GMRESQR for Example 5.

Example 6: Similar behavior of the algorithms GMRES, GMRESQR and DEFLGMRES as in the previous example can be observed, when we proceed

number of iterations	GMRES(35)		DEFLGMRES		GMRESQR	
	$\|r_s\|$	time	$\|r_s\|$	time	$\|r_s\|$	time
1075	8,87E−1		8,02E−8		3,20E−13	226,08
1435	8,86E−1		8,09E−11	128,64		

Table 4.11: Convergence times of GMRES, DEFLGMRES, GMRESQR for Example 5.

to larger dimensions. Discretizing (4.1) on a 320 × 320 grid yields the stiffness matrix $\mathbf{A} \in \mathbb{R}^{102\,400\times 102\,400}$ with 510 720 nonzero elements.

Figure 4.11 illustrates convergence results for restart $m := 50$. The restarted GMRES(50) fully stagnates. In the GMRESQR method an invariant subspace of dimension 10 was computed and in DEFLGMRES one vector was added in each of the 20 first restarts.

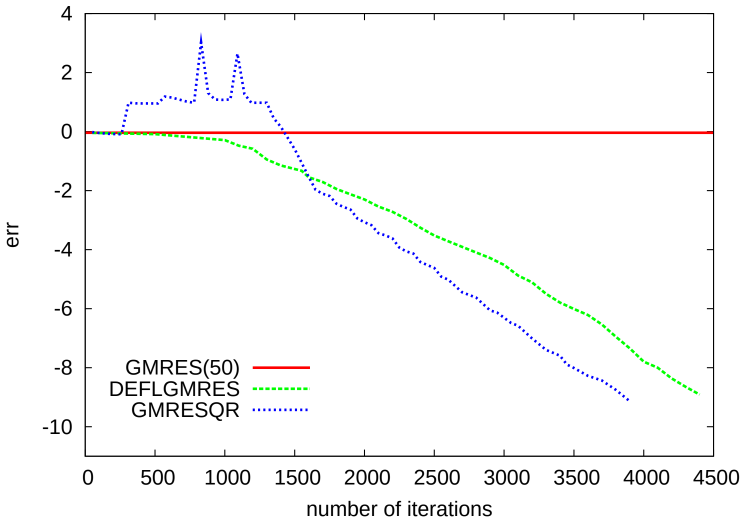

Figure 4.11: Convergence of GMRES, DEFLGMRES and GMRESQR for Example 6.

Part II

Core problem in errors-in-variables modeling

Chapter 5

Basic concepts

In this chapter we are interested in unitarily invariant linear approximation problems, represented by systems of linear algebraic equations with a generally rectangular and possibly rank deficient matrix. Such systems arise in many scientific and technical areas and various techniques are used to solve them. When the matrix representing the model in the approximation problem is large, which is often the case, we need to consider iterative methods with some appropriate stopping criteria. Main types of linear approximation problems and methods for solving them are summarized in this chapter.

5.1 Linear approximation problems

Consider estimating x from the system of linear algebraic equations

$$\mathbf{A}x \approx b, \quad \mathbf{A} \in \mathbb{R}^{n\times m}, b \in \mathbb{R}^{n} \tag{5.1}$$

with a nonzero matrix $\mathbf{A}$ and a nonzero vector b. The system can be compatible, i.e. $b \in \mathcal{R}(\mathbf{A})$, or incompatible, i.e. $b \notin \mathcal{R}(\mathbf{A})$. The uninteresting case is excluded by the assumption $\mathbf{A}^T b \neq 0$ (otherwise it is meaningless to approximate b by the columns of $\mathbf{A}$ and the system (5.1) has a trivial solution $x \equiv 0$). We assume, for simplicity of notation, that $\mathbf{A}$ and b are real. The extension to complex data is straightforward.

Sources of approximation problems

Linear approximation problems of the form (5.1) arise in a broad class of scientific and technical disciplines and applications. The first group of such problems comes from *errors-in-variables modeling*. Here $\mathbf{A}$ represents the data matrix determining the model, b is the observation vector and x is the unknown vector of true solution. The vector b represents some observed or measured variables and thus it can contain errors. Furthermore, in many applications the sampling or modeling errors may also imply inaccuracies in the matrix $\mathbf{A}$. Least squares [6] (also called linear regression in statistics) and total least squares [44] (orthogonal regression in statistics) techniques can be used to solve such system in case they are well-posed, e.g., in statistical applications. The corresponding approaches are summarized in Section 5.2.

Additional difficulty appears when the system (5.1) is *ill-posed.* Ill-posed systems appear in many applications - medical image deblurring (tomography), bioelectrical inversion problems, geophysics (seismology, radar or sonar imaging), astronomical observations. Here the matrix $\mathbf{A}$ is ill conditioned and a small perturbation on b typically causes large changes in the estimated solution x. Moreover, the matrix $\mathbf{A}$ is often numerically rank deficient and/or it has small singular values, but without a well defined numerical rank. We illustrate this on testing matrices from Matlab Regularization Tools [37].

Example 1: Figure 5.1 shows the singular values of 100×100 matrix Foxgood (left figure) and 100×23 matrix Parallax (right figure). Both matrices correspond to ill-posed problems, but the singular values of Foxgood decay fast to machine precision, while the singular values of Parallax decay gradually without a noticeable gap.

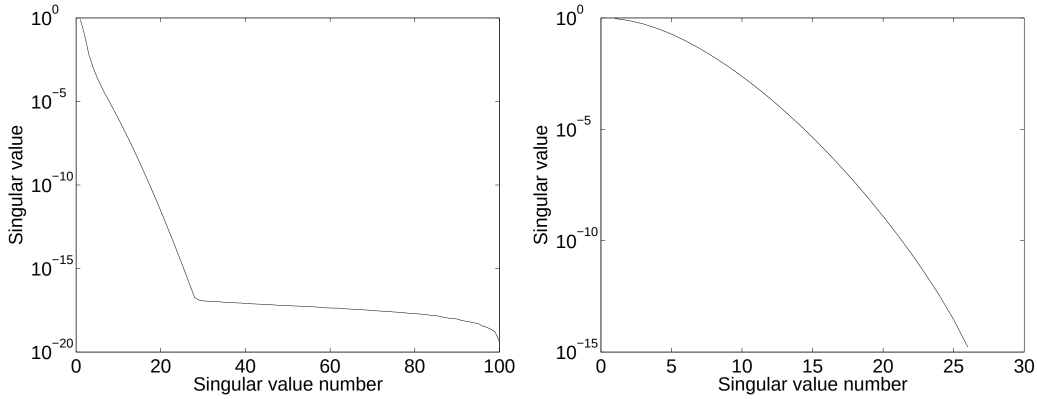

Figure 5.1: Singular values of the matrices Foxgood (left) and Parallax (right).

Usually linear models have the form

$$\mathbf{A}x \approx b, \quad \text{where} \;\; b = b^{exact} + b^{error}$$

and possibly also $\mathbf{A} = \mathbf{A}^{exact} + \mathbf{A}^{error}$. In such cases the least squares, total least squares or similar techniques might give a solution that is absolutely meaningless, because it is dominated by errors present in the data b, $\mathbf{A}$ and possibly also by computational errors. The *regularization techniques* must be used in order to obtain a meaningful solution. For more discussion see Section 5.3.

Model reduction represents another important area of applications. The idea is to approximate high order system (5.1) by a lower order one while approximating well the behavior of the whole system. *Truncation and projection techniques* (that may be viewed also as a type of regularization) are used here to reduce the dimensions of the linear system, see Section 5.3.

5.2 Least squares and related techniques

Ordinary least squares

The ordinary least squares (LS) method is used to solve the system (5.1) when errors are confined to the right hand side b but not to the matrix $\mathbf{A}$. The LS method seeks a vector $g \in \mathbb{R}^n$ satisfying

$$\begin{aligned} &\text{a) } \tilde{b} \in \mathcal{R}(\mathbf{A}), \text{ where } \tilde{b} = b + g \\ &\text{b) } g \;=\; \arg_{\tilde{g} \in \mathbb{R}^n} \min \|\tilde{g}\| \text{ subjected to a),} \end{aligned} \tag{5.2}$$

i.e. a minimal perturbation of the right-hand side b is searched such that $\mathbf{A}x \approx \tilde{b}$ is compatible. From the definition it follows that $\tilde{b}$ is the orthogonal projection of b on the space generated by the columns of $\mathbf{A}$. The LS solution always exists and is equal to

$$x^{LS} = \mathbf{A}^{+}b. \tag{5.3}$$

Numerical methods for computing LS solution are direct (based on singular value decomposition or QR-factorization of $\mathbf{A}$), or iterative, see [6].

Total least squares

The total least squares (TLS) method is used to solve the system (5.1) when errors are confined both to b and $\mathbf{A}$. The TLS method seeks a vector $g \in \mathbb{R}^n$ and a matrix $\mathbf{E} \in \mathbb{R}^{n \times m}$ such that

$$\begin{aligned} &\text{a) } \tilde{b} \in \mathcal{R}(\tilde{\mathbf{A}}), \text{ where } \tilde{b} = b + g,\ \tilde{\mathbf{A}} = \mathbf{A} + \mathbf{E} \\ &\text{b) } [g, \mathbf{E}] \;=\; \arg_{[\tilde{g}, \tilde{\mathbf{E}}]} \min \|[\tilde{g}, \tilde{\mathbf{E}}]\|_F \text{ subjected to a),} \end{aligned} \tag{5.4}$$

i.e. a minimal perturbation of the right-hand side b and the matrix $\mathbf{A}$ is searched such that $\tilde{\mathbf{A}}x \approx \tilde{b}$ is compatible.

