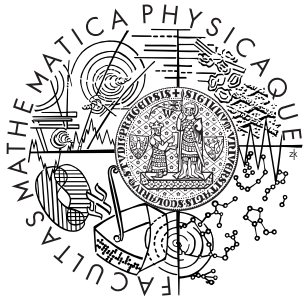
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Otto Strnad

Theoretical questions in behaviour of Krylov subspace methods

Katedra numerické matematiky

Vedoucí diplomové práce: prof. Ing. Zdeněk Strakoš, DrSc.

Studijní program: Matematika, obor matematické modelování ve fyzice a technice

2010

Rád bych poděkoval vedoucímu své diplomové práce Zdeňku Strakošovi za jeho čas a trpělivost, Gérardu Meurantovi za jeho pomoc s numerickými výpočty a poskytnutí mnoha nepublikovaných výsledků. Dále bych rád poděkoval Petru Tichému, Ctiradu Matonohovi a Ladislavu Lukšanovi za mnoho užitečných rad a pomoc s numerickými výpočty a paní Haně Bílkové za pomoc s grafickým zpracováním práce.

Prohlašuji, že jsem svou diplomovou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V Praze dne 9. 12. 2010 Otto Strnad

Contents

Název práce: Teoretické otázky popisu chování Krylovovských metod
Autor: Otto Strnad
Katedra: Katedra numerické matematiky
Vedoucí diplomové práce: prof. Ing. Zdeněk Strakoš, DrSc.
e-mail vedoucího: strakos@cs.cas.cz

Abstrakt: Předkládaná diplomová práce se zabývá analýzou konvergence metody GMRES. Vysvětluje základní principy metod CG, MINRES a GMRES. Práce shrnuje některé známé konvergenční výsledky týkající se těchto metod. Shrnuje také známé charakterizace matic a pravých stran generujících shodné Krylovovské reziduální prostory. Jsou ukázány souvislosti a rozdly mezi různými úhly pohledu na analýzu rychlosti konvergence metody GMRES. Předpokládáme, že pokud se konvergenční křivka metody GMRES aplikované na matici A, jež není normální, a pravou stranu b chová, jako by byla určena vlastními čísly matice A, potom existuje téměř normální matice, jež má shodné spektrum, jako matice A a pro pravou stranu b, shodnou GMRES konvergenční křivku, jako matice A (Předpokládáme, že počáteční aproximace $x_0 = 0$). K prozkoumání tohoto předpokladu je provedeno několik numerickch experimentů. Předkládaná práce popisuje nepublikovaný výsledek Gérarda Meuranta, vzorec pro normu k-té chyby metody GMRES aplikované na matici A a pravou stranu b a odvození tohoto vzorce. Dále je odvozen horní odhad k-té chyby GMRES. Tento odhad je minimalizován přes spektrum.

Klíčová slova: GMRES, analýza konvergence, Krylovovský podprostor

Title: Theoretical questions in behaviour of Krylov subspace methods
Author: Otto Strnad
Department: Department of numerical mathematics
Supervisor: prof. Ing. Zdeněk Strakoš, DrSc.
Supervisor's e-mail address: strakos@cs.cas.cz

Abstract: The presented thesis is focused on the GMRES convergence analysis. The basic principles of CG, MINRES and GMRES are briefly explained. The thesis summarizes some known convergence results of these methods. The known characterizations of the matrices and the right hand sides generating the same Krylov residual spaces are summarized. Connections and the differences between the different points of view on GMRES convergence analysis are shown. We expect that if the convergence curve of GMRES applied to the nonnormal matrix A and the right hand side b seems to be determined by the eigenvalues of the matrix A then exists a matrix that is close to normal and has the same spectrum as the matrix A and for the right hand side b has the same GMRES convergence curve (We assume that the initial approximation $x_0 = 0$). Several numerical experiments are done to examine this assumption. This thesis describes an unpublished result of Gérard Meurant which is the formula for the norm of the k-th error of GMRES applied to the matrix A and right hand side b and its derivation. The upper estimate of the k-th GMRES error is derived. This estimate is minimized via spectrum.

Keywords: GMRES, convergence analysis, Krylov subspace

Notation

$\mathbb{C}^N$	linear vector space of N dimensional complex vectors
$\mathbb{F}^N$	N dimensional linear vector space of real or complex vectors
$\mathbb{F}^{N\times N}$	linear vector space of N times N real or complex matrices
$K_k(A,b)$	k-th Krylov subspace determined by the matrix A and the vector b
$AK_k(A,b)$	k-th Krylov residual subspace determined by the matrix A and the vector b
x_k	k-th approximation of the solution
$r_k = b - Ax_k$	k-th residual
$\epsilon_k = x - x_k$	k-the error
$\mathcal{S}_k$	k-th search space
$\mathcal{C}_k$	k-th constraint space
$\mathcal{F}(A)$	field of values of the matrix A
P_k	linear vector space of all polynomials of degree at most k having the value one at origin
HPD	Hermitian positive definite matrix
$\langle .,. \rangle$	scalar product
$\langle .,. \rangle_A$	A-scalar product
$\|.\|_A$	A-norm
$\mathrm{diag}\{x_j\}$	diagonal matrix with the entry x_j on the position $(.)_{j,j}$
$\kappa(A)$	condition number of the matrix A
$\sigma(A)$	spectrum of the matrix A
$\mathcal{L}(\Gamma)$	length of curve or curves Γ
$\mathrm{Co}(\sigma(A))$	convex hull of $\sigma(A)$
$(A)_{\bullet i}$	i-th column of matrix A
$\mathrm{Ker}(A)$	nullspace of the operator A

Chapter 1

Introduction

Krylov subspace methods are very popular for solving large sparse systems of linear equations, because of their low memory requirements and good approximation properties. Such systems

$$Ax = b, \quad A \in \mathbb{F}^{N\times N}, \quad b \in \mathbb{F}^N, \tag{1.1}$$

where $\mathbb{F}$ denotes real or complex numbers, often arise from discretization of partial differential equations. The use of Krylov subspaces in iterative methods for linear systems is even counted among the "Top 10" algorithmic ideas of the 20th century [8].

In Krylov subspace methods the system (1.1) is projected onto a sequence of suitable nested subspaces of much smaller dimensionality than the original problem. These projections give a sequence of the approximate solutions of the problem. Krylov subspace methods build up nested subspaces using increasing powers of the system matrix with respect to the given starting vector. We denote the k-th Krylov subspace determined by the matrix A and the vector b as $K_k(A, b)$ (or just K_k, if A and b are obvious)

$$K_k(A, b) = \text{span}\{b, Ab, A^2b, ..., A^{k-1}b\}. \tag{1.2}$$

If the original problem contains some dominant information then the powers transfer this information quickly into the projected problem. On the other hand, if the problem does not contain any dominant information then the projection of the problem on a Krylov subspace of small dimensionality approximates the original problem poorly. Krylov subspace methods are naturally highly nonlinear, even though the solved problem (1.1) is linear, because the powers represent the nonlinearity. We would like to mention that dominance has a negative effect on the loss of information due to performing the computations inaccurately. In other words, the effect of rounding errors can be significant, so numerical stability analysis is very important.

Krylov subspace methods in exact arithmetic find an exact solution in a finite number of steps, or break down in some step l, $l \leq \dim A$. Breaking down means that an iteration x_l does not exist or it is not unique. A convergence of these methods is meant differently from the standard asymptotic meaning. From the computational point of view this difference is not important, because we want to find a sufficiently accurate approximate solution much before the N-th step. For Krylov subspace methods the rate of convergence means the speed of achieving a

sufficiently accurate solution, which is a slight abuse of the standard terminology. In other words: "generally speaking, fast convergence of Krylov subspace methods means that Krylov subspace of small dimensionality (significantly smaller than the size of the problem) contain enough information about the solution", see [57]. Event hough the finite character of Krylov subspace methods is unimportant from computational point of view, the theoretical finite termination property affects the practical behavior of the iterates from the very first iteration. At any step Krylov subspace methods use the information of all previous iterations and anticipate the behavior in future iterations.

The purpose of chapters 2 and 3 is to describe the basic principle of the most popular Krylov subspace methods and briefly summarize known convergence results. In most of these chapters we have followed the exposition and used some results as expressed in [45]. We have also used [44]. Many useful pieces of information can also be found in [57], [55] and [48].

Chapter 4 follows the exposition of the closely related papers [35], [34] and [3] and describes the conditions, when different matrices have the same convergence curve or the same spectrum and the convergence curve.

Chapters 5, 6 and 7 discuss a few theoretical questions about convergence of Krylov subspace methods and describe the results of our numerical experiments.

Chapter 2

Krylov subspace methods

Consider a linear algebraic system of the form (1.1) and suppose the k-th iterate to be of the form

$$x_k \in x_0 + \mathcal{S}_k, \tag{2.1}$$

where $\mathcal{S}_k$ is a k-dimensional space called search space. For Krylov subspace methods $\mathcal{S}_k = K_k$ and x_0 is an initial guess, so we can write

$$x_k \in x_0 + K_k. \tag{2.2}$$

It is obvious that there are k degrees of freedom, so we need k constraints to determine the unique approximate solution. We can do that by choosing k-dimensional space $\mathcal{C}_k$ called constraint space and demanding that k-th residual r_k, which is defined like

$$r_k \equiv b - Ax_k \tag{2.3}$$

is orthogonal to this space.

$$r_k \in r_0 + A\mathcal{S}_k, \quad r_k \perp \mathcal{C}_k, \tag{2.4}$$

where $r_0 \equiv b - Ax_0$ is an initial residual. It's easy to see that

$$r_k \in r_0 + AK_k, \tag{2.5}$$

where AK_k is for the lack of a better name called Krylov residual space.

Individual methods differ in choice of constraint space. There are two most important choices of $\mathcal{C}_k$ [54][1].

$$\begin{array}{ll} \mathcal{C}_k = \mathcal{S}_k & \text{(Galerkin method)} \\ \mathcal{C}_k = A\mathcal{S}_k & \text{(Minimal residual method)} \end{array} \tag{2.6}$$

As we mentioned above, for Krylov subspace methods $\mathcal{S}_k = K_k$. In this case the Galerkin method generates orthogonal residuals $r_k = b - Ax_k$ and in this context

[1]The exact definition of Krylov subspace methods is a matter of taste. It is possible to consider the search space not to be $K_k(A, r_0)$, but some linear space related to K_k. For example $\mathcal{S}_k = AK_k(A, r_0)$ or $\mathcal{S}_k = K_k(A^*, r_0)$. Furthermore, one may consider Krylov subspace methods as such methods, that they search their iterates in Krylov subspaces, but these methods are not defined by projection property.

this method is also called orthogonal residual (OR) method. OR Krylov subspace method computes uniquely defined solution, if zero is outside the field of values of matrix A, which is defined as

$$\mathcal{F}(A) = \{v^*Av : \|v\| = 1, v \in C^N\}. \tag{2.7}$$

The proof of this statement is well known. We show the idea of the proof at the end of the section 2.3.

In this thesis we will consider conjugate gradients method (CG) which represents Galerkin method applied to Hermitian positive definite (HPD) matrices. The uniqueness of the solution is satisfied automatically. Minimal residual methods are well defined if the matrix A is regular. The most popular implementations are MINRES (minimal residual method) for Hermitian matrices and GMRES (generalized minimal residual method) for general matrices.

Let p_k by element of P_k, the vector space of all the polynomials of degree at most k and with value one at the origin, the conditions (2.2) and (2.3) imply that the error $\epsilon_k \equiv x - x_k$ and the residual r_k can be written in the polynomial form

$$x - x_k = p_k(A)(x - x_k), \tag{2.8}$$

$$r_k = p_k(A)r_0. \tag{2.9}$$

This formulation of the k-th error and the residual is widely used for the convergence analysis of Krylov subspace methods. Next section briefly describes Arnoldi's algorithm, which is the most important procedure for computing the orthogonal basis of the Krylov subspace. Another two sections describe the principle of the most popular Galerkin and minimal residual methods.

2.1 Arnoldi's algorithm

Arnoldi's algoritm computes in the k-th step the k-dimensional orthonormal basis of Krylov subspace $K_k(A, b)$. The k-th vector v_k is computed as an orthogonalisation of Av_{k-1} against the first $(k-1)$ vectors. From one point of view, the Arnoldi's algorithm is a variant of the Gram-Schmidt orthogonalisation method applied to the Krylov sequence in order to computer an orthonormal basis of the Krylov subspace. For detail description of Arnoldi algorithm, see [54, section 6.3]. The result of this process is so-called Arnoldi's decomposition

$$AV_k = V_k H_k + \xi_{k+1} v_{k+1} e_k^T \quad \xi_{k+1} \in \mathbb{R} \tag{2.10}$$

or

$$AV_k = V_{k+1} H_{k+1,k}, \tag{2.11}$$

where V_k is a matrix with columns $(v_1, \ldots, v_k)$, H_k is an upper Hessenberg matrix which is defined as a matrix with zeros under first subdiagonal, e_k is the k-th column of the k-dimensional identity matrix and $H_{k+1,k}$ is H_k with added $(0, \ldots, \xi_{k+1})$ as the last row. After N steps is Arnoldi's decomposition of the form

$$AV_N = H_N V_N. \tag{2.12}$$

The relation (2.10) implies that

$$h_{k+1,k}v_{k+1} = Av_k - h_{1,k}v_1 - \ldots - h_{k,k}v_k \tag{2.13}$$

If the matrix A is Hermitian, the upper Hessenberg matrix $H = V_N^* A V_N$ must also be Hermitian and consequently tridiagonal. Therefore the relation (2.13) is reduced to the form

$$h_{k+1,k}v_{k+1} = Av_k - h_{k-1,k}v_{k-1} - h_{k,k}v_k. \tag{2.14}$$

The recurrence (2.14) shows, that for A Hermitian it is nor necessary to orthogonalise the vector Av_k against all the vectors $v_1, \ldots, v_k$ to get global orthogonality of $v_1, \ldots, v_{k+1}$. It is sufficient to orthogonalise the vector Av_k against the two most recent vectors v_{k-1} and v_k. Arnolid's algorithm applied to the Hermitian matrix is called Lanczos algorithm.