The existence and uniqueness of the TLS solution depends on the right singular vector subspace of the matrix $[b, \mathbf{A}]$, corresponding to the smallest nonzero singular value σ. If σ is simple and the first component of the corresponding right singular vector s is nonzero, it can be proved that the TLS solution *exists and is unique* and has closed-form

$$x^{TLS} = (\mathbf{A}^T\mathbf{A} - \sigma^2\mathbf{I})^{-1}\mathbf{A}^T b,$$

see [44] p. 53. If $e_1^T s = 0$ (or more generally $e_1^T \mathbf{S}^{'} = 0$, where the columns of $\mathbf{S}^{'}$ generate the right singular vector subspace of $[b, \mathbf{A}]$ corresponding to multiple singular value σ) the TLS solution *does not exist.* This unpleasant situation can be illustrated on the following simple example.

Example 2: Consider the linear approximation problem

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \approx \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

The TLS method seeks the smallest perturbation $[g, \mathbf{E}]$ such that the resulting system is compatible. If we put

$$[g, \mathbf{E}] = \left[\begin{array}{c|cc} 0 & 0 & 0 \\ 0 & 0 & \theta^{-1} \end{array} \right],$$

then the Frobenius norm of this correction is equal to θ^{-1} and the corresponding solution is $x = (1, \theta)^T$. Thus $\theta \to \infty$ implies $\|[g, \mathbf{E}]\|_F \to 0$ and $\|x\| \to \infty$, i.e. only the non-optimal TLS solution can be obtained. Moreover, the above solution vector needs have nothing to do with the original approximation problem because θ is an arbitrary number.

Example 2 indicates that in case that $e_1^T \mathbf{S}^{'} = 0$ the TLS problem (5.4) is not well formulated. S. Van Huffel and J. Vandewalle analyzed this difficulty and defined a *nongeneric* TLS solution as a solution of the TLS problem (5.4) with the restriction

$$[g, \mathbf{E}] \perp \mathbf{S}^{'},$$

see [44] pp. 66-84. However, the nonexistence of the generic TLS solution in this case was at first proved in [64] and follows from the *core theory.* The core reduction avoids the nonuniqueness and nongenericity of the solution, by transforming the original data $b, \mathbf{A}$ to core data $b_1, \mathbf{A}_{11}$ smaller in size, that contain the necessary and sufficient information for solving the original problem. More discussion about the core reduction is given in Chapter 6, thus we omit the details here.

Remark: The condition $\sigma < \tilde{\sigma}$, where $\tilde{\sigma}$ is the smallest nonzero singular value of $\mathbf{A}$, is often used to guarantee the existence of the solution. This condition is equivalent only to the case that σ is simple and $e_1^T s \neq 0$. The proof follows from interlacing theorem (see [44] p. 35). Therefore this condition is only sufficient but not necessary for existence of the TLS solution.

For more details about the TLS solution in nongeneric case and also in case that σ is a multiple singular value see, e.g., [44], where also efficient numerical methods for computation of TLS solution are discussed. They are usually based on the singular value decomposition. Discussion for the TLS problem with multiple right-hand side can be found, e.g., in [44], [77]. For comparison of the TLS and the LS solution of (5.1) see [91], [92].

Scaled total least squares

All approaches presented above can be unified by considering the following general scaled total least squares (ScTLS) method. For a given real $\gamma > 0$, the ScTLS seeks a vector $g \in \mathbb{R}^n$ and a matrix $\mathbf{E} \in \mathbb{R}^{n \times m}$ such that

$$\begin{aligned} &\text{a) } \tilde{b} \in \mathcal{R}(\gamma \tilde{\mathbf{A}}), \text{ where } \tilde{b} = \gamma b + g,\ \tilde{\mathbf{A}} = \mathbf{A} + \mathbf{E} \\ &\text{b) } [g, \mathbf{E}] \ = \ \arg_{[\tilde{g}, \tilde{\mathbf{E}}]} \min \|[\tilde{g}, \tilde{\mathbf{E}}]\|_F \text{ subjected to a).} \end{aligned} \tag{5.5}$$

When $\gamma \to 0$ the ScTLS solution approaches the LS solution, $\gamma = 1$ coincides with the TLS formulation and when $\gamma \to \infty$ the ScTLS solution approaches data least squares solution, where the correction is allowed only in the matrix $\mathbf{A}$.

5.3 Truncation and regularization methods

Truncation methods for linear approximation problems are applicable when an (usually SVD-based) expansion of the system is computable. The idea is to reduce the rank of the original linear system, i.e. to approximate the system by a well conditioned compatible one with smaller rank l, such that the components of the solution corresponding to unwanted subspaces are ignored.

Consider the singular value decomposition (SVD)

$$\mathbf{A} \ = \ \tilde{\mathbf{R}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{S}}^T \ = \sum_{i=1}^{l} \tilde{r}_i \tilde{\sigma}_i \tilde{s}_i^T,$$

where $l \equiv \text{rank}(\mathbf{A})$, $\tilde{\sigma}_1 \geq \ldots \geq \tilde{\sigma}_l > 0$ are singular values and r_i, s_i left and right singular vectors of $\mathbf{A}$, respectively. Then the LS solution of (5.1) can be expressed in the form

$$x^{LS} = \sum_{i=1}^{l} \frac{\tilde{r}_i^T b}{\tilde{\sigma}_i} \tilde{s}_i. \tag{5.6}$$

The matrix $\mathbf{A}_k \equiv \sum_{i=1}^{k} \tilde{r}_i \tilde{\sigma}_i \tilde{s}_i^T$ is the nearest rank k approximation to the matrix $\mathbf{A}$, see [20], [54]. The simplest type of truncation methods - *truncated LS* (T-LS) (also called truncated SVD) can be obtained by replacing $\mathbf{A}$ by $\mathbf{A}_k$ in (5.2), i.e. the T-LS solution of (5.1) has the form

$$x_k^{T-LS} = \sum_{i=1}^{k} \frac{\tilde{r}_i^T b}{\tilde{\sigma}_i} \tilde{s}_i. \tag{5.7}$$

Similarly the *truncated TLS* (T-TLS) method can be obtained by completing (5.4) by the condition $\operatorname{rank}([\tilde{b}, \tilde{\mathbf{A}}]) = k$. The T-LS and the T-TLS solution are closely connected, see [24].

The expansion (5.6) indicates that the components of the LS solution corresponding to small singular values may be dominated by errors in the right-hand side b and these components can be eliminated by choosing good truncation level k in (5.7). Thus truncation methods as truncated LS, truncated TLS etc. have *regularizing properties* and the important question is how to identify a good truncation level here. The choice is clear if the matrix $\mathbf{A}$ has well defined numerical rank (i.e. if there is a large gap between "small" and "large" singular values) or in applications where a noise level is apriori known. Then it is reasonable to eliminate all information below this level. Unfortunately, in many cases the choice is more complicated, e.g., in ill-posed problems where singular values decay gradually to zero and/or numerical rank of the matrix $\mathbf{A}$ is not well defined. Therefore more sophisticated regularization methods are required to suppress the effect of errors in the data and extract the essential information about the system.

Regularization methods

Regularized (or filtered) solution can be generally formulated as

$$x = \sum_{i=1}^{l} f_i \, \frac{\tilde{r}_i^T b}{\tilde{\sigma}_i} \tilde{s}_i,$$

where f_i are filter factors. For example in T-LS the filter factor are simply $f_i = 1$ for $i = 1, \ldots, k$ and $f_i = 0$ for $i = k+1, \ldots, l$. Thus k can be understood as a filtration or *regularization level*. The form of the filter factors for T-TLS was derived in [25] and indicates the filtering property, even if there is no marked gap between large and small singular values (see pp. 1230-1231). Regularizing properties of this method were widely discussed also in [24], [35].

Many regularization techniques control not only the residual of the solution but also the solution norm. Tikhonov regularization is originally connected with LS. Here the condition b) in (5.2) is replaced by

$$g \;=\; \arg_{\tilde{g} \in \mathbb{R}^n} \min\{\|\tilde{g}\|^2 + \lambda \|\mathbf{L}x\|^2\} \text{ subjected to a)}, \tag{5.8}$$

where the matrix $\mathbf{L}$ is often equal to $\mathbf{I}$ or represents a discretized differential operator, e.g., approximate first order derivative operator, and the number λ is *regularization (or penalty) parameter* controlling the trade-off between the LS distance and the norm of the solution. If $\mathbf{L} = \mathbf{I}$ then

$$x^{Tik} = \sum_{i=1}^{l} \frac{\tilde{\sigma}_i^2}{\tilde{\sigma}_i + \lambda^2} \, \frac{\tilde{r}_i^T b}{\tilde{\sigma}_i} \tilde{s}_i.$$

Tikhonov type regularization can be similarly used to regularize the TLS solution, see, e.g., [28], [39], [71]. In [28] connection between the LS and the TLS solution of Tikhonov type is analyzed. Computation of the Tikhonov solution in case that $\mathbf{L} \neq \mathbf{I}$ usually consists of transforming the problem to standard form (section 2.3 in [35]), then using truncated TLS method to compute the regularized solution and transforming it back to the original variables. Recently also methods working directly with the matrix $\mathbf{L}$ have been developed.

A lot of regularization methods is based on Golub-Kahan (also called Lanczos) bidiagonalization, see [30], [61]. *LSQR* (see [62], [63]) is a frequently used conjugate-gradient type method for solving both sparse linear systems discussed in the first part of this thesis, and sparse least squares problems in the Tikhonov form (5.8) with $\mathbf{L} = \mathbf{I}$. The method is algebraically equivalent to applying symmetric conjugate gradient (CG) method to normal equations

$$(\mathbf{A}^T\mathbf{A} + \lambda^2\mathbf{I})x \; = \; \mathbf{A}^T b$$

corresponding to the problem (5.8), but has better numerical properties, especially if $\mathbf{A}$ is ill conditioned. The idea is the following. First the problem (5.1) is projected on the Krylov subspace using the lower Golub-Kahan bidiagonalization of the matrix $\mathbf{A}$ starting from the vector $u_1 \equiv b/\beta_1$, where $\beta_1 \equiv \|b\| \neq 0$ (for algorithm see Section 6.2). After k steps of the bidiagonalization we obtain

$$\mathbf{A}\mathbf{V}_k \; = \; \mathbf{U}_{k+1}\mathbf{L}_{k+},$$

where $\mathbf{V}_k \in \mathbb{R}^{m\times k}$, $\mathbf{U}_{k+1} \in \mathbb{R}^{n\times(k+1)}$ have orthonormal columns, $u_1 = \mathbf{U}_{k+1}e_1$ and $\mathbf{L}_{k+} \in \mathbb{R}^{(k+1)\times k}$ is lower bidiagonal matrix. The CG method seeks a vector x_k in the kth Krylov subspace $\mathcal{K}_k(\mathbf{A}^T\mathbf{A}, \mathbf{A}^T b) = \operatorname{span}\{\mathbf{V}_k\}$ that minimizes $\|\mathbf{A}x - b\|$. The solution can be written in the form $x_k = \mathbf{V}_k y$ and

$$\min_{x\in\mathcal{K}_k(\mathbf{A}^T\mathbf{A},\mathbf{A}^T b)} \|\mathbf{A}x - b\| = \min_{y\in\mathbb{R}^k} \|\mathbf{L}_{k+}y - \beta_1 e_1\|.$$

The LSQR algorithm in fact produces the bidiagonal decomposition above and the approximation vector x_k is computed recursively from x_{k-1} using only the reduced problem $\mathbf{L}_{k+}y \approx \beta_1 e_1$. *Hybrid methods* developed, e.g., in [7], [9], [34], [35], [60], are based on similar ideas as LSQR. The outer bidiagonalization is combined with an inner regularization applied to the problem $\mathbf{L}_{k+}y \approx \beta_1 e_1$.

Remark: CG is only one of the Krylov subspace methods that have regularizing properties. Regularizing properties of QMR method [26], [76] or GMRES [75] were explored in series of papers, e.g., [17], [18].

5.4 Stopping criteria

If there is no apriori knowledge about the required truncation or regularization level, several stopping criteria can be used. The first is based on estimation of the L-curve (see [38], [35], [36]), i.e. the curve where the norm of the approximate solution $\|x_k\|$ is plotted against the norm of residual error $\|\mathbf{A}x_k - b\|$ for various k and the corner in the log-log scale is chosen. The discrepancy principle requires knowledge about the properties of the noise in the perturbed data. Generalized cross validation combines the residual error with the effective number of used parameters (see, e.g., [29]). These techniques are widely discussed in literature, comparison is given in [35], [78], [77].

Methods used to solve problems arising from errors-in-variables modeling and problems of model reduction are the same, but choice of optimal stopping criteria can be different. Let us concentrate only on methods using bidiagonalization, where the *core theory* may be found useful. Computational method for reducing $b, \mathbf{A}$ to core data $b_1, \mathbf{A}_{11}$ involves (partial) upper bidiagonalization of the matrix $[b, \mathbf{A}]$; more precisely $b_1 \equiv \beta_1 e_1$ and $\mathbf{A}_{11} \equiv \mathbf{L}_p$ if (5.1) is compatible, or $\mathbf{A}_{11} \equiv \mathbf{L}_{p+}$ if (5.1) is incompatible. This transformation is widely described in Chapter 6. If the bidiagonalization is used in model reduction in order to obtain a lower rank approximation to (5.1), it can be stopped at step $k < p$ and a lower rank approximation of rank k can be computed from this partial bidiagonalization.

If the bidiagonalization is used in ill-posed problems, the situation is more complicated. In exact arithmetics the necessary and sufficient information from (5.1) is extracted to $b_1, \mathbf{A}_{11}$ as soon as a *zero value* is encountered either on the diagonal or on the superdiagonal of the matrix $\mathbf{L}$. This gives an important theoretical background for the LSQR and hybrid methods. At each step $k < p$ these techniques compute the system $\mathbf{L}_{k+}y \approx \beta_1 e_1$ that is "approximation" to the core system and thus the approximate solution cannot contain redundant informations. On the other hand, stopping the bidiagonalization process when a zero value is encountered is numerically not suitable and one should ask how to decide about the "truncation level" for the bidiagonalization, i.e. how to numerically indicate the separation of the core problem. Then the question arises whether and when the main submatrix of the bidiagonal matrix $\mathbf{A}_{11}$ can be considered a sufficiently good approximation to the whole core matrix. Many recent results about submatrices of bidiagonal matrices, as interlacing property of the eigenvalues, can give some ideas. In Chapter 6 we give a motivation, how the relationship between the Golub-Kahan bidiagonalization and the Lanczos tridiagonalization, together with the properties of Jacobi matrices can be used in further investigation of efficient stopping criteria for regularization methods.

Chapter 6

Core reduction

In this chapter we concentrate on the concept of core reduction in linear algebraic systems. First we briefly summarize the main idea of the core theory given by C.C. Paige and Z. Strakoš, that shows how the core problem extracts the necessary and sufficient information for solving the original linear algebraic system. The computation is connected with the (partial) bidiagonalization and this fact can be used to solve efficiently the approximation problem with the original data. Then we relate the core problem formulation to the Lanczos tridiagonalization and derive its fundamental characteristics from the relationship between the Golub-Kahan bidiagonalization, the Lanczos tridiagonalization and the well known properties of Jacobi matrices. Finally we outline some directions for further research.

6.1 Basic ideas of core reduction

C.C. Paige and Z. Strakoš proposed, in a sequence of papers [65], [67], [64], to orthogonally transform the original problem (5.1) to the block form that allows to separate the necessary and sufficient information present in the data $b, \mathbf{A}$ from the redundancies. This transformation leads to better understanding of several well known regularization techniques, that were briefly summarized in the previous chapter, and can also lead to the improvement in the stopping criteria for these methods. In this section, we explain the idea of the core reduction and summarize its fundamental characteristics.

Assuming that the approximation problem (5.1) is unitarily invariant, it was proved in [64] that there exists an orthogonal transformation of the form

$$\mathbf{P}^T \left[\, b \,\|\, \mathbf{A}\mathbf{Q} \,\right] = \left[\begin{array}{c||c|c} b_1 & \mathbf{A}_{11} & 0 \\ \hline 0 & 0 & \mathbf{A}_{22} \end{array}\right], \tag{6.1}$$

where $\mathbf{P} \in \mathbb{R}^{n\times n}$, $\mathbf{Q} \in \mathbb{R}^{m\times m}$ are orthogonal matrices, $b_1 = \beta_1 e_1$ and $\mathbf{A}_{11}$ is a lower bidiagonal matrix with nonzero bidiagonal elements. The matrix $\mathbf{A}_{11}$ is either square, when (5.1) is compatible, or rectangular, when (5.1) is incompatible. Depending on the relation between b and $\mathbf{A}$, the matrix $\mathbf{A}_{22}$ and the corresponding block row and/or column in (6.1) can be nonexistent. This situation will be explained in the following section.

Using this transformation, the original problem is decomposed into the approximation problem

$$\mathbf{A}_{11}x_1 \approx b_1\,, \tag{6.2}$$

which contains all necessary and sufficient information for solving the problem (5.1), and the remaining part $\mathbf{A}_{22}x_2 \approx 0$ containing the redundancies from the original data. The problem (6.2) is therefore called a *core problem* within (5.1). In [64], it was suggested to find x_1 from (6.2), set $x_2 \equiv 0$, and substitute

$$x \equiv \mathbf{Q}\begin{bmatrix} x_1 \\ 0 \end{bmatrix} \tag{6.3}$$

for the solution of (5.1). The fact that the problem (6.2) extracts all necessary and sufficient information from the problem (5.1) follows easily from the properties of the orthogonal transformation summarized in the following theorem.

Theorem 6.1: *Let $\mathbf{A} \in \mathbb{R}^{n\times m}$ be a nonzero matrix, $b \in \mathbb{R}^n$ and $\mathbf{A}^T b \neq 0$. Then there exists a decomposition*

$$\mathbf{P}^T\left[\, b \,\|\, \mathbf{A}\mathbf{Q} \,\right] = \left[\begin{array}{c||c|c} b_1 & \mathbf{A}_{11} & 0 \\ \hline 0 & 0 & \mathbf{A}_{22} \end{array}\right],$$

where $\mathbf{P} \in \mathbb{R}^{n\times n}$, $\mathbf{Q} \in \mathbb{R}^{m\times m}$ are orthogonal matrices, $b_1 = \beta_1 e_1$ and $\mathbf{A}_{11}$ is a lower bidiagonal matrix with nonzero bidiagonal elements. Moreover:

(i) The matrix $\mathbf{A}_{11}$ has full column rank and its singular values are simple. Consequently, any zero singular values or multiplicities of singular values that $\mathbf{A}$ has, must appear in $\mathbf{A}_{22}$.

(ii) The matrix $\mathbf{A}_{11}$ has minimal dimensions, and $\mathbf{A}_{22}$ has maximal dimensions, over all orthogonal transformations giving the block structure above, without any additional assumptions on the structure of $\mathbf{A}_{11}$ and b_1.

(iii) All components of $b_1 = \beta_1 e_1$ in the left singular vector subspaces of $\mathbf{A}_{11}$ (i.e. the first components of all left singular vectors of $\mathbf{A}_{11}$) are nonzero.

Proof: Proofs of *(i)–(iii)* are given in [64], see Theorems 2.2, 3.2 and 3.3. They are based on the singular value decomposition of $\mathbf{A}$ and on the properties of the upper bidiagonal form $[b_1, \mathbf{A}_{11}]$ with nonzero bidiagonal elements. □

The properties *(i)* and *(iii)* imply that the TLS solution (5.4) of the system (6.2) always exists and is unique, while the property *(ii)* shows that the system has minimal dimensions. Thus the problem (5.1) can always be reduced to the core problem (6.2) of minimal dimensions that is uniquely solvable.