2.2 Conjugate gradients

Even thought this thesis is mostly about GMRES and its convergence analysis, CG is probably the most important Krylov subspace method and it is desirable to explain, how this method works, see [38]. Conjugate gradient method is the most common and powerful method for solving the linear problem $Ax = b$ with Hermitian positive definite matrix. There are more possible approaches of derivation of CG (see [54, section 6.7]). One of these approaches is described bellow.

2.2.1 Derivation

From one point of view CG is a method for minimizing the quadratic problem

$$\min_{x \in R^n} J(x), \tag{2.15}$$

where

$$J(x) = \frac{1}{2}x^* A x - b^* x. \tag{2.16}$$

and A is a Hermitian positive definite matrix. Each HPD matrix defines a so-called an A-scalar product

$$\langle x, y \rangle_A = x^* A y \tag{2.17}$$

and an A-norm

$$\|x\|_A = \langle x, x \rangle_A^{\frac{1}{2}}. \tag{2.18}$$

The key point in the derivation is a clever use of these basic tools. Let y be some approximation to the solution of the linear problem $Ax = b$ and let's define an error $\epsilon = x - y$. We can easily derive

$$J(y) = \frac{1}{2}\|\epsilon\|_A^2 - \frac{1}{2}\|x\|_A^2. \tag{2.19}$$

We can see, that $J(y)$ is minimal, if and only if $\epsilon = 0$ i. e. $Ay = b$.

As stated in [45], the A-norm of the error often has a counterpart in the underlying real-world problem. For example, when the linear system comes from finite element approximations of self-adjoint elliptic PDEs, then the A-norm of the error can be interpreted as the discretized measure of energy which is to be minimized. That's the reason why we often call the A-norm as an energy-norm [1],[2].

Consider the k-th approximation x_k and minimization of $J(y)$ along the line $y = x_k + \gamma p_k$, i. e., in the direction p_k. New approximation x_{k+1} we compute as a minimum of the A-norm of an error in this direction. It means that

$$x_{k+1} = x_k + \gamma_k p_k \tag{2.20}$$

$$\gamma_k = \frac{\langle r_k, p_k \rangle}{\langle p_k, p_k \rangle_A}, \tag{2.21}$$

where r_k is k-th residual vector $r_k = b - Ax_k$. Now we choose p_0 as r_0 and

$$p_k = r_k + \omega_k p_{k-1}, \tag{2.22}$$

where

$$\omega_k = -\frac{p_{k-1}^* A r_k}{p_{k-1}^* A p_{k-1}} \tag{2.23}$$

This choice of ω_k ensure that $p_k \perp_A p_{k-1}$.

The surprising, but well known result is that this local A-orthogonality is a sufficient condition for global A-orthogonality of all vectors p_k.

Theorem 2.1 *Let the $(k+1)$-th approximation of CG be defined by the relations (2.20), (2.21), (2.22), (2.23) and let $p_0 \equiv r_0$. Then*

$$r_{k+1}^* p_k = 0, \quad k = 1, 2, \ldots \tag{2.24}$$

$$r_k^* r_j = 0, \quad k \neq j \tag{2.25}$$

$$p_k^* A p_j = 0, \quad k \neq j. \tag{2.26}$$

Proof:

$$x_{k+1} = x_k + \gamma_k p_k \Rightarrow r_{k+1} = r_k - \gamma_k A p_k$$
$$\langle r_{k+1}, p_k \rangle = \langle r_k - \gamma_k A p_k, p_k \rangle = \langle r_k, p_k \rangle - \gamma \langle A p_k, p_k \rangle = 0$$

This proves that $r_{k+1} \perp p_k$.
Now we can see that

$$r_k^* r_{k-1} = 0, \quad k = 1, 2, \ldots, \tag{2.27}$$

because

$$r_k^* r_{k-1} = r_k^* (p_{k-1} - \omega_{k-1} p_{k-2}) = -\omega_{k-1} r_k^* p_{k-2} = -\omega_{k-1} (r_{k-1} - \gamma_{k-1} A p_{k-1})^* p_{k-2} = 0 \tag{2.28}$$

The rest is by induction. Holds

$$\begin{gathered} r_1^* r_0 = (r_0 - \gamma_0 A p_0)^* r_0 = \|r_0\|^2 - \gamma_0 r_0^* A r_0 = 0, \\ r_2^* r_0 = (r_1 - \gamma_1 A p_1)^* r_0 = -\gamma p_1^* A r_0 = -\gamma_1 r_0^* A p_1 = -\gamma_1 p_0^* A p_1 = 0, \\ p_1^* A r_0 = p_1^* A p_0 = 0, \\ p_2^* A p_0 = (r_2 + \omega_2 p_1)^* A p_0 = r_2^* A p_0 = r_2^* \left(r_0 - r_1/\gamma_0\right) = 0 \\ p_2^* A r_1 = p_2^* A (p_1 - \omega_0 p_0) = 0. \end{gathered}$$

With induction assumptions

$$r_k^* r_i = 0, \quad p_k^* A p_i = 0, \quad p_k^* A r_i = 0, \quad i < k,$$

we prove the rest:

$$\begin{gathered} r_{k+1}^* r_i = 0, \quad i = k, \\ r_{k+1}^* r_i = (r_k - \gamma_k A p_k)^* r_i = 0, \quad i < k \\ p_{k+1}^* A p_i = 0, \quad i = k, \\ p_{k+1}^* A p_i = (r_{k+1} + \omega_{k+1} p_k)^* A p_i = r_{k+1}^* A p_i = r_{k+1}^* \left(r_i - r_{i+1}/\gamma_i\right) = 0, \quad i < k, \\ p_{k+1}^* A r_i = p_{k+1}^* A (p_i - \omega_{i-1} p_{i-1}) = 0, \quad i \leq k. \end{gathered}$$

□

Let $\epsilon_{k+1} = x - x_{k+1}$ be the $(k+1)$-th error. This error is A-orthogonal to p_k, because of the relation (2.24) and $A\epsilon_{k+1} = r_{k+1}$. Observe that

$$\epsilon_{k+1} = \epsilon_k - \gamma_k p_k \tag{2.29}$$

so $(k+1)$-th error can be interpreted as an A-orthogonalization of ϵ_k on $\mathrm{span}\{p_k\}$ and $\epsilon_k = \epsilon_{k+1} + \gamma_k p_k$ is orthogonal decomposition of ϵ_k with respect to the A-scalar product. As we know $p_0, \ldots, p_k$ are globally A-orthogonal, it follows

$$\epsilon_{k+1} = \epsilon_0 - \sum_{l=0}^{k} \gamma_l p_l \tag{2.30}$$

is an A-orthogonalization of e_0. This means

$$\|x - x_{k+1}\|_A = \min_{y \in x_0 + \mathrm{span}\{p_0, \ldots, p_k\}} \|x - y\|_A, \tag{2.31}$$

and because of the global A-orthogonality p_n=0 necessarily, so the method ends at least in the n-th step.

The relation (2.31) shows very important fact, that because of a very smart way of choosing the directions p_k, CG minimizes the A-norm of the error in each step k over linear space of the dimension k.

Furthermore, after j steps of CG (in case $r_k \neq 0$ in every step) holds

$$\mathrm{span}\{p_0, \ldots, p_j\} = \mathrm{span}\{r_0, \ldots, r_j\} = K_{j+1}(A, r_0). \tag{2.32}$$

The relation (2.32) follows from the relations above. For detailed proof see e. g. [62, Theorem 7.7.14]. The relation (2.32) shows that CG computes simultaneously A-orthogonal basis $\{p_0, \ldots, p_j\}$ and euclidian orthogonal basis $\{r_0, \ldots, r_j\}$ of $K_{j+1}(A, r_0)$.

2.3 Minimal residual methods (MINRES, GMRES)

The most popular implementations of minimal residual methods, defined by (2.1) and (2.6) are MINRES (Minimal residual method) for Hermitian matrices and GMRES (general residual method) for general matrices. Mathematical principle of MINRES is the very same as the principle of GMRES, but MINRES was developted earlier and independently from GMRES [2], so it is convenient to consider MINRES and GMRES as two different methods.

Mathematical characterization of MINRES and GMRES by optimality property is

$$\|r_k\| = \min_{z \in x_0 + \mathcal{K}_k(A, r_0)} \|b - Az\|. \tag{2.33}$$

Using (2.9)

$$\|r_k\| = \min_{p \in P_k} \|p(A) r_0\|. \tag{2.34}$$

GMRES usually computes by modified Gram-Schmidt process Arnoldi decomposition

$$AV_k = V_{k+1} H_{k+1,k}, \tag{2.35}$$

where $v_1 = r_0/\|r_0\|$ and minimizes a norm of k-th residual $\|b - Ax_k\|$. Approximations are of the form $x_k = x_0 + V_k x_k$, where y is the solution of minimization problem

$$\begin{aligned} \|r_k\| &= \min_y \|b - Ax_0 - AV_k y\| = \min_y \|\|r_0\| v_1 - V_{k+1} H_{k+1,k} y\| = \\ &= \min_y \|V_{k+1} \|r_0\| e_1 - V_{k+1} H_{k+1,k} y\| = \min_y \|\|r_0\| e_1 - H_{k+1,k} y\|. \end{aligned}$$

If A is Hermitian matrix, then we use instead of Arnoldi algorithm Lanczos algortihm, using short recurences that generates instead of upper Hessenberg matrix, tridiagonal matrix $T_{k+1,k}$. The k-th approximation of MINRES is determined by the solution of the problem

$$y_k = \operatorname{argmin}\{\|r_k\|\} = \operatorname{argmin}\{\|\|r_0\| e_1 - T_{k+1,k} y\|\}.$$

It is well known that if zero is outside the field of values, then OR method computes uniquely defined solution. We would like to show the idea of the proof of this statement. The proof has two steps.

The first step is to show that

$$0 \notin \mathcal{F}(A) \Rightarrow \|r_{k-1}\| > \|r_k\|, \tag{2.36}$$

where r_k is k-th residual of GMRES $r_k \perp AK_k(A, r_0)$. If $\|r_{k-1}\| = \|r_k\|$, then $r_{k-1} = r_k$ and consequently $r_{k-1} \perp AK_k(A, r_0)$ as well as $r_{k-1} \perp AK_{k-1}(A, r_0)$, i. e. $r_{k-1}^* A r_{k-1} = 0$. This implies that zero is an element of $\mathcal{F}(A)$. We have proved

$$\|r_{k-1}\| = \|r_k\| \Rightarrow 0 \in \mathcal{F}(A)$$

[2] MINRES was developted by Ch. Paige and M. Saunders in 1975, see [52].

thus (2.36) holds.

The second step is a bit more complicated. The OR method applied to the nonsymmetric matrix is called Full orthogonal method (FOM). Let H_k be the matrix obtained from $H_{k+1,k}$ (2.35) by deleting its last row. The k-th approximation of FOM is not defined if and only if H_k is singular and H_k is singular if and only if $\|r_{k-1}\| = \|r_k\|$ (see e. g. [54, page 181]). The proof is now complete.

2.4 Restarted GMRES

As number of the GMRES iterations grows, the memory and computational requirements grow as well. One possible solution of this inconvenience is based on restarting GMRES. We borrowed following basic restarted GMRES algorithm from [54]

1. Compute $r_0 = b - Ax_0$, $\beta = \|r_0\|$ and $v_1 = r_0/\beta$
2. Generate the Arnoldi basis and the matrix $H_{k_1,k}$ using the Arnoldi algorithm starting with v_1
3. Compute y_k witch minimizes $\|\beta e_1 - H_{k+1,k}y\|$ and $x_k = x_0 + V_k y_k$
4. If satisfied then Stop, else set $x_0 := x_m$ and GoTo 1

For more informations about restarted GMRES and the convergence of restarted GMRES see [50], [5], [67], [66], [65] and [64].

Chapter 3

Convergence analysis of Krylov subspace methods

In exact arithmetics, any well defined Krylov subspace method, finds solution in finite number of steps, thus the "rate of convergence" loses it's classical meaning. The goal of the convergence analysis is to describe the convergence rate, mainly the early stage of the process, in terms of input data, i. e. system matrix, right hand side and initial guess, because it is typically required, to find an acceptable approximate solution x_n in $n \ll N$ steps.

Convergence analysis for normal matrices is connected with their eigenvalues and quite well understood. A is normal if and only if the A is unitarily diagonalizable, which means, that there exists unitary transformation between A and some diagonal matrix.

$$A = Q\Lambda Q^*, \quad Q^*Q = I, \quad \Lambda = \mathrm{diag}\{\lambda_j\} \tag{3.1}$$

In other words $A \in \mathbb{F}^{N\times N}$ is normal if and only if its normalized eigenvectors form orthonormal basis of $\mathbb{F}^N$. This fact dramatically simplifies the problem. Formula (3.1) is spectral decomposition of matrix A, $\lambda_1, \ldots, \lambda_N$ are the eigenvalues of matrix A.

Convergence analysis for nonnormal matrices is highly nonlinear and consequently a very difficult and open problem.