Computation of core problem

The core problem (6.2) can be obtained from the singular value decomposition of the extended matrix $[b, \mathbf{A}]$, but for computational efficiency reasons a bidiagonal transformation is usually used. The transformed data b_1 and $\mathbf{A}_{11}$ can be computed by the (possibly partial) upper bidiagonalization of the matrix $[b, \mathbf{A}]$. If $\mathbf{A}$ has small dimensions, bidiagonalization can be performed directly using Householder transformations [31]. Otherwise, when $\mathbf{A}$ is large and sparse, Golub-Kahan bidiagonalization (see [30], [61]) is suggested in [64] as the algorithm for computing the core problem. The bidiagonalization is stopped at the first zero element, giving the block structure in (6.1). Note that the remaining part (giving the matrix $\mathbf{A}_{22}$) need not be bidiagonalized, because it is not needed for computation of the solution (6.3).

If the Golub-Kahan bidiagonalization is used, at any iteration step the computed left principal part of $\mathbf{A}_{11}$ represents an approximation to the core problem matrix. Practical applications may require stopping the computation before the full decomposition (6.1) is reached, see discussion in Section 5.4. Therefore it is important to study iterative approximations to the core problem decomposition. It is well known, that the Golub-Kahan bidiagonalization is closely related to the Lanczos tridiagonalization [48], which has been throughly investigated as a tool for computation of a few dominant eigenvalues. We believe that the knowledge about the partial Lanczos tridiagonalization may prove useful in future investigation of the partial core problem decomposition. Therefore, in the following section, we briefly summarize the relationship of the core problem decomposition with the Lanczos tridiagonalization. In Section 6.3, we present a new proof of the fundamental characteristics *(i)–(iii)* of the core problem formulated in Theorem 6.1, from the connection between the Golub-Kahan bidiagonalization and the Lanczos tridiagonalization.

6.2 Connection between core problem and Lanczos tridiagonalization

Consider the *partial* lower Golub-Kahan bidiagonalization of the matrix $\mathbf{A}$ in the following form. Given the initial vectors $v_0 \equiv 0, u_1 \equiv b/\beta_1$, where $\beta_1 \equiv \|b\| \neq 0$, the algorithm computes for $i = 1, 2, \ldots$

$$\alpha_i v_i = \mathbf{A}^T u_i - \beta_i v_{i-1}, \quad \|v_i\| = 1, \tag{6.4}$$

$$\beta_{i+1} u_{i+1} = \mathbf{A} v_i - \alpha_i u_i, \quad \|u_{i+1}\| = 1 \tag{6.5}$$

until $\alpha_i = 0$ or $\beta_{i+1} = 0$, or until $i = \min\{n, m\}$.

We present, for completeness, the basic properties of the Golub-Kahan bidiagonalization as given in [61]. Consider $\alpha_i \beta_i \neq 0$ for $i = 1, \ldots, k+1$ and denote by $\mathbf{U}_k \equiv (u_1, \ldots, u_k), \mathbf{V}_k \equiv (v_1, \ldots, v_k)$,

$$\mathbf{L}_k \equiv \begin{pmatrix} \alpha_1 & & & \\ \beta_2 & \alpha_2 & & \\ & \ddots & \ddots & \\ & & \beta_k & \alpha_k \end{pmatrix}, \quad \mathbf{L}_{k+} \equiv \begin{pmatrix} \mathbf{L}_k \\ \beta_{k+1} e_k^T \end{pmatrix}.$$

Then (6.4)–(6.5) can be rewritten in the matrix form

$$\begin{aligned} \mathbf{A}^T \mathbf{U}_k &= \mathbf{V}_k \mathbf{L}_k^T, & (6.6)\\ \mathbf{A}\mathbf{V}_k &= [\mathbf{U}_k, u_{k+1}]\, \mathbf{L}_{k+}, & (6.7) \end{aligned}$$

giving

$$\begin{aligned} \mathbf{U}_k^T \mathbf{A}\mathbf{V}_k &= (\mathbf{A}^T \mathbf{U}_k)^T \mathbf{V}_k = \mathbf{L}_k \mathbf{V}_k^T \mathbf{V}_k \\ &= \mathbf{U}_k^T\, [\mathbf{U}_k, u_{k+1}]\, \mathbf{L}_{k+} = \mathbf{U}_k^T \mathbf{U}_k \mathbf{L}_k \;+\; \beta_{k+1} \mathbf{U}_k^T u_{k+1} e_k^T, \end{aligned}$$

and thus

$$\mathbf{L}_k \mathbf{V}_k^T \mathbf{V}_k \;=\; \mathbf{U}_k^T \mathbf{U}_k \mathbf{L}_k + \beta_{k+1} \mathbf{U}_k^T u_{k+1} e_k^T. \qquad (6.8)$$

Similarly, (6.4) gives for $i = k+1$

$$\mathbf{A}^T [\mathbf{U}_k, u_{k+1}] \;=\; \mathbf{V}_k \mathbf{L}_{k+}^T + \alpha_{k+1} v_{k+1} e_{k+1}^T, \qquad (6.9)$$

and therefore

$$\begin{aligned} \mathbf{V}_k^T \mathbf{A}^T [\mathbf{U}_k, u_{k+1}] &= \mathbf{V}_k^T \mathbf{V}_k \mathbf{L}_{k+}^T + \alpha_{k+1} \mathbf{V}_k^T v_{k+1} e_{k+1}^T \\ &= (\mathbf{A}\mathbf{V}_k)^T [\mathbf{U}_k, u_{k+1}] \;=\; \mathbf{L}_{k+}^T [\mathbf{U}_k, u_{k+1}]^T [\mathbf{U}_k, u_{k+1}], \end{aligned}$$

which yields

$$\mathbf{L}_{k+}^T [\mathbf{U}_k, u_{k+1}]^T [\mathbf{U}_k, u_{k+1}] \;=\; \mathbf{V}_k^T \mathbf{V}_k \mathbf{L}_{k+}^T + \alpha_{k+1} \mathbf{V}_k^T v_{k+1} e_{k+1}^T. \qquad (6.10)$$

As a direct consequence we get the following well known properties of the matrices $\mathbf{U}_k$ and $\mathbf{V}_k$.

Lemma 6.2: *Assume that algorithm (6.4)–(6.5) does not stop before step $k+1$. Then it generates the vectors $u_1, u_2, \ldots, u_{k+1}$ and $v_1, v_2, \ldots, v_{k+1}$ such that $u_i^T u_j = v_i^T v_j = 0$ for $i \neq j$.*

Proof: Follows immediately by induction. The induction assumption $\mathbf{U}_k^T \mathbf{U}_k = \mathbf{I}$, $\mathbf{V}_k^T \mathbf{V}_k = \mathbf{I}$ gives from (6.8)

$$\mathbf{L}_k \;=\; \mathbf{L}_k + \beta_{k+1} \mathbf{U}_k^T u_{k+1} e_k^T,$$

and thus $\mathbf{U}_k^T u_{k+1} = 0$, because $\beta_{k+1} \neq 0$. Similarly, (6.10) and $\alpha_{k+1} \neq 0$ yield

$$\mathbf{L}_{k+}^T \;=\; \mathbf{L}_{k+}^T + \alpha_{k+1}\mathbf{V}_k^T v_{k+1} e_{k+1}^T,$$

that gives $\mathbf{V}_k^T v_{k+1} = 0$. □

Summarizing, the Golub-Kahan bidiagonalization (6.4)–(6.5) of the matrix $\mathbf{A}$ with $u_1 = b/\|b\|$ results in one of the two following situations, which will be distinguished throughout this chapter:

Case 1. $\alpha_i\beta_i \neq 0$ for $i = 1, \dots, p$; $\beta_{p+1} = 0$ or $p = n$. Then (6.6) gives

$$\mathbf{U}_p^T \mathbf{A}\mathbf{V}_p = \mathbf{L}_p,$$

$$\mathbf{U}_p^T [b, \mathbf{A}\mathbf{V}_p] = \left[\begin{array}{c|cccc} \beta_1 & \alpha_1 & & & \\ & \beta_2 & \alpha_2 & & \\ & & \ddots & \ddots & \\ & & & \beta_p & \alpha_p \end{array}\right] \equiv [b_1 | \mathbf{A}_{11}] \qquad (6.11)$$

and $\mathbf{A}_{11}x_1 \equiv \mathbf{L}_p\, x_1 \approx \beta_1 e_1 \equiv b_1$ is the *compatible* core problem. The matrices $\mathbf{U}_p, \mathbf{V}_p$ represent the first p columns of the matrices $\mathbf{P}, \mathbf{Q}$ respectively, see (6.1).

Case 2. $\alpha_i\beta_i \neq 0$ for $i = 1, \dots, p$, and $\beta_{p+1} \neq 0$; $\alpha_{p+1} = 0$ or $p = m$. Then (6.7) gives

$$[\mathbf{U}_p, u_{p+1}]^T \mathbf{A}\mathbf{V}_p = \mathbf{L}_{p+},$$

$$[\mathbf{U}_p, u_{p+1}]^T [b, \mathbf{A}\mathbf{V}_p] = \left[\begin{array}{c|cccc} \beta_1 & \alpha_1 & & & \\ & \beta_2 & \alpha_2 & & \\ & & \ddots & \ddots & \\ & & & \beta_p & \alpha_p \\ & & & & \beta_{p+1} \end{array}\right] \equiv [b_1 | \mathbf{A}_{11}] \quad (6.12)$$

and $\mathbf{A}_{11}x_1 \equiv \mathbf{L}_{p+}\, x_1 \approx \beta_1 e_1 \equiv b_1$ is the *incompatible* core problem. The matrices $\mathbf{U}_{p+1}$ and $\mathbf{V}_p$ represent the first $(p+1)$ and p columns of the matrices $\mathbf{P}$ and $\mathbf{Q}$ respectively.

Remark: For clarity of exposition we review the situations when the bidiagonalization is not stopped until the maximum number of steps is reached. If $p = n = m$, then $\mathbf{U}_p = \mathbf{U}_n = \mathbf{P}$, $\mathbf{V}_p = \mathbf{V}_m = \mathbf{Q}$ and

$$\mathbf{P}^T \left[\; b \,\|\, \mathbf{A}\mathbf{Q} \;\right] = \left[\; b_1 \,\|\, \mathbf{A}_{11} \;\right].$$

If $p = n < m$, then $\mathbf{U}_p = \mathbf{U}_n = \mathbf{P}$, and completing $\mathbf{V}_p$ by $(m - n)$ additional columns into the orthogonal matrix $\mathbf{Q}$ gives

$$\mathbf{P}^T \left[\; b \,\|\, \mathbf{A}\mathbf{Q} \;\right] = \left[\; b_1 \,\|\, \mathbf{A}_{11} \,|\, 0 \;\right].$$

If $p = m < n$, then $\mathbf{V}_p = \mathbf{V}_m = \mathbf{Q}$, and completing $\mathbf{U}_{p+1}$ by $(n-m-1)$ additional columns into the orthogonal matrix $\mathbf{P}$ gives

$$\mathbf{P}^T \left[\ b \,\|\, \mathbf{AQ}\ \right] = \left[\begin{array}{c||c} b_1 & \mathbf{A}_{11} \\ \hline 0 & 0 \end{array}\right].$$

Lanczos tridiagonalization

First we remind the basic terminology used in this section.

Definition 6.1: *Let $\mathbf{B} \in \mathbb{R}^{t\times t}$ be a tridiagonal matrix with elements $b_{i+1,i} \neq 0$, $b_{i,i+1} \neq 0$ for $i = 1, \dots, t-1$. Then it is called a Jacobi matrix.*

The bidiagonalization algorithm is closely connected with the Lanczos tridiagonalization. Let $\mathbf{B} \in \mathbb{R}^{t\times t}$ be a symmetric matrix. Given the initial vector $w_1 \in \mathbb{R}^t$ such that $\|w_1\| = 1$; $w_0 \equiv 0$, $\delta_1 \equiv 0$, the partial tridiagonalization algorithm computes for $i = 1, 2, \dots$

$$y_i = \mathbf{B}w_i - \delta_i w_{i-1}, \tag{6.13}$$

$$\gamma_i = (y_i, w_i), \tag{6.14}$$

$$\delta_{i+1} w_{i+1} = y_i - \gamma_i w_i, \quad \|w_{i+1}\| = 1 \tag{6.15}$$

until $\delta_{i+1} = 0$, or until $i + 1 = t$. Consider $\delta_i \neq 0$ for $i = 1, \dots, k+1$ and denote by $\mathbf{W}_k \equiv (w_1, \dots, w_k)$,

$$\mathbf{T}_k \equiv \begin{pmatrix} \gamma_1 & \delta_2 & & \\ \delta_2 & \gamma_2 & \ddots & \\ & \ddots & \ddots & \delta_k \\ & & \delta_k & \gamma_k \end{pmatrix}.$$

Then $\mathbf{W}_k$ has orthonormal columns and $\mathbf{T}_k$ represents the symmetric tridiagonal matrix with positive elements on the subdiagonal, i.e. a Jacobi matrix. The Lanczos algorithm can be written in the matrix form

$$\mathbf{BW}_k = \mathbf{W}_k\mathbf{T}_k + \delta_{k+1} w_{k+1} e_k^T, \qquad \mathbf{W}_k^T w_{k+1} = 0. \tag{6.16}$$

Given a real symmetric $\mathbf{B}$, (6.16) is fully determined by the starting vector w_1. Moreover, the properties of Jacobi matrices yield the fundamental properties of the matrix $\mathbf{T}_k$.

Lemma 6.3: *Let $\mathbf{B} \in \mathbb{R}^{t\times t}$ be a symmetric matrix, $w_1 \in \mathbb{R}^t$ and $\|w_1\| = 1$. Assume that algorithm (6.13)–(6.15) does not stop before step k. Then:*

(I) The matrix $\mathbf{T}_k$ has distinct eigenvalues.

(II) *If* $\mathbf{B}$ *is real symmetric positive semidefinite and* $w_1 \perp \ker(\mathbf{B})$, *then all eigenvalues of* $\mathbf{T}_k$ *are positive.*

(III) The first (as well as the last) components of all eigenvectors of $\mathbf{T}_k$ *are nonzero.*

Proof: The properties *(I)* and *(III)* are the basic properties of Jacobi matrices, see [68] Lemma 7.7.1, Theorem 7.9.3. Further, *(II)* follows from the fact that the final Jacobi matrix $\mathbf{T}_l$, for which $\mathbf{B}\mathbf{W}_l = \mathbf{W}_l\mathbf{T}_l$, must be nonsingular (and, using the assumption in *(II)*, symmetric positive definite) and from the interlacing property (see [68] Theorem 10.1.1). □

The relationship between the Lanczos tridiagonalization and the Golub-Kahan bidiagonalization can be described in several ways, see [7] pp. 662–663, [9] pp. 513–515, [30] pp. 212–214 and also [61] pp. 199–200, [47] pp. 115–118. Consider the coefficients of the Golub-Kahan bidiagonalization $\alpha_i\beta_i \neq 0$ for $i = 1, \ldots, k+1$. Then, (6.6) multiplied by $\mathbf{A}$ and combined with (6.7) gives

$$\mathbf{A}\mathbf{A}^T\,\mathbf{U}_k \;=\; \mathbf{A}\mathbf{V}_k\,\mathbf{L}_k^T = [\mathbf{U}_k, u_{k+1}]\,\mathbf{L}_{k+}\mathbf{L}_k^T = \mathbf{U}_k\,\mathbf{L}_k\mathbf{L}_k^T \;+\; \alpha_k\beta_{k+1}\,u_{k+1}e_k^T, \qquad (6.17)$$

where

$$\mathbf{L}_k\mathbf{L}_k^T = \begin{pmatrix} \alpha_1^2 & \alpha_1\beta_1 & & \\ \alpha_1\beta_2 & \alpha_2^2+\beta_2^2 & \ddots & \\ & \ddots & \ddots & \alpha_{k-1}\beta_k \\ & & \alpha_{k-1}\beta_k & \alpha_k^2+\beta_k^2 \end{pmatrix}.$$

In short, (6.17) represents k steps of the Lanczos tridiagonalization of the matrix $\mathbf{A}\mathbf{A}^T$ with the starting vector $u_1 = b/\beta_1 = b/\|b\|$. Here, according to the notation in (6.16), we have $\mathbf{B}^{(1)} \equiv \mathbf{A}\mathbf{A}^T$, $\mathbf{W}_k^{(1)} \equiv \mathbf{U}_k$, $\mathbf{T}_k^{(1)} \equiv \mathbf{L}_k\mathbf{L}_k^T$ and $\delta_{k+1}^{(1)} \equiv \alpha_k\beta_{k+1}$. Similarly, (6.7) together with (6.9) gives

$$\mathbf{A}^T\mathbf{A}\,\mathbf{V}_k \;=\; \mathbf{A}^T\,[\mathbf{U}_k, u_{k+1}]\,\mathbf{L}_{k+} = \mathbf{V}_k\,\mathbf{L}_{k+}^T\mathbf{L}_{k+} \;+\; \alpha_{k+1}\beta_{k+1}\,v_{k+1}e_k^T, \qquad (6.18)$$

where

$$\mathbf{L}_{k+}^T\mathbf{L}_{k+} = \mathbf{L}_k^T\mathbf{L}_k + \beta_{k+1}^2 e_k e_k^T = \begin{pmatrix} \alpha_1^2+\beta_2^2 & \alpha_2\beta_2 & & \\ \alpha_2\beta_2 & \alpha_2^2+\beta_3^2 & \ddots & \\ & \ddots & \ddots & \alpha_k\beta_k \\ & & \alpha_k\beta_k & \alpha_k^2+\beta_{k+1}^2 \end{pmatrix}.$$

The identity (6.18) represents k steps of the Lanczos tridiagonalization of the matrix $\mathbf{A}^T\mathbf{A}$ with the starting vector $v_1 = \mathbf{A}^T u_1/\alpha_1 = \mathbf{A}^T b/\|\mathbf{A}^T b\|$. Here we have $\mathbf{B}^{(2)} \equiv \mathbf{A}^T\mathbf{A}$, $\mathbf{W}_k^{(2)} \equiv \mathbf{V}_k$, $\mathbf{T}_k^{(2)} \equiv \mathbf{L}_{k+}^T\mathbf{L}_{k+}$ and $\delta_{k+1}^{(2)} \equiv \alpha_{k+1}\beta_{k+1}$.

Remark: The relationship between (6.4)-(6.5) and (6.16) can be similarly described using the following relation. The Lanczos tridiagonalization applied to the augmented matrix

$$\mathbf{B} \equiv \begin{pmatrix} 0 & \mathbf{A} \\ \mathbf{A}^T & 0 \end{pmatrix}$$

with the starting vector $w_1 \equiv (u_1, 0)^T$ yields in $2k$ steps the orthogonal matrix

$$\mathbf{W}_{2k} = \begin{pmatrix} u_1 & 0 & \dots & u_k & 0 \\ 0 & v_1 & \dots & 0 & v_k \end{pmatrix}$$

and the Jacobi matrix $\mathbf{T}_{2k}$ with the zero main diagonal and the subdiagonals equal to $(\alpha_1, \beta_2, \dots, \beta_k, \alpha_k)$.