3.1 Convergence analysis of CG

We know, that CG minimizes the A-norm of the error of Hermitian positive definite matrix. Using (2.8) we easily derive

$$\begin{aligned} \|x - x_k\|_A &= \min_{p\in P_k} \|p(A)(x - x_0)\|_A \\ \|\epsilon_k\|_A &= \min_{p\in P_k} \|p(A)\epsilon_0\|_A. \end{aligned} \tag{3.2}$$

A simple algebraic manipulation shows:

$$\|p(A)\epsilon_0\|_A^2 = \epsilon_0^* p(A)Ap(A)\epsilon_0 = (A^{\frac{1}{2}}\epsilon_0)^* p(A)^2 A^{\frac{1}{2}}\epsilon_0 \le \|\epsilon_0\|_A^2 \|p(A)\|^2, \tag{3.3}$$

where $A^{\frac{1}{2}} = Q\ \mathrm{diag}\{\sqrt{\lambda_j}\}\ Q^*$. Using (3.1), (3.2) and (3.3):

$$\frac{\|\epsilon_k\|_A}{\|\epsilon_0\|_A} \le \min_{p\in P_k} \|p(A)\| = \min_{p\in P_k} \|Qp(\Lambda)Q^*\| = \min_{p\in P_k} \|p(\Lambda)\|$$

and finally

$$\frac{\|\epsilon_k\|_A}{\|\epsilon_0\|_A} \leq \min_{p \in P_k} \max_{i=1,\ldots,n} |p(\lambda_i)|. \tag{3.4}$$

The relative A-norm of the error is bounded by min-max approximation problem on the discrete set of eigenvalues. The rate of convergence depends on eigenvalue distribution of the A strongly. There exists right hand side b or initial guess x_0, such that for each iteration step k equality holds in (3.4).

We often don't know the whole spectrum, but only the estimates of the largest and the smallest eigenvalues. Then we can replace the discrete set of eigenvalues in estimate (3.4) by continuous interval $[\lambda_1, \lambda_N]$ and use Chebyshev polynomials of the first kind to estimate this approximation with result

$$\frac{\|\epsilon_k\|_A}{\|\epsilon_0\|_A} \leq 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k, \quad \kappa = \frac{\lambda_{max}}{\lambda_{min}}, \tag{3.5}$$

see [47]. This result is the well-known upper bound based on the condition number κ of the A. We would like to stress, that the estimate (3.5) is often a large overestimate of the worst case upper bound (3.4). On the other hand, the formula (3.5) shows, that a small condition number implies fast convergence, that justifies classical goal of preconditioning, but a large condition number doesn't imply slow convergence of CG method.

Similar convergence bound can be derived for the steepest descent method which might seem closely related to CG, the difference is in the choice of directions p_k (2.20). In each step, the steepest descent method minimizes functional $J(x)$ (2.16) in the direction of $\nabla J(x)$, i. e. in the direction of the greatest rate of the decrease of the functional $J(x)$. This choice of direction is quite simple, but not that efficient as choice of CG. CG uses in every step information from previous iterations, steepest descent doesn't and that causes usually slower convergence in comparison to CG. The convergence bound for the steepest descent derived similarly like the bound (3.5) for CG is

$$\frac{\|\epsilon_k\|_A}{\|\epsilon_0\|_A} \leq \left(\frac{\kappa - 1}{\kappa + 1} \right)^k, \tag{3.6}$$

see [54, Theorem 5.9]. We would also like to mention an interesting phenomenon called superlinear convergence, that can be observed in many practical applications. When CG converges superlinearly, the rate of convergence measured by A-norm of the error increases during the iteration. There were some attempts to explain this behavior, for more information see [9], [49], [63], [60], [61], [6] and [7].

In summary, the convergence behavior of the CG method is relatively well understood, but some open problems still remain.

3.2 MINRES convergence analysis

MINRES is the method which is well defined for Hermitian nonsingular matrices. For the definition of this method see section 2.3. For any Hermitian (not

necessarily positive definite) system matrix A, we will derive again min-max approximation of the worst-case convergence behavior.

$$\|r_k\| = \|p_k(A)r_0\| = \min_{p\in P_k} \|p(A)r_0\| = \min_{p\in P_k} \|Q[p(\Lambda)Q^* r_0]\|$$

$$= \min_{p\in P_k} \|p(\Lambda)Q^* r_0\| = \min_{p\in P_k} \left\{ \sum_i |(q_i^* r_0)p(\lambda_i)|^2 \right\}^{\frac{1}{2}} \leq \|r_0\| \min_{p\in P_k} \max_i |p(\lambda_i)|.$$

And finally

$$\frac{\|r_k\|}{\|r_0\|} \leq \min_{p\in P_k} \max_i |p(\lambda_i)|, \tag{3.7}$$

It is clear, that the bound (3.7) determines the worst-case behavior for GMRES applied to any normal matrix also. This bound is the earliest upper bound on the GMRES relative residual norm for normal matrices and is sometimes called as "standard bound". The biggest problem of estimate (3.7) is that the exact solution of the min-max problem is known only for some special cases. We can try to estimate the bound (3.7) similarly as in Hermitian positive definite case for CG, by replacing the discrete spectrum by the continuous interval including the spectrum. However, the estimation of this min-max approximation becomes much more complicated now see e. g. [45] and such estimates are often misleading.

3.3 GMRES convergence analysis

In this section, we would like to show several different approaches of analysing GMRES convergence. These methods are described in scientific literature and well known among scientific community.

3.3.1 Convergence analysis based on eigenvalues

If system matrix A is normal, but not Hermitian, we can derive for GMRES convergence bound (3.7) in the very same way as for MINRES in previous section. If the system matrix is diagonalizable we can derive a slightly different estimate, see [16].

$$\|r_k\| = \|p_k(A)r_0\| = \min_{p\in P_k} \|p(A)r_0\| = \min_{p\in P_k} \|Q[p(\Lambda)Q^{-1} r_0]\|$$

$$\leq \|Q\| \min_{p\in P_k} \|p(\Lambda)Q^* r_0\| = \|Q\| \min_{p\in P_k} \left\{ \sum_i |[Q^{-1}r_0]_i p(\lambda_i)|^2 \right\}^{\frac{1}{2}}$$

$$\leq \|Q\|\|Q^{-1}r_0\| \min_{p\in P_k} \max_i |p(\lambda_i)| \leq \|r_0\|\kappa(Q) \min_{p\in P_k} \max_i |p(\lambda_i)|,$$

where $\kappa(Q) = \|Q\|\|Q^{-1}\|$ denotes the condition number of the matrix Q. And finally

$$\frac{\|r_k\|}{\|r_0\|} = \min_{p\in P_k} \frac{\|Qp(\Lambda)Q^{-1}r_0\|}{\|r_0\|} \leq \kappa(Q) \min_{p\in P_k} \max_i |p(\lambda_i)|. \tag{3.8}$$

This bound typically provides a good estimate, when $\kappa(Q)$ small. When Q is far from unitary, (3.8) often fails to give any reasonable convergence information. When the eigenvector matrix Q is ill-conditioned, then some components of the vector $Q^{-1}r_0$ can be potentially much larger than $\|r_0\|$. Linear combination $Q[p(\Lambda)Q^{-1}r_0]$ might contain a significant cancellation, which is not reflected in the min-max problem on the right hand side of (3.8). That is the reason why this minimization problem need not have any connection with GMRES convergence curve for given diagonalizable nonnormal matrix.

This approach can be generalized to defective matrices. For detail derivation of following bound see [57].

$$\|r_j\| = \min_{p\in P_k} \|p(A)r_0\| \leq \|r_0\| N^{1/2} e\kappa(S) \min_{p\in P_k} \max_{i,l} |p^{(l)}(\lambda_i)|, \tag{3.9}$$

where e is Euler's number, $p^{(l)}$ is the l-th derivetive of polynomial p and index l goes form 0 to the size of corresponding Jordan block decreased by one.

3.3.2 Worst-case GMRES vs. ideal GMRES

The minimization (2.34) is in general case not easy to analyze, so natural approach is to approximate this problem by some less complicated problem. In practice, however, unless the right hand side b has special properties, it appears to be usually system matrix A that predominantly determines the convergence rate [58]. We eliminate the influence of the right hand side by excluding the initial residual.

$$\frac{\|r_n\|}{\|r_0\|} = \min_{p\in P_k} \frac{\|p(A)r_0\|}{\|r_0\|} \leq \max_{\|v\|=1} \min_{p\in P_k} \|p(A)v\|. \tag{3.10}$$

The last expression in (3.10) is so-called the worst-case GMRES and represents a sharp upper bound. In any step k there exists initial residual r_0 for which stands equality. The formula (3.10) can be bounded by so-called ideal GMRES introduced by Greenbaum and Trefethen [36]:

$$\frac{\|r_n\|}{\|r_0\|} \leq \max_{\|v\|=1} \min_{p\in P_k} \|p(A)v\| \leq \min_{p\in P_k} \|p(A)\|. \tag{3.11}$$

The worst-case GMRES is equal to the ideal GMRES for many classes of matrices. For example equality hold for normal matrices [42] [33] or for any matrix in the first step of GMRES [42] [33]. Greenbaum and Trefethen were led by such positive results to formulate conjecture, that the ideal GMRES convergence curve is equal to the worst case GMRES convergence curve [36], but there exist counterexamples showing, that (3.11) is not equality for any matrix A in any step. One of these examples introduced K.-Ch. Toh in [58]. For matrix A such that

$$A = \begin{pmatrix} 1 & \varepsilon & & \\ & -1 & c/\varepsilon & \\ & & 1 & \varepsilon \\ & & & -1 \end{pmatrix}, \varepsilon > 0, \quad 0 < c < 2.$$

the ration of the worst-case GMRES and ideal GMRES tends to 0 as parameter ε tends to zero in step 3.

Further we would like to generalize the bound (3.8) for normal matrices by determining the set $\Omega \subset \mathbb{C}$ that is associated with the matrix A and provide an upper bound for $\|r_n\|/\|r_0\|$ in the sense

$$\frac{\|r_n\|}{\|r_0\|} \leq c \min_{p \in P_k} \max_{z \in \Omega} |p(z)| \tag{3.12}$$

or just an upper bound for ideal GMRES in the same sense

$$\min_{p \in P_k} \|p(A)\| \leq c \min_{p \in P_k} \max_{z \in \Omega} |p(z)| \tag{3.13}$$

For symmetric matrices is Ω spectrum of A and $c = 1$. Approximations of these forms based on ε-pseudospectrum and field of values are described in the following two sections. Polynomial numerical hull also described later provides the lower bound of ideal GMRES in the form similar to (3.13).

3.3.3 Convergence analysis based on the ε-pseudospectrum

The first approximation of the ideal GMRES of the form (3.13) that we mention suggested Trefethen in [59]. He suggested to use ε-pseudospectrum of system matrix A. There are three equivalent definitions of ε-pseudospectrum.

Definition 1: The ε-pseudospectrum of A, denoted $\Lambda_\varepsilon(A)$, is the set of points $z \in \mathbb{C}$ such that $\|(zI - A)^{-1}\| \geq \varepsilon^{-1}$.

Definition 2: The ε-pseudospectrum of A is the set of points $z \in \mathbb{C}$ which are eigenvalues of some matrix $A + E$ with $\|E\| \leq \varepsilon$.

Definition 3: Given $\varepsilon > 0$ the number $\lambda \in \mathbb{C}$ is an ε-pseudoeigenvalue of A if $\exists u \in \mathbb{F}^N$ with $\|u\| = 1$ such that $\|(\lambda I - A)\| \leq \varepsilon$. [53]

Any analytic function can be written in the form

$$f(A) = \frac{1}{2\pi i} \int_\Gamma (zi - A)^{-1} f(z) dz, \tag{3.14}$$

where Γ is a simple closed curve resp. union of simple closed curves containing the spectrum of matrix A (see [11]). To estimate $\|f(A)\|$ we can replace the norm of integral by length of the curve resp. curves times the maximum norm of the integrand factors.

$$\|f(A)\| \leq \frac{1}{2\pi} \mathcal{L}(\Gamma) \max_{z \in \Gamma} \|(zI - A)^{-1}\| \max_{z \in \Gamma} |f(z)| \tag{3.15}$$

We want this inequality to be close equality. Free parameter in this estimate is the choice of the curves, so we choose the curves Γ_ε on which the resolvent $(zI - A)^{-1}$ has constant norm $1/\varepsilon$, i. e., the boundaries of the ε-pseudospectrum of A. In this case we can reasonably hope, that cancellation does not cause the norm of the integral to be much smaller then the estimate.

Taking the function f to be polynomial $p \in P_k$ for which is bound (3.15) minimal and using (3.11) the bound (3.15) becomes

$$\frac{\|r_k\|}{\|r_0\|} \leq \frac{\mathcal{L}(\Gamma_\varepsilon)}{2\pi\varepsilon} \min_{p_k \in P_k} \max_{z \in \Gamma_\varepsilon} |p_k(z)|. \tag{3.16}$$

The parameter ε gives some flexibility, but choice of good value is found to be little tricky. For large ε is $\mathcal{L}(\Gamma_\epsilon)/2\pi\varepsilon$ small, but value of min-max problem can be too large and vice versa. In some cases the bound (3.16) gives for properly chosen value of ε much better estimate than (3.8), but there are cases in which even this bound is large overestimate of relative norm of the residual. We would like to demonstrate this problem on the example introduced in [35]. Consider the matrix A of the form $Z\Lambda Z^{-1}$, where

$$Z = \begin{pmatrix} 1 & \sqrt{1-\delta} & 0 & \dots & 0 \\ 0 & \sqrt{\delta} & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ . & . & . & . & . \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}, \quad \delta \ll 1 \quad \Lambda = \begin{pmatrix} 20 & & & & \\ & 10 & & & \\ & & 5 & & \\ & & & . & \\ & & & & 1 \end{pmatrix}. \tag{3.17}$$

There are two large and well separated eigenvalues 20 and 10. These two eigenvalues correspond to an ill-conditioned block in Z. The other eigenvalues are uniformly distributed in the interval $[1, 5]$. If A is reduced to Schur form, then the Frobenius norm of the strict upper triangle of the Schur form is approximately 10^5, so we have to deal with a nonnormal matrix. Event hough the GMRES iteration will behave almost as it would if applied to the diagonal matrix Λ and converge quite quickly. Bound (3.16) predicts a very slow convergence. For the matrix $A \in \mathbb{F}^{19\times 19}$ with $\delta = 10^{-8}$, and $\varepsilon = 10^{-5}$ or 10^{-4} is the factor $\mathcal{L}(\Gamma_\epsilon)/2\pi\varepsilon$ on the order $2 \cdot 10^4$ and low degree minimizing polynomial $p_k(A)$ could not make convergence bound (3.16) reasonable. For larger $\varepsilon \gtrsim 10^{-2}$, ε-pseudospectrum contains the origin, $p_k(0) = 1$ and bound (3.16) is again a large overestimate. The condition number of the matrix Z is approximately $2/\sqrt{\delta}$, so the bound (3.8) is not useful also. This example presents diagonalizable nonnormal matrix for which convergence analysis based on spectrum or ε-pseudospectrum is not descriptive.

3.3.4 Field of values

The field of values $\mathcal{F}(A)$ (see (2.7)) is another basic tool to obtain the upper bound of the size of the relative residual norm of GMRES. The very basic properties of the field of values are:

1. The field of values is a closed set.
   Matrix A is finite dimensional operator and $\mathcal{F}(A)$ is the image of compact unit ball.

2. The field of values contains the spectrum of A.
   $Aq = \lambda q,\ \|q\| = 1 \Rightarrow q^*Aq = \lambda$

3. The field of values is a convex set
   This statement is called Toeplitz-Hausdorff theorem [41].

We would like to mention two basic approaches based on field of values.

The first approach is described in [15]. Eiermann and Ernst shows in this paper, that ideal GMRES is bounded by the distance of $\mathcal{F}(A)$ from the origin $\nu(\mathcal{F}(A)) = \min_{z\in\mathcal{F}(A)}|z|$ in the following way

$$\frac{\|r_k\|}{\|r_0\|} \leq (1 - \nu(A)\nu(A^{-1}))^{m/2}. \tag{3.18}$$

Further let A be real and positive[1], i. e. the Hermitian part of A, $M = (A+A^*)/2$ is positive definite. The field of values of M is a projection of the field of values of A to the real axis [41]. If A is normal, the field of values is a convex hull of the spectrum of A, $\mathcal{F}(A) = \mathrm{Co}(\sigma(A))$ [41]. Summarizing the facts forementioned, we get $\nu(A) = \lambda_{\min}(M) > 0$ and

$$\nu(A^{-1}) = \min_{v\in\mathbb{F}^N} \frac{\langle A^{-1}v, v\rangle}{\langle v, v\rangle} = \min_{v\in\mathbb{F}^N} \frac{\langle v, Av\rangle}{\langle v, v\rangle}\frac{\langle v, v\rangle}{\langle Av, Av\rangle} \geq \frac{\lambda_{\min}(M)}{\|A\|^2}.$$

In this special case (3.18) yields a bound first given by Elman (1982)

$$\frac{\|r_m\|}{\|r_0\|} \leq \left(1 - \frac{\lambda_{\min}(M)^2}{\lambda_{max}(A^TA)}\right)^{m/2}. \tag{3.19}$$

We require 0 to be outside the field of values, otherwise (3.18) provides only trivial estimate that $\|r_k\|/\|r_0\| \leq 1$. Since $\mathcal{F}(A)$ is convex set containing the eigenvalues of A, this requirement makes the estimate (3.18) useless for many matrices.

The second estimate [14] [13] [30] of the ideal GMRES using $\mathcal{F}(A)$ is

$$\min_{p\in P_k} \|P(A)\| \leq c_k \min_{p\in P_k} \max_{z\in\hat{\Omega}} |p(z)|, \tag{3.20}$$

where $\hat{\Omega}$ is a compact convex set containing $\mathcal{F}(A)$ and the constant c_k depends on $\hat{\Omega}$ and doesn't depend on A. This bound also provides useful information only if $0 \notin \mathcal{F}(A)$ and $\hat{\Omega}$.

Further, there exists so called Crouzeix's conjecture [10] that for any function f analytic in $\mathcal{F}(A)$

$$\|f(A)\| \leq 2 \max_{z\in\mathcal{F}(A)} |f(z)|. \tag{3.21}$$

This conjecture implies that $c_k = 2$ for any $k = 1, \ldots, N$ if $\hat{\Omega} = \mathcal{F}(A)$. The conjecture was proved for matrices $A \in \mathbb{F}^{2\times 2}$. For general N by N matrices was proved that

$$\|f(A)\| \leq 11.08 \max_{z\in\mathcal{F}(A)} |f(z)|, \tag{3.22}$$

nevertheless this bound is not necessarily sharp and there are some efforts to prove Crouzeix's conjecture for general matrices.

[1] A positive matrix is a matrix in which all the elements are greater than zero.

3.3.5 Polynomial numerical hull

The largest set providing lower bound of the form

$$\|p(A)\| \geq \max_{z\in\Omega} |p(z)|, \quad p \in \hat{P}_k, \tag{3.23}$$

where $\hat{P}_k$ is the set of all polynomials of degree k or less is the polynomial numerical hull introduced by Nevanlinna [51], and defined as

$$\mathcal{H}_k(A) = \{z \in \mathbb{C} : \|p(A)\| \geq |p(z)| \text{ for all } p \in \hat{P}_k\}. \tag{3.24}$$

The polynomial numerical hull is a generalization of the field of values in sense that $\mathcal{F}(A) = \mathcal{H}_1(A)$. Further, let m by the degree of the minimal polynomial of the matrix A. For $m \leq k$, $\mathcal{H}(A)_k$ is nothing but a spectrum of A. For $1 < k < m$ the $\mathcal{H}_k(A)$ is the intermediate between the field of values and the spectrum.

The bound provided by $\mathcal{H}_k(A)$ is

$$\min_{p\in P_k} \max_{z\in\mathcal{H}_k(A)} |p(z)| \leq \min_{p\in P_k} \|p(A)\|. \tag{3.25}$$

The bound (3.25) often provides very good estimate of ideal GMRES. The catch is that determination of $\mathcal{H}_k(A)$ is a difficult and open problem. For some special classes of matrices (Jordan blocks, banded triangular Toeplitz[2] matrices and block diagonal matrices with triangular Toeplitz blocks) A. Greenbaum and her co-workers introduced a several theoretical results about $\mathcal{H}_k(A)$ [25], [30], [31] and [32]. Needless to say, for larger applicability the bound (3.25) is necessary to extend the class of matrices for which $\mathcal{H}_k(A)$ is known.

[2] A Toeplitz matrix or diagonal-constant matrix, named after Otto Toeplitz, is a matrix in which each diagonal is constant.

Chapter 4

GMRES convergence curve

As has been shown in previous chapter, convergence analysis of GMRES in the terms of some simple properties of system matrix A, like eigenvalues or ε-pseudospectrum is in general unsuccessful, or at least insufficient. Described methods also don't incorporate the right hand side b or the initial guess x_0 in the analysis, but the convergence does depend on them, so we can consider this as a weakness of all these methods.

Following sections do not describe any convergence bound, but give a very useful insight into the behavior of the GMRES method. Basic idea is that if matrices A and B generate the same Krylov residual spaces $AK_k(A, r_0)$ and $BK_k(B, r_0)$ in each step k for the same initial residual, GMRES applied to the linear systems $Ax = b_1$ and $Bx = b_2$, with initial residuals $r_0 = b_1 - Ax_{01}$ and $r_0 = b_2 = Ax_{02}$, generates the same residual vectors in each step. We will express this by writing

$$\text{GMRES}(A, r_0) = \text{GMRES}(B, r_0). \tag{4.1}$$

We are interested in characterization of matrices that generates the same residual spaces. This approach was developed by A. Greenbaum and Z. Strakoš and introduced in a paper from 1994 "Matrices that Generate the Same Krylov Residual Spaces" [35]. Another two papers "Any nonincreasing convergence curve is possible for GMRES" from 1996, written by A. Greenbaum, V. Pták and Z. Strakoš [34] and "Krylov sequences of maximal length and convergence of GMRES" from 1998, written by M. Arioli, V. Pták and Z. Strakoš [3] develop this idea even further. We would like to follow the flow of the ideas of these three papers in the following four sections. We will assume that matrices are nonderogatory, which means that minimal polynomial of a matrix is identical with characteristic polynomial, consequently GMRES can converge to exact solution in N-th step, where N is the dimension of the problem. This assumption simplifies the notation. The modification to the general case is straightforward. We also assume without the loss of generality that the initial guess x_0 is zero and the right hand side b is the initial residual, $b \equiv r_0$. We would like to emphasize, that $x_0 = 0$ is in many cases the only reasonable choice [56]. A random x_0 might provide completely misleading information about the convergence measured by the relative residual norm $\|r_n\|/\|r_0\|$. The initial guess x_0 containing no useful information about an exact solution might lead to $\|r_0\| \gg \|b\|$ and the relative residual norm lacks backward error interpretation. In fact random choice potentially creates an illusion of fast

convergence measured by relative residual norm. A simple and computationally cheap trick to avoid this difficulties is the rescaling of the initial approximation. Authors of [56] describe this trick and refer to a private communication with C. Hegedüs. Let x_p be a preliminary initial guess and ζ_{min} scaling parameter such that

$$\|r_0\| = \|b - Ax_p\zeta_{min}\| = \min_{\zeta} \|b - Ax_p\zeta\|, \quad \zeta_{min} = \frac{b^*Ax_p}{\|Ax_p\|} \tag{4.2}$$

For $x_0 = x_p\zeta_{min}$ is ensured $\|r_0\| \leq \|b\|$ and relative residual norm is a proper measure of convergence.

4.1 Matrices that generate the same Krylov residual spaces

In the beginning of this section we would like to remind a few simple facts. It's suitable to keep this facts in mind during reading the following sections as it will prevent unnecessary misunderstandings.

- The fact that two Krylov spaces $K_k(A, r_0)$ and $K_k(B, r_0)$ are equal doesn't mean, that two Krylov residual spaces $AK_k(A, r_0)$ and $BK_k(B, r_0)$ are also the same. Consider for example the Krylov spaces $K_k(A, r_0)$ and $K_k(A + \alpha I, r_0)$, where α is arbitrary scalar. They are the same for all k, but $AK_k(A, r_0)$ and $(A+\alpha I)K_k((A+\alpha I), r_0)$ are different spaces and GMRES behaves differently when applied to A and $(A + \alpha I)$.

- The fact, that two Krylov residual spaces are the same $AK_k(A, r_0) = BK_k(B, r_0)$ doesn't mean that individual vectors $Ar_0, \ldots, A^kr_0$ and $Br_0, \ldots, B^kr_0$ are necessarily the same, only the spans of these vectors must be equal $\text{span}\{Br_0, \ldots, B^kr_0\} = \text{span}\{Br_0, \ldots, B^kr_0\}$.

- As we mentioned before, k-th residual can by written in the form (2.9) $r_k = P_k(A)r_0$. The fact that $P_k(A)r_0 = \hat{P}_k(B)r_0$ doesn't mean that polynomials P_k and $\hat{P}_k$ are equal.

Let $\mathcal{W} = \{w_1, \ldots, w_k\}$ be an orthonormal basis of $\text{span}\{Ar_0, \ldots, A^kr_0\}$ and let W be the matrix with orthonormal columns $w_1, \ldots, w_k$. It is well known that Arnoldi process with starting vector $w_1 = Ab/\|Ab\|$ generates unreduced upper Hessenberg matrix H [1] such that

$$AW = WH. \tag{4.3}$$

Following theorem and its proof can be found in [35, Theorem 2.1]. We just slightly modified the notation. It describes full parametrization of the set of matrices B, that generates the same Krylov residual spaces.

[1] Unreduced upper Hessenberg matrix is a matrix with zero elements under the first subdiagonal and nonzero elements on the first subdiagonal.

Theorem 4.1 *Using the above notation, let B be of the form*

$$B = W\check{R}HW^*, \tag{4.4}$$

where $\check{R}$ is any nonsingular upper triangular matrix. Then

$$BK_k(B, v) = AK_k(A, v), \quad k = 1, \ldots, n. \tag{4.5}$$

Conversely, any matrix B that satisfies (4.5) *is of the form* (4.4).

Proof: Suppose the order one Krylov residual spaces are the same; i. e.,

$$Bv = cAv \tag{4.6}$$

for some nonzero scalar c. The higher order spaces $BK_k(B, v)$ and $AK_k(A, v)$, $k > 1$, will be the same if B satisfies

$$BW = W\tilde{H} \tag{4.7}$$

for some unreduced upper Hessenberg matrix $\tilde{H}$. To see that this is so, assume that $BK_{k-1}(B, v) = AK_{k-1}(A, v)$. The $(k-1)^{st}$ column of equation (4.7) can be written as

$$Bw_{k-1} = \sum_{j=1}^{k} w^j \tilde{H}_{j,k-1}. \tag{4.8}$$

It follows that Bw_{k-1}, and hence $B^k v$, is a linear combination of $w_1, \ldots, w_k$, and since the coefficient of w_k is nonzero, $B^k v$ is linearly independent of span$\{w_1, \ldots, w_{k-1}\}$ = span$\{Bv, \ldots, B^{k-1}v\}$. Thus the order k spaces are the same and the proof is by induction. Conversely, it is clear that (4.6) and (4.7) are necessary conditions in order for (4.5) to be satisfied.

Taking B to be of the form $B = W\tilde{H}W^*$, then, the condition (4.6) becomes upon multiplying each side by W^*,

$$\tilde{H}W^*v = cW^*Av. \tag{4.9}$$

But W^*Av is just a constant times e_1, the first unit vector,

$$W^*Av = \hat{c}e_1 \tag{4.10}$$

and W^*v satisfies

$$W^*v = W^*A^{-1}Av = W^*(WH^{-1}W^*)Av = \hat{c}H^{-1}e_1.$$

Therefore (4.9) can be written as

$$\tilde{H}H^{-1}e_1 = ce_1. \tag{4.11}$$

Clearly, if $\tilde{H}$ is of the form $\check{R}H$, where $\check{R}$ is a nonsingular upper triangular matrix, then $\tilde{H}$ is an unreduced upper Hessenberg matrix satisfying (4.11). Conversely, if $\tilde{H}$ is an unreduced upper Hessenberg matrix satisfying (4.11), and if we write $\tilde{H}$ in the form XH, then the first column of X must be scalar multiple of e_1. But the requirement that XH be upper Hessenberg then implies that the

elements below the diagonal in subsequent columns of X are also zero, and the requirement that XH be unreduced implies that the diagonal elements of X are nonzero. Thus, the only matrices B satisfying (4.5) are of the form $B = W\tilde{H}W^*$, where $\tilde{H} = \check{R}H$.