6.3 Proof of the core problem characteristics

In this section, we prove Theorem 6.1 by relating the characteristics *(i)–(iii)* of the core problem to the well known properties of the Lanczos tridiagonalization, and the properties *(I)–(III)* of the Jacobi matrices, see Lemma 6.3. We distinguish two cases described in the previous section.

Case 1. $\alpha_i\beta_i \neq 0$ for $i = 1, \dots, p$; $\beta_{p+1} = 0$ or $p = n$ (i.e. $n \leq m$), see (6.11). The square matrix $\mathbf{A}_{11} \equiv \mathbf{L}_p$ represents a Cholesky factor of $\mathbf{T}_p^{(1)} \equiv \mathbf{L}_p\mathbf{L}_p^T$, which we see by (6.17) results from the Lanczos tridiagonalization of $\mathbf{B}^{(1)} \equiv \mathbf{A}\mathbf{A}^T$ with the starting vector $u_1 = b/\|b\|$, which stops exactly in p steps, i.e.

$$\mathbf{A}\mathbf{A}^T\,\mathbf{U}_p = \mathbf{U}_p\,\mathbf{L}_p\mathbf{L}_p^T. \tag{6.19}$$

Consider the singular value decomposition $\mathbf{L}_p = \mathbf{R}\mathbf{\Sigma}\mathbf{S}^T$, where $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, .., \sigma_p)$, and $\mathbf{R}, \mathbf{S} \in \mathbb{R}^{p \times p}$ are orthogonal matrices. Then

$$\mathbf{T}_p^{(1)} = \mathbf{L}_p\mathbf{L}_p^T = \mathbf{R}\mathbf{\Sigma}^2\mathbf{R}^T$$

is the spectral decomposition of the matrix $\mathbf{T}_p^{(1)}$, σ_i^2 are its eigenvalues and $r_i = \mathbf{R}e_i$ its eigenvectors, $i = 1, \dots, p$. Consequently, from *(I)* the values σ_i^2 are distinct and thus the singular values of $\mathbf{L}_p$ are distinct. The matrix $\mathbf{L}_p$ is square with positive elements on its diagonal. Therefore all its singular values must be positive, which proves *(i)*. Moreover *(iii)* follows from *(III)*, since

$$b_1^T r_i = \beta_1 e_1^T r_i \neq 0 \qquad \text{for} \quad i = 1, \dots, p.$$

The minimality property *(ii)* can be proved by contradiction. For some orthogonal matrices $\tilde{\mathbf{P}}$, $\tilde{\mathbf{Q}}$ let

$$\tilde{\mathbf{P}}^T[\; b \,\|\, \mathbf{A}\tilde{\mathbf{Q}} \;] = \left[\begin{array}{c||c|c} \tilde{b}_1 & \tilde{\mathbf{A}}_{11} & 0 \\ \hline 0 & 0 & \tilde{\mathbf{A}}_{22} \end{array}\right],$$

where $\tilde{\mathbf{A}}_{11} \in \mathbb{R}^{q\times q}$ with $q < p$. (The system (6.1) is compatible, see (6.11), and therefore, for example by considering the QR–decomposition of $\tilde{\mathbf{A}}_{11}$, we can with no loss of generality assume that $\tilde{\mathbf{A}}_{11}$ is square.) Substituting

$$\mathbf{A} \,=\, \tilde{\mathbf{P}} \left[\begin{array}{c|c} \tilde{\mathbf{A}}_{11} & 0 \\ \hline 0 & \tilde{\mathbf{A}}_{22} \end{array}\right] \tilde{\mathbf{Q}}^T$$

into the Lanczos tridiagonalization (6.19) gives

$$\tilde{\mathbf{P}} \left[\begin{array}{c|c} \tilde{\mathbf{A}}_{11} & 0 \\ \hline 0 & \tilde{\mathbf{A}}_{22} \end{array}\right] \left[\begin{array}{c|c} \tilde{\mathbf{A}}_{11}^T & 0 \\ \hline 0 & \tilde{\mathbf{A}}_{22}^T \end{array}\right] \tilde{\mathbf{P}}^T \mathbf{U}_p \,=\, \mathbf{U}_p \, \mathbf{T}_p^{(1)},$$

i.e.

$$\left[\begin{array}{c|c} \tilde{\mathbf{A}}_{11}\tilde{\mathbf{A}}_{11}^T & 0 \\ \hline 0 & \tilde{\mathbf{A}}_{22}\tilde{\mathbf{A}}_{22}^T \end{array}\right] (\tilde{\mathbf{P}}^T\mathbf{U}_p) \,=\, (\tilde{\mathbf{P}}^T\mathbf{U}_p)\, \mathbf{T}_p^{(1)}, \tag{6.20}$$

with

$$\tilde{\mathbf{P}}^T u_1 \,=\, \tilde{\mathbf{P}}^T b/\|b\| \,=\, \begin{pmatrix} \tilde{b}_1/\|b\| \\ 0 \end{pmatrix}.$$

Since $\tilde{\mathbf{A}}_{11}\tilde{\mathbf{A}}_{11}^T \in \mathbb{R}^{q\times q}$ and $\tilde{b}_1 \in \mathbb{R}^q$, the Lanczos tridiagonalization represented by (6.20) must stop in at most q steps, and $\mathbf{T}_p^{(1)}$ must have $\delta_{q+1}^{(1)} = 0$, which contradicts the fact that $\mathbf{T}_p^{(1)}$ is a Jacobi matrix.

Case 2. $\alpha_i\beta_i \neq 0$ for $i = 1, \ldots, p$, and $\beta_{p+1} \neq 0$; $\alpha_{p+1} = 0$ or $p = m$ (i.e. $n \geq m$), see (6.12). The rectangular matrix $\mathbf{A}_{11} \equiv \mathbf{L}_{p+}$ can be linked to the matrix $\mathbf{T}_p^{(2)} \equiv \mathbf{L}_{p+}^T\mathbf{L}_{p+}$, which we see by (6.18) results from the Lanczos tridiagonalization of $\mathbf{B}^{(2)} \equiv \mathbf{A}^T\mathbf{A}$ with the starting vector $v_1 = \mathbf{A}^T b/\|\mathbf{A}^T b\|$. It stops exactly in p steps, i.e.

$$\mathbf{A}^T\mathbf{A}\,\mathbf{V}_p = \mathbf{V}_p\,\mathbf{L}_{p+}^T\mathbf{L}_{p+}. \tag{6.21}$$

Consider the singular value decomposition $\mathbf{L}_{p+} = \mathbf{R}\boldsymbol{\Sigma}\mathbf{S}^T$, where $\mathbf{R} \in \mathbb{R}^{(p+1)\times p}$ is now a rectangular matrix with orthonormal columns, $\mathbf{S} \in \mathbb{R}^{p\times p}$ is orthogonal matrix. Then

$$\mathbf{T}_p^{(2)} = \mathbf{L}_{p+}^T\mathbf{L}_{p+} = \mathbf{S}\boldsymbol{\Sigma}^2\mathbf{S}^T$$

is the spectral decomposition of the matrix $\mathbf{T}_p^{(2)}$, σ_i^2 are its eigenvalues and $s_i = \mathbf{S}e_i$ its eigenvectors, $i = 1, \ldots, p$. Similarly to the previous case, from *(I)* it follows that the singular values of $\mathbf{L}_{p+}$ are distinct. Since by construction v_1 does not have any nonzero component in the nullspace of $\mathbf{A}^T\mathbf{A}$, *(II)* yields that the singular values of $\mathbf{L}_{p+}$ are positive, which proves *(i)*. Moreover, $e_1^T s_i \neq 0$ by *(III)*, $i = 1, \ldots, p$. Considering $\mathbf{L}_{p+}\mathbf{S} = \mathbf{R}\boldsymbol{\Sigma}$ and the fact that $\mathbf{L}_{p+}$ is lower bidiagonal with nonzero bidiagonal elements, $e_1^T r_i \neq 0,\ i = 1, \ldots, p$. Consequently

$$b_1^T\, r_i = \beta_1 e_1^T\, r_i \neq 0 \qquad \text{for} \quad i = 1, \ldots, p$$

which proves *(iii)*.

The minimality property *(ii)* can be proved by contradiction, similarly to the previous Case 1. For some orthogonal matrices $\hat{\mathbf{P}}$, $\hat{\mathbf{Q}}$ let

$$\hat{\mathbf{P}}^T [\, b \,\|\, \mathbf{A}\hat{\mathbf{Q}} \,] = \left[\begin{array}{c||c|c} \hat{b}_1 & \hat{\mathbf{A}}_{11} & 0 \\ \hline 0 & 0 & \hat{\mathbf{A}}_{22} \end{array}\right],$$

where $\hat{\mathbf{A}}_{11} \in \mathbb{R}^{(q+1)\times q}$ with $q < p$. (The system (6.1) is incompatible and therefore we can with no loss of generality assume that $\hat{\mathbf{A}}_{11}$ is rectangular of the given dimensions.) Substituting

$$\mathbf{A} \;=\; \hat{\mathbf{P}} \left[\begin{array}{c|c} \hat{\mathbf{A}}_{11} & 0 \\ \hline 0 & \hat{\mathbf{A}}_{22} \end{array}\right] \hat{\mathbf{Q}}^T$$

into the Lanczos tridiagonalization (6.21) gives

$$\left[\begin{array}{c|c} \hat{\mathbf{A}}_{11}^T\hat{\mathbf{A}}_{11} & 0 \\ \hline 0 & \hat{\mathbf{A}}_{22}^T\hat{\mathbf{A}}_{22} \end{array}\right] (\hat{\mathbf{Q}}^T\mathbf{V}_p) \;=\; (\hat{\mathbf{Q}}^T\mathbf{V}_p)\, \mathbf{T}_p^{(2)}, \qquad (6.22)$$

with

$$\hat{\mathbf{Q}}^T v_1 \;=\; \hat{\mathbf{Q}}^T\mathbf{A}^T b/\|\mathbf{A}^T b\| \;=\; \left[\begin{array}{c|c} \hat{\mathbf{A}}_{11}^T & 0 \\ \hline 0 & \hat{\mathbf{A}}_{22}^T \end{array}\right] \hat{\mathbf{P}}^T b/\|\mathbf{A}^T b\| \;=\; \left(\begin{array}{c} \hat{\mathbf{A}}_{11}^T \hat{b}_1/\|\mathbf{A}^T b\| \\ 0 \end{array} \right),$$

which leads to a contradiction exactly in the same way as in Case 1.

Summarizing, we have shown in this chapter that the fundamental properties of the core problem can be proved in an elegant way without using the singular value decomposition of the whole matrix $[b, \mathbf{A}]$. Here the Golub-Kahan bidiagonalization and the Lanczos tridiagonalization are used as very strong mathematical tools for constructing proofs.

Possible directions for further research

The presented relationship may be found useful in applications of the core problem formulation discussed in the previous chapter, in particular in connection with large ill-posed problems. From the core problem point of view one should particularly ask whether and when the matrix $\mathbf{L}_{k+}$ for $k < p$ (possibly $k \ll p$) can be considered a *sufficiently good approximation* to the core matrix $\mathbf{L}_{p+}$. When $p \ll m$, one must ask how to *numerically* indicate the separation of the core problem, since in finite precision computation α_{p+1} will hardly be identically zero. Similarly, one can ask when the tridiagonal matrix $\mathbf{T}_k$, $k < p$, sufficiently approximates the matrix $\mathbf{T}_p$ discussed above. It might be useful to study in this context perturbation theory of Jacobi matrices, in particular the specific perturbations when the off-diagonal element $\delta_{k+1} = \alpha_{k+1}\beta_{k+1}$ is replaced by zero, see [46].

Conclusions

In part I of this thesis, we have discussed techniques for constructing invariant subspaces of a general square matrix $\mathbf{A}$. First, we have considered the frequently used IRA process with shifts beeing the smallest Ritz or harmonic Ritz values of $\mathbf{A}$, see [81]. Then we have presented an alternative technique in which the roots of transformed and scaled Tchebychev polynomials are taken for the shifts in the IRA process. The convergence of the IRA process has been studied in [81] for some special cases, see also [49]. We have generalized some of these results and described the convergence by the angle between the updated starting vector of the Arnoldi process and the searched invariant subspace, and by the convergence of subdiagonal elements of the upper Hessenberg matrix to zero. We have tested assumptions of these theorems on numerical example to demonstrate their fulfillment.

We have analyzed the case when the smallest in magnitude eigenvalue λ of $\mathbf{A}$ has geometric multiplicity greater then one. We have shown that in this case the IRA process produces a Hessenberg matrix with a small subdiagonal element on the position corresponding to the dimension of some of the Jordan blocks corresponding to λ. Thus it is not always possible to construct the invariant subspace of prescribed dimension and it is important to modify the stopping criterion for determining an invariant subspace. The error analysis is an extra open problem, which is not solved here. However, the discussion based on our observation indicates how to avoid the difficulties with defective and/or derogatory matrices in practice.

We have compared the classical IRA process and the new technique numerically and we have found that on the demonstrated examples the new technique is usually more efficient. The remaining issue is how to construct Tchebychev filters in some special cases, e.g., for matrices having eigenvalues with a small real part and a very big imaginary part.

In Part II, we have considered the concept of core reduction in general algebraic unitarily invariant linear approximation problem. We have presented alternative proofs of fundamental properties of the core problem based on the relationship between the Golub-Kahan bidiagonalization, the Lanczos tridiagonalization and the well known properties of Jacobi matrices. We have discussed possible applications of the core problem formulation; especially in regularization methods for solving large scale ill-posed problems, where the outer Golub-Kahan bidiagonalization is combined with an inner regularization applied to the reduced problem. We have outlined possible directions for further research and formulated several open questions arising from the core theory. We believe that the relationships presented in this thesis, together with known results on Jacobi matrices, can be used in further investigation of effective stopping criteria in regularization methods.

Bibliography

[1] M. Arioly, V. Pták, Z. Strakoš: Krylov sequences of maximal length and convergence of GMRES. BIT **38** (1998), pp. 636–643.

[2] A.H. Baker, E.R. Jessup, T. Manteuffel: A technique for accelerating the convergence of restarted GMRES. SIAM J. Matrix Anal. Appl. **26** (2005), pp. 962–984.

[3] J. Baglama, D. Calvetti, G.H. Golub, L. Reichel: Adaptively preconditioned GMRES algorithms. SIAM J. Sci. Comput. **20** (1998), pp. 243–269.

[4] Ch. Beattie, M. Embree, J. Rossi: Convergence of restarted Krylov subspaces to invariant subspaces. SIAM J. Matrix Anal. Appl. **25** (2004), pp. 1074–1109.

[5] M. Bennani, T. Braconnier: Comparative behaviour of eigensolvers on highly nonnormal matrices. Tech. Rep. TR-PA-93-23, CERFACS, Toulouse, France 1994.

[6] A. Björck: *Numerical Methods for Least Squares Problems*. Philadelphia, SIAM second ed 1996.

[7] A. Björck: A bidiagonalization algorithm for solving large and sparse ill-posed systems of linear equations. BIT **28** (1988), pp. 659–670.

[8] A. Björck: Solving linear least squares problems by Gram-Schmidt orthogonalization. BIT **7** (1967), pp. 1–21.

[9] A. Björck, E. Grimme, P. Van Dooren: An implicit shift bidiagonalization algorithm for ill-posed systems. BIT **34** (1994), pp. 510–534.

[10] A. Björck, C.C. Paige: Loss and recapture of orthogonality in the modified Gram-Schmidt algorithm. SIAM J. Matrix Anal. Appl. **13** (1992), pp. 176–190.

[11] T. Braconnier, F. Chatelin, V. Fraysse: The influence of large nonnormality on the quality of convergence of iterative methods in linear algebra. Tech. Rep. TR-PA-94-07, CERFACS, Toulouse, France 1994.

[12] K. Burrage, J. Erhel: On the performance of various adaptive preconditioned GMRES strategies. Numer. Lin. Algebra with Appl. **5** (1998), pp. 101–121.

[13] B. Carpentieri, I.S. Duff, L. Giraud: A class of spectral two-level preconditioners. SIAM J. Scient. Comp. **25** (2003), pp. 749–765.

[14] A. Chapman, Y. Saad: Deflated and augment Krylov subspace techniques. Numer. Lin. Algebra with Appl. **4** (1997), pp. 43–66.

[15] F. Chaitin-Chatelin, V. Toumazou, E. Traviesas: Accuracy assessment for eigencomputations: Variety of backward errors and pseudospectra. Lin. Alg. Appl. **309** (2000), pp. 73–83.

[16] F. Chatelin: *Eigenvalues of Matrices.* New York, J. Wiley and Sons 1993.

[17] D. Calvetti, B. Lewis, L. Reichel: Gmres, L-curves and discrete ill-posed problems. BIT **42** (2002), pp. 44–65.

[18] D. Calvetti, B. Lewis, L. Reichel: On the regularizing properties of the GMRES method. Numer. Math. **91** (2002), pp. 605-625.

[19] J. Drkošová, A. Greenbaum, M. Rozložník, Z. Strakoš: Numerical stability of GMRES. BIT, **35** (1995), pp. 309–330.

[20] G. Eckart, G. Young: The approximation of one matrix by another of lower rank. Psychometrica **1** (1936), pp. 211–218.

[21] J. Erhel, K. Burrage, B. Pohl: Restarted GMRES preconditioned by deflation. J. Comput. and Appl. Math. **69** (1996), pp. 303–318.

[22] M. Eiermann, O. G. Ernst: Geometric aspects in the theory of Krylov subspace methods. Acta Numerica (2001), pp. 251–312.

[23] M. Eiermann, O. G. Ernst, O. Schneider: Analysis of acceleration strategies for restarted minimal residual methods. J. Comput. and Appl. Math. **123** (2000), pp. 261–292.

[24] R.D. Fierro, J.R. Bunch: Collinearity and total least squares. SIAM J. Matrix Anal. Appl. **15** (1994), pp. 1167–1181.

[25] R.D. Fierro, G.H. Golub, P.Ch. Hansen, D.P. O'Leary: Regularization by truncated total least squares. SIAM J. Scient. Comp. **18** (1997), pp. 1223–1241.

[26] R.W. Freund, N.M. Nachtigal: QMR: a Quasi-minimal residual method for non-hermitian linear systems. Numer. Math. **60** (1991), pp. 315–339.

[27] L. Giraud, J. Langou: When modified Gram-Schmidt generates a well-conditioned set of vectors. IMA J. of Num. Anal. **22** (2002), pp. 521–528.

[28] G.H. Golub, P.C. Hansen, D.P. O'Leary: Tikhonov regularization and total least squares. SIAM J. Matrix Anal. Appl. **21** (1999), pp. 185–194.

[29] G.H. Golub, M.T. Heath, G. Wahba: Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics **21** (1979), pp. 215–223.

[30] G. H. Golub, W. Kahan: Calculating the singular values and pseudo-inverse of a matrix. SIAM J. Numer. Anal. Ser. B **2** (1965), pp. 205–224.

[31] G.H. Golub, Ch.F. Van Loan: *Matrix Computations.* Baltimore, The John Hopkins University Press 1984.

[32] S. Goossens, D. Roose: Ritz and harmonic Ritz values and convergence of FOM and GMRES. Numer. Lin. Algebra with Appl. **6** (1999), pp. 281–293.