□

There are many parametrizations of matrices generating the same sequences of Krylov residual subspaces. Another characterization provides the decomposition of the form

$$B = W\hat{R}\hat{H}W^*, \tag{4.12}$$

where $\hat{R}$ is any regular upper triangular matrix and $\hat{H}$ is of the form

$$\hat{H} = \begin{pmatrix} 0 & . & . & 0 & 1/\langle v, w_N\rangle \\ 1 & & & . & -\langle v, w_1\rangle/\langle v, w_N\rangle \\ & . & & . & . \\ & & . & . & . \\ 0 & . & . & 1 & -\langle v, w_{N-1}\rangle/\langle v, w_N\rangle \end{pmatrix}. \tag{4.13}$$

If and only if the matrices A and B are of the form (4.12), then

$$BK_k(B, v) = AK_k(A, v),\ k = 1, \ldots, N. \tag{4.14}$$

We would like to point out that the decomposition (4.12) isolate the information about the convergence curve in the matrix $\hat{H}$. The decomposition shows the GMRES convergence curve almost explicitly. To see that, we need to realize, that if we use the right hand side b as an initial residual and express it in the orthogonal basis $\mathcal{W}$:

$$b = \sum_{j=1}^{N} \langle b, w_j\rangle w_j, \tag{4.15}$$

then from the minimization property (2.33), results that the absolute value of the j-th coordinate of b in this basis is equal to the root of the difference of the squares of the norms of the $(j-1)$-st and j-th residual

$$|\langle b, w_j\rangle| = \sqrt{\|r_{j-1}\|^2 - \|r_j\|^2}. \tag{4.16}$$

Read below for the proof that (4.12) is really the parametrization of all matrices that generate the same Krylov residual spaces. The proof follows the ideas from [35]. Let $\hat{K}$ be the matrix with columns $(v, w_1, \ldots, w_{N-1})$. The condition (4.14) is equivalent to the condition, that spans of the first j, $j = 1, \ldots, N$ columns of $A\hat{K}$ and $B\hat{K}$ are the same. Further, the spans of the first j, $j = 1, \ldots, N$ columns of $A\hat{K}$ and W are the same. These facts can be written in the form of

$$A\hat{K}R' = B\hat{K}, \tag{4.17}$$

$$A\hat{K} = WR'', \tag{4.18}$$

where R' and R'' are some regular upper triangular matrices. Furthermore

$$\hat{K} = WD, \text{ where } D = \begin{pmatrix} \langle v, w_1 \rangle & 1 & 0 & . & 0 \\ \langle v, w_2 \rangle & 0 & 1 & & . \\ . & . & & . & . \\ . & . & & & 1 \\ \langle v, w_N \rangle & 0 & . & . & 0 \end{pmatrix} \tag{4.19}$$

Note that $D^{-1} = \hat{H}$, then using (4.17), (4.18), (4.19) and denoting $\hat{R} = R''R'$ gives relation $B = W\hat{R}\hat{H}W^*$ (4.12) and the proof is complete.

Let us turn our attention to the Theorem 4.1 again. This theorem provides full parametrization of the set of all pairs $\{A, b\}$ with the same GMRES convergence curve f for a given initial residuum. Let's denote this set $\mathcal{S}_1(f)$. As we mentioned before, convergence behavior of GMRES is well understood for normal matrices in the terms of eigenvalues. An interesting question is: For which matrices and initial residuals, $\mathcal{S}_1(f)$ contains a unitary matrix? Slightly different formulation of this question is: For which matrices H we can find such a matrix $\dot{R}$, that $\dot{R}H$ is normal? The answer is surprisingly easy. For any regular matrix H we can compute RQ decomposition $H = \ddot{R}Q$, where $\ddot{R}$ is regular upper triangular and Q is a unitary matrix, i. e. $\ddot{R}^{-1}$ is the desired matrix $\dot{R}$. Thus, any behavior that can be seen with the GMRES algorithm applied any matrix can be seen with the GMRES algorithm applied to a unitary matrix. Unfortunately, there is no known useful relationship between eigenvalues of matrix Q and some simple property of matrix A. Thus this approach can not be used to analyze GMRES convergence.

4.2 Any nonincreasing convergence curve is possible for GMRES

In this section we would like to show the way how to construct the matrix A in such manner that for the given right hand side b and the given spectrum $\sigma(A)$, the GMRES computes the given convergence curve. The choice of the convergence curve is restricted only by identity $\|b\| = \|r_0\|$ and assumption $r_0 = 0$. As we had shown before, the right hand side b expressed in orthogonal basis of $AK_N(A, b)$ is

$$b = \sum_{j=1}^{N} \langle b, w_j \rangle w_j \tag{4.20}$$

and

$$|\langle b, w_j \rangle| = \sqrt{\|r_{j-1}\|^2 - \|r_j\|^2}. \tag{4.21}$$

Let the nonincreasing positive sequence $f(0) \geq f(1) \geq \ldots \geq f(N-1) > 0$, $f(N) \equiv 0$ be the desired convergence curve, i. e. $f(j) = \|r_j\|$, $j = 1, \ldots, N$. Putting (4.21)together with the condition $f(j) = \|r_j\|$ we obtain

$$W^*b = (g(1), \ldots, g(N))^T \equiv h, \tag{4.22}$$

where $g(j) = \sqrt{\|r_{j-1}\|^2 - \|r_j\|^2}$ are the differences. Note that for any b and h ($\|b\| = \|h\|$), there exists a unitary matrix satisfying (4.22).

Let $\sigma(A) = \{\lambda_1, \ldots \lambda_N\}$ be the given spectrum of the matrix A and let C be the companion matrix corresponding to these eigenvalues. Further, lets define the set of vectors $\hat{\mathcal{K}} = \{b, w_1, \ldots, w_{N-1}\}$ and the matrix $\hat{K}$ with columns $(b, w_1, \ldots, w_{N-1})$. The matrix with desired eigenvalues and generating desired sequence of Krylov residual subspaces (subspaces with orthogonal basis $w_1, \ldots, w_j$) is defined by the equations

$$\begin{aligned} Ab &\equiv w_1 \\ Aw_1 &\equiv w2 \\ &\vdots \\ Aw_{N-2} &\equiv w_{N-1} \\ Aw_{N-1} &\equiv \alpha_0 b + \alpha_1 w_1 + \ldots + \alpha_{N-1} w_{N-1} \end{aligned}$$

The representation of this homomorphism in the basis $\mathcal{B}$ is the companion matrix C. Finally, the matrix A is defined by relation

$$A = \hat{K} C \hat{K}^{-1}. \tag{4.23}$$

This construction of the desired matrix A is summarized in [34, Theorem 2.1]:

Theorem 4.2 *Given a nonincreasing positive sequence*

$$f(0) \geq f(1) \geq \ldots \geq f(N-1) > 0$$

and a set of nonzero complex numbers $\{\lambda_1, \ldots, \lambda_N\}$*, there exists a matrix* A *with eigenvalues* $\lambda_1, \ldots, \lambda_N$ *and a right-hand side* b *with* $\|b\| = f(0)$ *such that the residual vectors* r_k *at each step of GMRES(A,b) satisfy*

$$\|r_k\| = f(k),\ k = 1, 2, \ldots, N-1.$$

Combining that (4.14) holds if and only if matrices A and B are of the form (4.12) and observation that $GMRES(QVQ^*, b)$ generates the same sequence of residual norms as $GMRES(C, Q^*b)$, we obtain following theorem (see [34, Theorem 3.2]).

Theorem 4.3 *Given a nonincreasing positive sequence*

$$f(0) \geq f(1) \geq \ldots \geq f(N-1) > 0,$$

the residual vectors r^k *at each step of* $GMRES(A, b)$ *satisfy*

$$\|r^k\| = f(k),\ k = 1, \ldots, N-1,$$

if and only if A *is of the form*

$$A = W \hat{R} \hat{H} W^*$$

and b *satisfies*

$$W^* b = (g(1), \ldots, g(N))^T,$$

where W *is a unitary matrix,* R *is a nonsingular upper triangular matrix,* $\hat{H}$ *is defined in* (4.13)*, and* $g(1), \ldots, g(N)$ *are defined as*

$$g(k) = \sqrt{(f(k-1))^2 - (f(k))^2},\ k = 1, \ldots, N.$$

4.3 Matrices that generate the same convergence curve and have the same spectra

Finally the following theorem summarizes the Theorem 2.1 and Corollary 2.4 and their proofs from [3]. This theorem provides two full parametrizations of matrices with the same spectrum and convergence curve. The second parametrization will be frequently used in chapters 5, 6 and 7.

Theorem 4.4 *Suppose we are given n positive numbers*

$$f(0) \geq f(1) \geq \ldots \geq f(N-1) > 0$$

and N complex numbers $\lambda_1, \ldots, \lambda_N$*, all different from zero. Let A be an N by N matrix, b an N-dimensional vector. Then the following assertions are equivalent:*

1. *The spectrum of A is* $\{\lambda_1, \ldots, \lambda_N\}$ *and GMRES applied to the pair* A, b *yields residuals* $r_0, \ldots, r_{N-1}$ *such that*

$$\|r_j\| = f(j), \quad j = 0, 1, \ldots, N-1$$

2. *The matrix A is of the form* $A = W\bar{R}C\bar{R}^{-1}W^*$ *and b=Wh, where C is the companion matrix corresponding to the polynomial q, W is unitary, and* $\bar{R}$ *nonsingular upper triangular such that* $\bar{R}s = h$*.*

   The polynomial q and the vectors s, h are constructed as follows:

$$\begin{aligned} q(z) &= (z-\lambda_1)\ldots(z-\lambda_N) = z^N - \sum_{j=0}^{N-1} \alpha_j z^j, \\ s &= (\xi_1, \ldots, \xi_N)^T \quad \textit{where} \quad 1 - (\xi_1 z + \cdots + \xi_N z^N) = \prod_{i=1}^{N} \left(1 - \frac{z}{\lambda_i}\right) \\ h &= (\eta_1, \ldots, \eta_N)^T \quad \textit{where} \quad \eta_j = (f(j-1)^2 - f(j)^2)^{1/2} \\ & \qquad\qquad\qquad\qquad\quad j = 1, \ldots, N \quad f(N) \equiv 0 \end{aligned}$$

3. *Matrix A is of the form*

$$A = WYCY^{-1}W^* \textit{ and } b = Wh, \tag{4.24}$$

   where W is a unitary matrix, Y is given by

$$Y = \begin{pmatrix} \boxed{h} & \boxed{R} \\ & \boxed{0} \end{pmatrix} \tag{4.25}$$

   and R is any $(N-1) \times (N-1)$ *nonsingular upper triangular matrix.*

Proof: Assume that condition 1 is satisfied. Since $f(j-1) > 0$, the dimension of each Krylov space $K_j(A,b) = \mathrm{span}\{b, \ldots, A^{j-1}b\}$ equals j, for $j = 1, \ldots, N$. In particular $\{b, Ab, \ldots, A^{n-1}b\}$ is the basis. Let B be the matrix containing the naive basis of Krylov residual subspace $AK_N(A,b)$ in its columns. Since q is a characteristic polynomial of matrix A and consequently q annihilates A, we obtain following identities

$$AB = BC \tag{4.26}$$

$$b = \sum_{1}^{N} \xi_j A^j b = Bs \tag{4.27}$$

Let $B = \tilde{W}\tilde{R}$ be the QR decomposition of the matrix B ($\tilde{W}$ is unitary matrix and $\tilde{R}$ is an upper triangular). The condition $\|r^j\| = f(j), j = 0, 1, \ldots, N-1$, implies $Bs = b = \tilde{W}\Gamma h$, where $\Gamma = \mathrm{diag}\{\gamma_j\}$, $|\gamma_j| = 1$, $j = 1, \ldots, N$. Set $W = \tilde{W}\Gamma$, $\bar{R} = \Gamma\tilde{R}$, then we obtain

$$AW\bar{R} = A\tilde{W}\tilde{R} = AB = BC = W\bar{R}C, \tag{4.28}$$

whence $A = W\bar{R}C\bar{R}^{-1}W^*$. Using (4.27)

$$W\bar{R}s = \tilde{W}\tilde{R}s = Bs = \tilde{W}\Gamma h = Wh, \tag{4.29}$$

so that $\bar{R}s = h$. this proves an implication 1. $\Rightarrow$ 2.

Assume that condition 2 is satisfied. A is similar to C, so A has the same eigenvalues as C $\lambda_1, \ldots, \lambda_N$. To prove an implication 2. $\Rightarrow$ 1. is sufficient to show that for the given A, b, and any j, $j = 1, \ldots, N$, the first j column vectors $w_1, \ldots, w_j$ of the matrix W represents the unitary basis of the Krylov residual subspace $AK_j(A,b)$. We will prove this by induction. Consider $b_j = A^j b$, $j = 1, \ldots, N$. From $b = Wh = W\bar{R}s$ follows that

$$b_1 = Ab = W\bar{R}Cs = W\bar{R}e_1 = (\bar{R})_{11}w_1.$$

Assume that $b_{j-1} = W\bar{R}e_{j-1}, j \leq N$. Then

$$b_j = Ab_{j-1} = W\bar{R}Ce_{j-1} = W\bar{R}e_j = W(\bar{R})_{\bullet j},$$

where $(\bar{R})_{\bullet j}$ denotes the j-th column of the matrix $\bar{R}$.