[33] A. Greenbaum, M. Rozložník, Z. Strakoš: Numerical behaviour of the modified Gram-Schmidt GMRES implementation. BIT, **37** (1997), pp. 706–719.

[34] M. Hanke: On Lanczos based methods for the regularization of discrete ill-posed problems. BIT Numer. Math. **41** (2001), pp. 1008–1018.

[35] P.C. Hansen: *Rank-Deficient and Discrete Ill-Posed Problems – Numerical Aspects of Linear Inversion.* Philadelphia, SIAM 1997.

[36] P.C. Hansen: The L-curve and its use in the numerical treatment of inverse problems. Advances in Comp. Bioen. **4** (2000), pp. 119–142.

[37] P.C. Hansen: Regulatization Tools, a Matlab package for analysis of discrete regularization problems. Numerical Algorithms **6** (1994), pp. 1–35.

[38] P.C. Hansen: Analysis of discrete ill-posed problems by means of the L-curve. SIAM Rev. **34** (1992), pp. 561–580.

[39] P.C. Hansen, D.P. O'Leary: Regularization algorithms based on total least squares. Recent Advances in Total Least Squares Techniques and Errors-in Variables Modeling, SIAM, Philadelphia (1996), pp. 127–137.

[40] Harwell-Boeing sparse matrix collection. Matrix Market, http://math.nist.gov/MatrixMarket/index.html.

[41] P. Henrici: *Elements of Numerical Analysis.* New York, J. Wiley and Sons 1964.

[42] I. Hnětynková, Z. Strakoš: Lanczos tridiagonalization and core problems. To appear in Lin. Alg. Appl.

[43] A.S. Householder: *The Theory of Matrices in Numerical Analysis.* New York, Blaisdell Publishing Co. 1964.

[44] S. Van Huffel, J. Vandewalle: *The Total Least Squares Problem – Computational Aspects and Analysis.* Philadelphia, SIAM 1991.

[45] I.C.F. Ipsen: Expressions and bounds for the GMRES residual. BIT **40** (2000), pp. 524–535.

[46] E.-X. Jiang: Perturbation in eigenvalues of a symmetric tridiagonal matrix. Lin. Alg. Appl. **399** (2005), pp. 91–107.

[47] C. Lanczos: *Linear Differential Operators.* London, Van Nostrand 1961.

[48] C. Lanczos: An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. J. Res. Nat. Bur. Standards **45** (1950), pp. 255–282.

[49] R.B. Lehoucq: *Analysis and Implementation of an Implicitly Restarted Arnoldi Iteration.* PhD. thesis, Houston 1995.

[50] J. Liesen, Z. Strakoš: GMRES convergence analysis for a convection-diffusion model problem. SIAM J. Scient. Comp. **26** (2005), pp. 1989–2009.

[51] J. Liesen, Z. Strakoš: Convergence of GMRES for tridiagonal Toeplitz matrices. SIAM J. Matrix Anal. Appl. **26** (2004), pp. 233–251.

[52] T.A. Manteuffel: Adaptive procedure for estimating parameters for the nonsymmetric Tchebychev iteration. Numer. Math. **31** (1978), pp. 183–208.

[53] T.A. Manteuffel: The Tchebychev iteration for nonsymmetric linear systems. Numer. Math. **28** (1977), pp. 307–327.

[54] L. Mirsky: Symmetric gauge functions and unitary invariant norms. Quart. J. Math. Oxford **11** (1960), pp. 50–59.

[55] R.B. Morgan: GMRES with deflated restarting. SIAM J. Scient. Comp. **24** (2002), pp. 20–37.

[56] R.B. Morgan: Implicitly restarted GMRES and Arnoldi methods for nonsymmetric systems of equations. SIAM J. Matrix Anal. Appl. **21** (2000), pp. 1112–1135.

[57] R.B. Morgan: On restarting the Arnoldi method for large nonsymmetric eigenvalue problems. Math. Comp. **65** (1996), pp. 1213–1230.

[58] R.B. Morgan: A restarted GMRES method augmented with eigenvectors. SIAM J. Matrix Anal. Appl. **16** (1995), pp. 1154–1171.

[59] R.B. Morgan, M. Zeng: Harmonic projection methods for large nonsymmetric eigenvalue problems. Numer. Lin. Algebra with Appl. **5** (1998), pp. 33–55.

[60] D.P. O'Leary, J.A. Simmons: A bidiagonalization-regularization procedure for large scale discretizations of ill-posed problems. SIAM J. Scient. Statist. Comp. **2** (1981), pp. 474–489.

[61] C.C. Paige: Bidiagonalization of matrices and solution of linear equations. SIAM J. Numer. Anal. **11** (1974), pp. 197–209.

[62] C.C. Paige, M.A. Saunders: LSQR: an algorithm for sparse linear equations and sparse least squares. ACM Trans. Math. Software **8** (1982), pp. 43–71.

[63] C.C. Paige, M.A. Saunders: Algorithm 583: sparse linear equations and least squares problems. ACM Trans. Math. Software **8** (1982), pp. 195–209.

[64] C.C. Paige, Z. Strakoš: Core problems in linear algebraic systems. SIAM J. Matrix Anal. Appl. **27** (2006), pp. 861–875.

[65] C.C. Paige, Z. Strakoš: Scaled total least squares fundamentals. Numer. Math. **91** (2002), pp. 117–146.

[66] C.C. Paige, Z. Strakoš: Residual and backward error bounds in minimum residual Krylov subspace methods. SIAM J. Scient. Comp. **23** (2002), pp. 1899–1924.

[67] C.C. Paige, Z. Strakoš: Unifying least squares, total least squares and data least squares. In "Total Least Squares and Errors-in-Variables Modeling", S. van Huffel and P. Lemmerling, Dordrecht, Kluwer Academic Publishers (2002), pp. 25–34.

[68] B.N. Parlett: *The Symmetric Eigenvalue Problem.* Philadelphia, SIAM second ed. 1998.

[69] A. Ralston: *A First Course in Numerical Analysis.* New York, McGraw-Hill Book Company 1965.

[70] A. Ruhe: The rational Krylov algorithm for nonsymmetric eigenvalue problems III: complex shifts for real matrices. BIT, **34** (1994), pp. 165–176.

[71] R.A. Renaut, H. Guo: Efficient algorithms for solution of regularized total least squares. SIAM J. Matrix Anal. Appl. **26** (2005), pp. 457–476.

[72] T.J. Rivlin: *The Tchebychev polynomials: from approximation theory to algebra and number theory.* New York, J. Wiley and Sons 1990.

[73] Y. Saad: *Iterative Methods for Sparse Linear Systems.* Philadelphia, SIAM 2000.

[74] Y. Saad: *Numerical Methods for Large Eigenvalue Problems.* Manchester, Manchester University Press 1992.

[75] Y. Saad, M. Schultz: A generalized minimal residual algorithm for solving nonsymmetric linear systems. SIAM J. Scientific and Stat. Comp. **7** (1986), pp. 856–869.

[76] M. Sauren, H.M. Bucker: On deriving the quasi-minimal residual method. SIAM Rev. **40** (1998), pp. 922–926.

[77] D. Sima: *Regularization techniques in model fitting and parameter estimation.* PhD. thesis, Faculty of Engineering, K.U. Leuven 2006.

[78] D. Sima, S. Van Huffel: Using core formulations for ill-posed linear systems. PAMM **5** (2005), pp. 795–796.

[79] H.D. Simon, H. Zha: Low rank matrix approximation using the Lanczos bidiagonalization process with applications. SIAM J. Scient. Comp. **21** (2000), pp. 2257–2274.

[80] V. Simoncini: A new variant of restarted GMRES. Numer. Lin. Algebra with Appl. **6** (1999), pp. 61–77.

[81] D.C. Sorensen: Implicit application of polynomial filters in a K-step Arnoldi method. SIAM J. Matrix Anal. Appl. **13** (1992), pp. 357–385.

[82] G.W. Steward, Ji-guang Sun: *Matrix Perturbation Theory.* London, Academic Press 1990.

[83] G.W. Stewart: *Matrix algorithms.* Philadelphia, SIAM 1998.

[84] E. Sturler: Truncation strategies for optimal Krylov subspace methods. SIAM J. Numer. Anal. **36** (1999), pp. 864–889.

[85] P. Tichý, J. Liesen: Worst-case and ideal GMRES for a Jordan block. Tech. Rep. 19-2005, Institut für Mathematik, TU Berlin, Germany 2005.

[86] L.N. Trefethen, D. Bau, III: *Numerical Linear Algebra.* Philadelphia, SIAM 1997.

[87] I. Ulrychová: *Acceleration of convergence of the GMRES method.* Diploma thesis, Faculty of Math. and Phys., Charles University, Prague 2003.

[88] R.S. Varga: *Matrix Iterative Analysis.* Berlin, Springer 2000.

[89] H.A. van der Vorst: *Iterative Krylov Methods for Large Linear Systems.* Cambridge, Cambridge Monographs on Appl. Comp. Math. 2003.

[90] M.F. Walker: Implementation of the GMRES method using Householder transformations. SIAM J. Scient. Comp. **9** (1988), pp. 152–163.

[91] M. Wei: The analysis for the total least squares problem with more than one solution. SIAM J. Matrix Anal. Appl. **13** (1992), pp. 746–763.

[92] M. Wei: Algebraic relations between the total least squares and least squares problems with more than one solution. Numer. Math. **62** (1992), pp. 123–148.

[93] J.H. Wilkinson: *The algebraic eigenvalue problem.* Oxford, Clarendon Press 1965.

[94] J. Zítko: Convergence conditions for a restarted GMRES method augmented with eigenspaces. Numer. Lin. Algebra with Appl. **12** (2005), pp. 373–390.

[95] J. Zítko: Generalization of convergence conditions for a restarted GMRES. Numer. Lin. Algebra with Appl. **7** (2000), pp. 117–131.