Before the proof that assertions 1 and 2 are equivalent to assertion 3, let's do following observation. Let $\mathcal{S}_2(f, \{\lambda_1, \ldots, \lambda_N\})$ be the set of all pairs $\{A, b\}$ such that $\sigma(A) = \{\lambda_1, \ldots, \lambda_N\}$ and the sequence of the norms of residuals of GMRES applied to $\{A, b\}$ is f. We already defined the set $\mathcal{S}_1(f)$ in section 4.1 as the set of all pairs $\{A, b\}$ with the same GMRES convergence curve f. It is obvious that any pair $\{A, b\} \in \mathcal{S}_1$ is parametrized by the decomposition of the form $A = W\hat{R}\hat{H}W^*$, see theorem 4.3. Further, any pair $\{A, b\} \in \mathcal{S}_2$ is parametrized by the relations

$$A = W\bar{R}C\bar{R}^{-1}W^*, \quad b = Wh \quad \text{and} \quad Rs = h. \tag{4.30}$$

It is interesting to see, how the parametrizations of these sets are related, i. e. what is the relation between the decomposition (4.30) and (4.12). Set $Y = \bar{R}C^{-1}$.

Considering $Rs = h$

$$Y = \bar{R}C^{-1} = \begin{pmatrix} \boxed{h} & 1 & & \\ 0 & \ddots & & \\ & \ddots & 1 & \\ & & 0 & \end{pmatrix} \begin{pmatrix} 1 & \boxed{0} \\ 0 & \boxed{R_{N-1}} \end{pmatrix},$$

where R is in this case the $(N-1)$st left principal submatrix of the matrix $\bar{R}$. Then

$$\bar{R}C\bar{R}^{-1} = \bar{R}(\bar{R}C^{-1})^{-1} = \bar{R} \begin{pmatrix} 1 & \boxed{0} \\ 0 & \boxed{R} \end{pmatrix}^{-1} \begin{pmatrix} \boxed{h} & 1 & & \\ 0 & \ddots & & \\ & \ddots & 1 & \\ & & 0 & \end{pmatrix}^{-1}$$

and now it's easy to see, that

$$\hat{R} = \bar{R} \begin{pmatrix} 1 & \boxed{0} \\ 0 & \boxed{R} \end{pmatrix}^{-1} \quad \text{and} \quad \hat{H} = \begin{pmatrix} \boxed{h} & 1 & & \\ 0 & \ddots & & \\ & \ddots & 1 & \\ & & 0 & \end{pmatrix}^{-1}$$

Previous considerations goes from the decomposition (4.30) to the decomposition (4.12) and thus from $\mathcal{S}_2$ to $\mathcal{S}_1$. One can get from $\mathcal{S}_1$ to $\mathcal{S}_2$ repeating all consideration above backwards. We would like to emphasize, that matrix W only represents the change of the basis and does not play a substantial role. As you can see, the number of free parameters describing $\mathcal{S}_1$, the nonzero entries of the matrix $\hat{R}$ is $(N+1)N/2$. By fixing the spectrum we decrease the number of free parameters describing $\mathcal{S}_1$ in comparison to $\mathcal{S}_2$ by N. The last column of the matrix $\bar{R}$ is determined using equality $Rs = h$ by following equation

$$\xi_N(\bar{R})_{\bullet N} = \begin{pmatrix} R(\xi_1, \ldots, \xi_{N-1})^T \\ 0 \end{pmatrix}, \tag{4.31}$$

where R is in this case the $(N-1)$st left principal submatrix of the matrix $\bar{R}$. The relation (4.31) says that any nonsingular upper triangular matrix $\bar{R}$ satisfying $Rs = h$ is determined by its $(N-1)$st left principal submatrix representing free parameters and the last column is determined by vectors h and s. Summarizing the observation above, we have proved the equivalence between assertions 1, 2 and 3.

The proof is complete now.

□

4.4 Krylov sequences of maximal length

Let m be the degree of the minimal polynomial of matrix A. According to the relation (2.34), the $GMRES(A, b)$ finds exact solution at most at m steps for any

right hand side. Moreover, for any matrix A there exists proper right hand side b, such that $GMRES(A, b)$ converges to the exact solution at the m-th step.

Let $J = F^{-1}AF$ be the Jordan canonical form of A, and

$$q_{\min}(\lambda) = (\lambda - \lambda_1)^{m_1} \dots (\lambda - \lambda_j)^{m_j} \qquad \lambda_i \neq \lambda_j \quad \forall\, i \neq j$$

be the minimal polynomial of A, i. e. $\lambda_1, \dots, \lambda_j$ are all the distinct eigenvalues and $m_1, \dots, m_j$ are all the sizes of the largest Jordan blocks corresponding to the proper eigenvalues. Operator $(\lambda_k I - A)^{m_k}$ projects Jordan canonical vectors corresponding to the eigenvalue λ_k to the zero vector. Nullspaces of these operators form invariant subspaces. A union of these subspaces forms whole $\mathbb{F}^N$. Thus any right hand side can be decomposed to it's components in invariant subspaces corresponding to the eigenvalues λ_k k=1,. . . ,j, i. e.

$$b = t_1 + t_2 + \dots t_j \quad t_k \in Ker(\lambda_k I - A)^{m_k}.$$

Putting these observations together with the fact that for any polynomial p holds equality $p(A) = Fp(J)F^{-1}$, we can formulate following statement (see Section 3 of [3]): The vector b yield the Krylov sequences of maximal length if and only if

$$(\lambda_k I - A)^{N-1} t_k \neq 0 \text{ for } k = 1, \dots, j. \tag{4.32}$$

Equations (4.32) are necessary and sufficient conditions for GMRES(A,b) to converge to the exact solution x at the last step N.

Chapter 5

The connections and differences between the spectral decomposition and the decomposition of the form $A = WYCY^{-1}W^*$

In this chapter, we would like to show a connection between the characterization of GMRES convergence, based on the spectral decomposition $A = Q\Lambda Q^*$, $A \in \mathbb{F}^{N\times N}$ (see (3.1)) of normal matrices, mentioned in Chapter 3 and decomposition of the form $A = WYCY^{-1}W^*$ (see (4.24)) described in section 4.3, Theorem 3.4. We will refer to this decomposition as to APS decomposition according to the authors M. Arrioli, V. Pták, and Z. Strakoš of the paper [3], where this decomposition was introduced for the first time.

A matrix $A \in \mathbb{F}^{N\times N}$ is similar to the companion matrix of its characteristic polynomial if and only if the minimal and characteristic polynomials of A are identical [40, Theorem 3.3.15]. Let's assume, that matrix A is nonderogatory and the GMRES(A,b) converges to the solution at the very last step. Consequently, the factors of the APS decomposition are square matrices. Further, let us assume that the companion matrix C of the matrix A is diagonalizable. This assumption is crucial.

Let S be the matrix containing the eigenvectors $s_k = (s_{1k}, \ldots, s_{Nk})^T$ of the matrix C as its columns. Because of the form of the companion matrix C

$$C = \begin{pmatrix} 0 & & & \alpha_0 \\ 1 & & & \alpha_1 \\ & \ddots & & \vdots \\ & & 1 & \alpha_{N-1} \end{pmatrix}, \tag{5.1}$$

the eigenvectors s_k are uniquely determined (apart from their normalization) by

the corresponding eigenvalues and the set of equations

$$\begin{aligned} \alpha_0 s_{Nk} &= \lambda_k s_{1k} \\ s_{1k} + \alpha_1 s_{Nk} &= \lambda_k s_{2k} \\ &\vdots \\ s_{(N-1)k} + \alpha_{(N-1)} s_{Nk} &= \lambda_k s_{Nk}. \end{aligned}$$

Thus the matrix $C = S\Lambda S^{-1}$ is diagonalizable if and only if C has all the eigenvalues distinct. Further, let D be diagonal matrix

$$D = \begin{pmatrix} \frac{1}{\|Ys_1\|} & & & \\ & \frac{1}{\|Ys_2\|} & & \\ & & \ddots & \\ & & & \frac{1}{\|Ys_N\|} \end{pmatrix}, \tag{5.2}$$

where Y is a factor of the APS decomposition. Using D to norm columns of the matrix S we obtain

$$C = SD\Lambda D^{-1}S^{-1}. \tag{5.3}$$

The relations (5.3) and (4.24) yield

$$A = WYSD\Lambda D^{-1}S^{-1}Y^{-1}W^*. \tag{5.4}$$

The columns of the matrix $WYSD$ are unit eigenvectors of the matrix A, because of the factor D. Comparing the decomposition (5.4) with spectral decomposition (3.1) shows, that

$$WYSD = Q, \tag{5.5}$$

i. e. $WYSD$ is a unitary matrix. Consequently

$$A = Q\Lambda Q^* = (WYSD)\Lambda(WYSD)^*. \tag{5.6}$$

The decomposition (5.6) connects the APS decomposition and the spectral decomposition of the matrix A. Please note that all the factors of the spectral decomposition (3.1) are normal matrices. On the other hand, the factors Y and C are nonnormal, because of their form (see (4.25) and (5.1)), even if the matrix A is normal, except for the case of the matrix C_1 (5.7) and the matrix Y_1 (5.8) multiplied by arbitrary constant, which are described below. Matrix C_1 corresponds to the characteristic polynomial of the form $p(\lambda) = \lambda^N - 1$.

$$C_1 = \begin{pmatrix} 0 & & & 1 \\ 1 & & & 0 \\ & \ddots & & \vdots \\ & & 1 & 0 \end{pmatrix} \tag{5.7}$$

$$Y_1 = \begin{pmatrix} 0 & 1 & & \\ \vdots & & \ddots & \\ 0 & & & 1 \\ 1 & & & 0 \end{pmatrix} \tag{5.8}$$

The matrix Y_1 is a factor of the APS decomposition of such a matrix A with the right hand side b that GMRES(A,b) does not make any progress until the last step, $\|r_{N-1}\| = 1$ and all the vectors $Ab, \ldots, A^{N-1}b$ are orthnormal. To clarify why all the vectors $Ab, \ldots, A^{N-1}b$ are orthnormal we would like to recall that the matrix Y is of the form (4.25) and R is the $(N-1)$-st left principal submatrix of the matrix $\bar{R}$. Further, $AK = W\bar{R}$ is a QR decomposition of matrix AK, which columns are $Ab, \ldots, A^N b$, i. e. the naive basis of Krylov residual subspace; see theorem 4.4. In this case $\mathrm{span}\{Ab, \ldots, A^N b\}$ is equivalent to $\mathbb{C}^{\mathbb{N}}$.

The spectral decomposition contains an information about system matrix A only. Eigenvalues of A are diagonal entries of Λ and eigenvalues are columns of matrix Q. The APS decomposition contains information about system matrix A and the right hand side, because the matrix C contains information about about the spectrum and the matrix Y contains the information about the convergence of GMRES(A,b) in the first column.

Chapter 6

The minimization of nonnormality

Throughout this chapter we assume that the initial GMRES approximation of the solution of the problem of the form (1.1) is $x_0 = 0$.

The papers [35], [34] and [3] introduce full parametrizations of matrices and initial residuals with the same GMRES convergence curve, or eigenvalues and GMRES convergence curve. These results are summarized in chapter 4. Particularly section 4.3 describes the decomposition of the form $A = WYCY^{-1}W^*$; see (4.24). We refer to this decomposition as to the APS decomposition; see chapter 5. At this point, we would like to remind a form of matrices which appear in the APS decomposition.

- Matrix W. Columns of the matrix W, $w_1, \ldots, w_N$ are orthonormal and for any $k = 1, \ldots, N$ vectors $w_1, \ldots, w_k$ form basis of Krylov residual subspace $\mathrm{span}\{Ab, \ldots, A^k b\}$.

- Matrix Y is of the form

$$Y = \left(\begin{array}{|c|c|} \hline \hat{h} & R \\ \hline \eta_N & 0 \\ \hline \end{array} \right) \tag{6.1}$$

  where $\hat{h} = (\eta_1, \ldots, \eta_{N-1})^T$. The convergence curve of GMRES(A,b) is determined by the set of equations $\|\eta_j\| = \sqrt{\|r_{j-1}\|^2 - \|r_j\|^2}$, $j = 1, \ldots, N$. Matrix R is the $(N-1)$st left principal submatrix of the matrix $\bar{R}$; see theorem 4.4.

- Matrix C is a companion matrix of the form (5.1) and contains information about eigenvalues only.

As we can see, any modification of the submatrix R of Y does not affect convergence curve of GMRES applied to the matrix A and the right hand side b or the spectrum of the matrix A. Consequently, the APS decomposition parametrizes the set of all the pairs $\{A, b\}$ such that the sequence f of the norms of the residuals of GMRES applied to any pair $\{A, b\}$ is the same and the spectrum of any matrix A included in this set is the same also $\sigma(A) = \{\lambda_1, \ldots, \lambda_N\}$. The free parameter is

the submatrix R. We denote this set as $\mathcal{S}_2(f, \{\lambda_1, \ldots, \lambda_N\})$. From another point of view, we can consider the APS decomposition as a tool to construct the set $\mathcal{S}_2(f, \{\lambda_1, \ldots, \lambda_N\})$ for any prescribed convergence curve and the spectrum $\sigma(A)$.

It is well known, that the worst-case convergence curve of GMRES applied to any normal matrix is bounded by the solution of min-max problem (3.7)

$$\frac{\|r_k\|}{\|r_0\|} \leq \min_{p \in P_k} \max_i |p(\lambda_i)|,$$

i. e. the convergence is determined by eigenvalues. A. Greenbaum and Z. Strakoš showed in paper [35] that if A is close to Hermitian, then the eigenvalues of A essentially determine the behavior of the GMRES iterations but that, in general, eigenvalue information alone is never sufficient to ensure rapid convergence of the GMRES algorithm. According to these facts we expect that if the convergence curve of GMRES applied to the nonnormal matrix A and the right hand side b seems to be determined by the eigenvalues $\lambda_1, \ldots, \lambda_N$ of the matrix A, than the set $\mathcal{S}_2(f, \{\lambda_1, \ldots, \lambda_N\})$ containing the matrix A and the right hand side b contains also a nearly normal matrix with the right hand side b.

We wanted to examine this assumption. We tried to find for several chosen matrices such a matrix that has the same eigenvalues and for the same initial residual the same convergence curve, but a minimal departure from normality. The rest of this chapter is devoted to these numerical experiments.

We would like to mention that we assume that if the set $\mathcal{S}_2(f, \{\lambda_1, \ldots, \lambda_N\})$ contains a pair of the highly nonnormal system matrix A and the right hand side b arisen by the discretization of some real world problem contains also a matrix close to normal with an arbitrary right hand side, the nonnormality of the matrix A can be caused by wrong formulation of the problem. The nonnormality can be just artificial problem in such a case.

6.1 The choice of measure of departure from normality

The first problem, that we had to solve was the choice of the suitable measure of departure from normality to minimize. There are many measures of nonnormality. A lot of measures of nonnormality and conditions of normality are discussed in papers [21], [37] and [20]. We selected two of them $\|A^*A - AA^*\|_F$ and $\|A^*A - AA^*\|_F / \|A\|_F^2$, where $\|.\|_F$ denotes the Frobenius norm. We selected these measures, because it is easy to compute them. We refer to $\|A^*A - AA^*\|_F$ as the absolute measure and to $\|A^*A - AA^*\|_F / \|A\|_F^2$ as the relative measure. These measures do not reflect any unitary transformations of matrix A, thus

$$\|A^*A - AA^*\|_F = \|YCY^{-1}Y^{-*}C^*Y^* - Y^{-*}C^*Y^*YCY^{-1}\|_F. \tag{6.2}$$

Analogical relation holds for the relative measure

$$\frac{\|A^*A - AA^*\|_F}{\|A\|_F^2} = \frac{\|YCY^{-1}Y^{-*}C^*Y^* - Y^{-*}C^*Y^*YCY^{-1}\|_F}{\|YCY^{-1}\|_F^2} \tag{6.3}$$

We will proceed in the following way:

- To compute for chosen matrix A and the right hand side b vector $h = (\eta_1, \ldots, \eta_N)$ which uniquely determines the GMRES convergence curve and the companion matrix C of the matrix A which uniquely determines the spectrum.

- To find such a matrix that is included with the right hand side b in the set $\mathcal{S}_2(f, \{\lambda_1, \ldots, \lambda_N\})$ corresponding to the chosen matrix A and the right hand side b mentioned in the previous point and its absolute departure from normality is minimal, i. e. to minimize the function (6.2) via submatrix R of Y using the data (matrix C and vector h) from previous point. Further, to find such a matrix that is included with the right hand side b in the set $\mathcal{S}_2(f, \{\lambda_1, \ldots, \lambda_N\})$ corresponding to the chosen matrix A and the right hand side b mentioned in the previous point and its relative departure from normality is minimal, i. e. to minimize the function (6.3) via submatrix R of Y using the data (matrix C and vector h) from previous point.

There is an upper estimate of the relative norm [12]:

$$\frac{\|A^*A - AA^*\|_F}{\|A\|_F^2} \leq \sqrt{2}. \tag{6.4}$$

We expected that the relative measure is more suitable for our purposes, because some change of submatrix R can enlarge the absolute measure, but only because it enlarges $\|A\|_F$. Even though the absolute measure for such a case increase, the relative measure can decrease. We constructed an example of such a behavior. Let the companion matrix $C_2 \in \mathbb{F}^{5\times 5}$ be of the form

$$C_2 = \begin{pmatrix} 0 & & & & 1 \\ 1 & & & & 1 \\ & 1 & & & 1 \\ & & 1 & & 1 \\ & & & 1 & 1 \end{pmatrix}. \tag{6.5}$$

Let the matrices $Y_2 \in \mathbb{F}^{5\times 5}$ and Y_3 be of the form

$$Y_2 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & & 1 & 1 & 1 \\ 1 & & & 1 & 1 \\ 1 & & & & 10 \\ 1 & & & & 0 \end{pmatrix} \tag{6.6}$$

and

$$Y_3 = \begin{pmatrix} 1 & 10 & 1 & 1 & 1 \\ 1 & & 1 & 1 & 1 \\ 1 & & & 1 & 1 \\ 1 & & & & 1 \\ 1 & & & & 0 \end{pmatrix}. \tag{6.7}$$

Matrices $A_2 = Y_2 C_2 Y_2^{-1}$ and $A_3 = Y_3 C_2 Y_3^{-1}$ has the same convergence curve and spectrum. The only difference between matrices A_2 and A_3 is a slight modification

of the submatrix R. Such a modification of the matrix R causes the increase of the absolute measure of nonnormality of the matrix A_2 from $\|A_2^*A_2 - A_2A_2^*\|_F = 287.0$ to the absolute measure of nonnormality of the matrix A_3, $\|A_3^*A_3 - A_3A_3^*\|_F = 304.2$. However, the relative measure decreases from $\|A_2^*A_2 - A_2A_2^*\|_F/\|A_2\|_F^2 = 1.38$ to $\|A_3^*A_3 - A_3A_3^*\|_F/\|A_3\|_F^2 = 1.30$. These results are summarized in the table 6.1

Table 6.1:

	A_2	A_3
Absolute measure of nonnormality	287.0	304.2
Relative measure of nonnormality	1.38	1.30

Nevertheless the absolute measure appears to be for some cases a better choice, because we were not able to minimize relative departure from normality as effectively as the absolute one; see the results in the section 6.3.

6.2 Computation of the decomposition

As we mentioned in the previous section, we wanted to find a minimum of functions (6.2) and (6.3) via submatrix R for some chosen matrices. For such chosen matrices, we needed to compute a companion matrix C and GMRES convergence curve, before the minimization. Further, we had to pick an initial point of minimization. The starting point of minimization can be any upper triangular matrix of proper dimension. The results of the minimization in section 6.3 show that it appears to be advantageous to use submatrix R computed by APS decomposition of the matrix A as a starting point of minimization. Thus we needed to compute APS decomposition before minimization.

We used a program `compapshouse` designed by Gérard Meurant in Matlab to compute the factors W, Y and C of APS decomposition. A code of this program is presented in Appendix, page 67.

Originally we wanted to compute APS decomposition and than minimize the departure from normality for some large matrix arisen by discretization of some real world problem. Unluckily, the factors of APS decomposition Y, C and Y^{-1} are usually highly nonnormal and the submatrix R is usually almost singular, thus the computation of APS decomposition and following minimization of functions (6.2) and (6.3) appeared to be extremely difficult. Let us recall, that the matrix R is the $(N-1)$st left principal submatrix of the matrix $\bar{R}$ and $AK = W\bar{R}$ is a QR decomposition of matrix AK, which columns are $Ab, \ldots, A^Nb$, i. e. the natural basis of Krylov residual subspace. These vectors are usually almost linearly dependant and that causes that R is nearly singular.

`Compapshouse` originally saved numbers using 8 bytes (double accuracy), which is default Matlab accuracy. We tried to compute APS decomposition of the matrix A motivated by the discretization of the convection-diffusion problem. We would like to emphasize that reasonable discretizations of the convection-diffusion

problem yields very large matrices. Nevertheless, because of the difficulties mentioned above and below, we were not able to compute the APS decomposition of such a large matrix. Consequently we were forced to do numerical experiments with very small matrices.

For our first numerical experiment we used a matrix that has been used and studied in many publications, see [18], [17], [19], [24], [26] and [43]. This matrix arose by discretization of convection-diffusion problem with Dirichlet boundary condition:

$$-\nu\triangle u + w \cdot \nabla u = 0 \ \text{ in } \ \Omega = (0,1) \times (0,1), \quad u = g \text{ on } \partial\Omega, \tag{6.8}$$

where ν is a scalar diffusion parameter. We chose the right hand side according to [43, example 2.2], i. e. the right hand side corresponds to nonzero boundary only on a part of the right hand side boundary of a unit square. In the context of notation in [43] we chose a parameter $k = \sqrt{N} - 1$, which is not substantial, details can be found in [43]. The program failed miserably to compute APS decomposition even for $A \in \mathbb{F}^{16\times 16}$, $\nu = 10^{-2}$. For this parameters is $\|A\| = 0.35$ and $\|A - A_{aps}\| = 224.63$, where $A_{aps} = WYCY^{-1}W^*$ is an approximation of the matrix A from the computed factors. To compute the matrix A_{aps} we used besides the program `compapshouse` which computes factors W,Y and C also a program `YinverseVPA` that computes the inverse of Y. The code of `YinverseVPA` is presented in the Appendix. `YinverseVPA` is able to use Symbolic toolbox and save the numbers during the computations with arbitrary precision. We did not use this feature for the computation mentioned above.

Because of this failure we decided to use Matlab Symbolic toolbox and slightly modify `compapshouse` in such a way that `compapshouse` saved numbers with 128 digits precision instead of the double precision during the computations of the factors of APS decomposition. Moreover, we set the program `YinverseVPA` to save numbers with 128 digits precision during the computation of Y^{-1}. At this point we would like to thank Petr Tichý for great help with using Symbolic toolbox. We were able to compute APS decomposition with symbolic toolbox even for matrix $A \in \mathbb{F}^{36\times 36}$, $\nu = 10^{-8}$ with accuracy $\|A - A_{aps}\| = 2.8 \cdot 10^{-72}$.

Program `compapshouse` computes the last column of the companion matrix numerically. We known the spectrum of all the matrices presented in this chapter analytically. Instead of using matrix C computed by program `compapshouse` as an input data for the minimization of the departure from normality of these matrices, we used a program `CompanionMatrixVPA` to construct the companion matrix C from the known spectrum as the input data for the experiments described in subsections 6.3.2, 6.3.3 and 6.3.4. `CompanionMatrixVPA` saved the data during the computations with 128 precision. The code of `CompanionMatrixVPA` is introduced in the Appendix.

6.3 The minimization

At this point we would like to thank Ctirad Matonoha for an implementation of the minimization.

As we mentioned in the previous section, we were able to compute the factors of APS decomposition W, Y, C, Y^{-1} with high accuracy. We saved the data

during computations of the APS decomposition with 128 digits precision. The input data of the minimization are matrix C and the first column of the matrix Y, i. e. vector h. However, we did the minimization using the system UFO [46]. The system UFO uses the double precision format to save data during the minimization and the format of input data have to be double also. Thus, after we precisely computed C and h, we were forced to save them in the format double. We had to check, if such a change of precision doesn't devaluate the input data. As we mentioned earlier, we tried to compute minimums of the functions (6.2) and (6.3) for several chosen matrices. We checked for every such a matrix the norm $\|A_I - A_{Idouble}\|$, where $A_I = WY_ICY_I^{-1}W^*$. The factors W and C correspond to the factors of the APS decomposition of the matrix A computed using 128 digits precision and the matrix Y_I corresponds to the matrix Y computed using 128 digits precision, but the submatrix R is replaced by identity. The factor C was computed using program `CompanionMatrixVPA`. The matrix $A_{Idouble}$ was computed the same way, but factors of the decomposition were saved with double precision. We replaced the submatrix R by identity matrix, because the computed matrix R corresponding to the matrix A often amplifies an error of the other factors. An example of such a behavior is the matrix $A \in \mathbb{F}^{16\times 16}$ and the right hand side b described in previous section, see [43, example 2.2]. For parameter $\nu = 10^{-3}$ is $\|A_I - A_{Idouble}\| = 4.77 \cdot 10^{-4}$ and $\|A_I\| = 3.72 \cdot 10^{13}$. On the other hand $\|A - A_{Rdouble}\| = 104.17$ and $\|A\| = 0.35$ for the same parameter $\nu = 10^{-3}$, where $A_{Rdouble}$ corresponds to $A_{Idouble}$ except the computed matrix R corresponding to the matrix A was not replaced by the identity I, i. e. $A_{Rdouble}$ is an approximation of the matrix A from the computed factors saved in double precision. As we can see, submatrix R corresponding to the matrix A amplifies an error of the other factors significantly. If $\|A_I - A_{Idouble}\|$ would be big in comparison to the $\|A_I\|$ for some matrix A, than saving C and h with double precision devaluate the input data for the minimization of (6.2) and (6.3) and we are not able to do minimization in such a case. We would like to point out that we did the minimization of the absolute and the relative departure from normality for any matrix separately. In general, the results of the minimization of the relative and absolute departure from normality are two different matrices.

6.3.1 The minimization of departure from normality for modified normal matrix

As we mentioned in the beginning of the section 6.3, we did the minimization by system UFO. We used nongradient heuristic method, pattern search [39]. As we showed higher, computation of the input data for minimization (matrix C, vector h) is a very sensitive problem. Slight perturbation of some factor of the APS decomposition of the matrix A usually causes a big change of the norm $\|A - WYCY^{-1}W^*\|$. Chapter 5 shows that factors of APS decomposition are often highly nonnormal. These facts make the minimization of the functions (6.2) and (6.3) via parameter R very difficult problem.

We did following numerical experiment to find out, if UFO is able to minimize the functions (6.2) and (6.3) effectively. At first we generated a diagonal matrix Λ with eigenvalues $1, \ldots, 10$. Than we generated a random orthonormal matrix Q and right hand side $b = (1, \ldots, 1)^T$. Further we computed matrix $A_{Normal} =$

$Q\Lambda Q^*$ and its APS decomposition $A_{Normal} = Q\Lambda Q^* = WYCY^{-1}W^* \in \mathbb{F}^{10\times 10}$. The matrix A_{Normal} is obviously normal, i. e. functions (6.2) and (6.3) are zero for the matrix A_{Normal}. Further, we replaced the submatrix R of the matrix Y by identity matrix I. The resulting matrix $A_I = WY_ICY_I^{-1}W^*$ had absolute departure from normality $4.23 \cdot 10^{16}$, relative departure from normality 1.41421. Matrix A_I is not normal, nevertheless A_I has the same convergence curve and eigenvalues as normal matrix A_{Normal}, i. e. the set $\mathcal{S}_2(f, \{\lambda_1, \ldots, \lambda_N\})$ corresponding to the matrix A_I and the right hand side $b = (1, \ldots, 1)^T$ contains the pair $\{A_{Normal}, b\}$. The result of the minimization of the functions (6.2) and (6.3) via the matrix R using data computed by APS decomposition of the matrix A_{Normal} and the identity as a starting point suppose to be zero. We tried if the UFO is able to reach this minimum. The result of the minimization of the absolute departure from normality was $6.59 \cdot 10^5$ after $3 \cdot 10^5$ steps. The minimization process made great progress, because it decreased the departure from normality by 11 orders. Nevertheless we were not able to find the global minimum.

The approximation of the matrix A_{Normal} from factors W, Y and C computed using 128 precision approximates the matrix A very well. As we mentioned earlier, even slight perturbation of some factor of the decomposition usually causes significant change of the norm $\|A - WYCY^{-1}W^*\|$ as well as the change of the departure from normality of the approximation. An example of such a perturbation is saving precisely computed factors in double precision, which is necessary before the minimization. The absolute departure from normality of the matrix $Y_{double}C_{double}Y_{double}^{-1}$ is not zero but $4.48 \cdot 10^{-7}$ and the relative departure from normality is $1.16 \cdot 10^{-9}$. We tried to minimize the departure from normality for this data. The problem is the same as the previous one (same matrix C and vector h), but the initial point of the minimization is not the identity, but the submatrix R saved with double precision and computed by APS decomposition of the matrix A_{Normal}. We denote this matrix as $R_{initial}$. The minimum that we found is not zero, but for the absolute measure it is $8.10 \cdot 10^{-8}$ and for the relative measure it is $1.69 \cdot 10^{-10}$.

We were not able to find global minimal departure from normality in previous two numerical experiments. We decided to do minimization again, but again from a different starting point. This time we chose matrix $R_{initial}$ shifted by 10^{-6} as the starting point of the minimization. Shifting $R_{initial}$ by 10^{-6} means, that we added to every nonzero entry of the matrix 10^{-6}. We denote this matrix as $R_{initialShift}$. We were not able to find zero as a global minimum. We significantly decreased the departure from normality though.

The results of all the experiments described above are summarized in the table 6.2. These results show that we are not able to compute exact global minimum of the functions (6.2) and (6.3), nevertheless we are able to significantly decrease the departure from normality.

We would like to mention a very interesting observation that we did during computation of the matrix $A_{Normal} = Q\Lambda Q^*$ and its APS decomposition. Matrix Q is a random unitary matrix, $\Lambda = diag\{1, \ldots, 10\}$. We needed the companion matrix C of the matrix A as a input data for the minimization. We can compute the last column of C numerically or using numbers $1, \ldots, 10$ as the eigenvalues. These two approaches suppose to be equivalent. Nevertheless, when the random unitary matrix Q is generated in double precision, than companion

Table 6.2: Results of the minimization of the departure from normality of initially normal matrix A_{Normal}

Starting point of the minimization process	The measure of the departure from normality	Initial departure from normality	Minimum
$R_{initial}$	Absolute	$4.48 \cdot 10^{-7}$	$8.10 \cdot 10^{-8}$
$R_{initial}$	Relative	$1.16 \cdot 10^{-9}$	$1.69 \cdot 10^{-10}$
R_{oShift}	Absolute	$2.26 \cdot 10^{-1}$	$166 \cdot 10^{-4}$
R_{oShift}	Relative	$5.89 \cdot 10^{-4}$	$4.31 \cdot 10^{-7}$
Identity	Absolute	$4.23 \cdot 10^{16}$	$6.59 \cdot 10^{5}$
Identity	Relative	1.41421	1.00

matrix computed using numbers $1, \ldots, 10$ approximates companion matrix of the matrix A_{Normal} poorly. Consequently, the approximation $A_{Normalaps}$ of the matrix A_{Normal} from the factors of the APS decomposition computed using 128 digits precision and computing C using numbers $1, \ldots, 10$ as eigenvalues approximate the matrix A_{Normal} poorly. The norm of the error of the approximation was in this case $\|A_{Normal} - A_{Normalaps}\| = 1.75 \cdot 10^{-12}$. On the other hand, the error of the approximation computed using numerically computed last column of C was very small $\|A_{Normal} - A_{Normalaps}\| = 1.75 \cdot 10^{-127}$. The reason of such a behavior was, that matrix Q generated in double precision was not enough orthonormal. The situation was different when Q was generated using 128 precision. Numbers $1, \ldots, 10$ approximated eigenvalues of the matrix A almost precisely in this case. The norm of the error computed using numbers $1, \ldots, 10$ as the eigenvalues was $\|A_{Normal} - A_{Normalaps}\| = 4.28 \cdot 10^{-127}$, which is even smaller than norm of the error $\|A_{Normal} - A_{Normalaps}\| = 1.29 \cdot 10^{-126}$ computed using numerically computed last column of C.

6.3.2 Minimization of the departure from normality of matrices motivated by discretization of the convection-diffusion problem

As we pointed out earlier, our original goal was to minimize the departure from normality for some large matrix arisen by discretization of some real world problem, but we were not able to compute input data for minimization sufficiently accurate for large matrices. We showed that we were not able to find exact global minimum of the functions (6.2) and (6.3) via R even for matrices of small dimensionality in previous section. We were able to decrease the departure from normality significantly though. Because of these complications, we decided to minimize the departure from normality of a small dimensional matrix $A_{conv-diff} \in \mathbb{F}^{16\times16}$ that is motivated by discretization of convection-diffusion problem. The parameter ν is the scalar diffusion parameter of the problem. We already mentioned this matrix in section 6.2.

The tables 6.3 and 6.4 summarize the results of the minimization of the departure from normality of the matrix $A_{conv-diff}$ with two different parameters

$\nu = 10^{-2}$ and $\nu = 10^{-3}$. We used two starting points of the minimization, the matrix $R_{initial}$, which is submatrix of the matrix Y computed by APS decomposition of the matrix $A_{conv-diff}$ and the identity.

Table 6.3: Results of the minimization of the departure from normality of the matrix $A_{conv-diff}$ motivated by discretization of convection-diffusion problem, $\nu = 10^{-2}$

Starting point of the minimization process	The measure of the departure from normality	Initial departure from normality	Minimum
$R_{initial}$	Absolute	1.72	0.18
$R_{initial}$	Relative	$9.77 \cdot 10^{-1}$	$2.69 \cdot 10^{-1}$
Identity	Absolute	$1.16 \cdot 10^{25}$	$4.93 \cdot 10^{16}$
Identity	Relative	1.41421	1.41421

Table 6.4: Results of the minimization of the departure from normality of the matrix $A_{conv-diff}$ motivated by discretization of convection-diffusion problem, $\nu = 10^{-3}$

Starting point of the minimization process	The measure of the departure from normality	Initial departure from normality	Minimum
$R_{initial}$	Absolute	$1.52 \cdot 10^{3}$	$4.53 \cdot 10^{1}$
$R_{initial}$	Relative	1.40	$9.99 \cdot 10^{-1}$
Identity	Absolute	$1.96 \cdot 10^{27}$	$2.96 \cdot 10^{17}$
Identity	Relative	1.41421	1.41421

The column "Initial departure from normality" of the tables 6.3 and 6.4 contains the departure from normality of the matrices YCY^{-1} and $Y_ICY_I^{-1}$, where Y and C corresponds to the factors of the APS decomposition of the matrix $A_{conv-diff}$ and Y_I is identical with Y except the submatrix R of Y is replaced by identity. As we mentioned earlier Y and C were computed using 128 precision and after computation saved with the double precision. Matrix Y^{-1} is the inverse of Y computed using the double precision. In other words, column "Initial departure from normality" contains the value of the departure from normality of the matrices YCY^{-1} and $Y_ICY_I^{-1}$, where all the factors are saved with double precision and consequently the departure from normality of these approximations can be different from the departure from normality of the approximated matrix $A_{conv-diff}$. The column "Initial departure from normality" of the tables presented in the subsections 6.3.3 and 6.3.4 contains analogously computed values of the departure from normality as column "Initial departure from normality" of the tables 6.3 and 6.4 in this section, i. e. the column "Initial departure from normality" of the tables presented in the subsections 6.3.3 and 6.3.4 contains the departure from normality of the matrices YCY^{-1} and $Y_ICY_I^{-1}$ corresponding to the matrices A_N, A_S,

A_{def1} and A_{def2} which are described in subsections 6.3.3 and 6.3.4. The absolute departure from normality of the matrix $A_{conv-diff}$, $\nu = 10^{-2}$ is $6.18 \cdot 10^{-1}$. The absolute departure from normality of the matrix $W_{128}Y_{128}C_{128}Y_{128}^{-1}W_{128}^{*}$, where all the factors are factors of the APS decomposition of the matrix $A_{conv-diff}$, $\nu = 10^{-2}$ computed and saved with 128 precision is $6.18 \cdot 10^{-1}$ also. As you can see, it is smaller than the result of the minimization of the absolute departure from normality of the matrix $A_{conv-diff}$, $\nu = 10^{-2}$. The situation is analogous for the parameter $\nu = 10^{-3}$. In this case the absolute departure from normality is $6.30 \cdot 10^{-2}$, but the absolute departure from normality of the approximation introduced in table 6.4 is $1.52 \cdot 10^{3}$ and the result of the minimization is the matrix with absolute departure from normality $4.53 \cdot 10^{1}$. This observation illustrates the sensitivity of the departure from normality to even slight change of the factors of the APS decomposition. The minimization program using double precision was not able to find matrix with smaller departure from normality that had precisely computed decomposition of the matrix $A_{conv-diff}$.

We were disappointed about the minimization of relative departure from normality with identity as the initial point. We were not able to achieve any progress.

6.3.3 Minimization of the departure from normality of some matrices used as a numericall example in paper How descriptive are GMRES convergence bounds

Chapter 3 describes some well known GMRES convergence bounds. Some of these bounds are based on the spectrum, field of values or ε-pseudospectrum of the system matrix A. Mark Embree introduced a few numerical examples of the matrices for which these bounds success or fail to predict convergence in his paper "How descriptive are GMRES convergence bounds" [22]. Let

$$A_N = \begin{pmatrix} 1 & \alpha & & & \\ & 1 & & & \\ & & c & & \\ & & & \ddots & \\ & & & & d \end{pmatrix}, \tag{6.9}$$

where $\alpha \gg 1$, $0 < c < d$ and numbers on the diagonal between c and d are uniformly distributed in the interval $[c, d]$. Matrix A_N is an example of the matrix where all the bounds based on the spectrum, field of values and ε-pseudospectrum fail together. This matrix is introduced in the paper [22] as an example B. We expect that the set $\mathcal{S}_2(f, \lambda_1, \ldots, \lambda_N)$ containing the matrix A_N with an arbitrary right hand side b does not contain any matrix close to normal, because the convergence bound based on the eigenvalues fails to predict the convergence of GMRES(A_N,b). We did the following numerical experiment to examine this assumption. We generated the random right hand side b and then we attempted to find the matrix with the same spectrum as the matrix A_N and for the right hand side b the same GMRES convergence curve and has the minimal departure from normality. We used two starting points of the minimization, the matrix $R_{initial}$, which is submatrix of the matrix Y computed by APS decomposition of the matrix A_N and the identity. We chose $A \in \mathbb{F}^{10\times 10}$, $\alpha = 10$ and $c = 2$, $d = 5.5$.

The results are summarized in the table 6.5. The result of the minimization was a matrix with relative departure from normality $2.56 \cdot 10^{-1}$. The discussion whether these results support our hypothesis is in subsection 6.3.4.

Table 6.5: Results of the minimization of the departure from normality of the matrix A_N

Starting point of the minimization process	The measure of the departure from normality	Initial departure from normality	Minimum
$R_{initial}$	Absolute	$1.41 \cdot 10^{2}$	$8.87 \cdot 10^{1}$
$R_{initial}$	Relative	$6.29 \cdot 10^{-1}$	$2.56 \cdot 10^{-1}$
Identity	Absolute	$276 \cdot 10^{13}$	$6.06 \cdot 10^{4}$
Identity	Relative	1.41421	$4.15 \cdot 10^{-1}$

Let

$$A_S = \begin{pmatrix} 1 & \delta & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & \ddots & \delta \\ & & & & 1 \end{pmatrix}, \tag{6.10}$$

where $0 < \delta \ll 1$. The matrix A_S is used in the paper [22] as an example D, which is an example of the matrix for which the bound based on the eigenvalues of the system matrix fails to predict convergence, but the bounds based on ε-pseudospektrum and field of values predict convergence successfully.

We once again expected that $\mathcal{S}_2(f, \lambda_1, \ldots, \lambda_2)$ corresponding to the matrix A_S and arbitrary right hand side b does not contain any matrix that is close to normal matrix, because bound based on the eigenvalues does not predict GMRES(A_S,b) convergence. We again attempted to find a matrix which has the same spectrum as the matrix A_S and has the same GMRES convergence curve for given right hand side b and has the minimal departure from normality. We chose $A_S \in \mathbb{F}^{10\times 10}$, $\delta = 1/2$. We did this numerical experiment using two right hand sides, random vector b ($\|b\| = 1$) and the vector e_N which the last column of the identity $I \in \mathbb{F}^{10\times 10}$. It is proven, that the GMRES applied to the A_S and e_1 is nearby the worst case GMRES, see [22]. The results for both right hand sides are summarized in the tables 6.6 and 6.7. The discussion whether these results support our expectation is in subsection 6.3.4.

Please notice that the columns "Initial departure from normality" (see subsection 6.3.2) of the tables 6.6 and 6.7 presents different values of the initial departure from normality, even though these columns suppose to be identical. The absolute departure from normality of the matrix A_S is $3.54 \cdot 10^{-1}$ which is the same as the absolute departure from normality of the matrix YCY^{-1} presented in the table 6.6. The relative departure from normality of the matrix A_S is also the same as the departure from normality of the matrix YCY^{-1} presented in the table 6.6. The value $1.95 \cdot 10^{5}$ of the absolute departure from normality of

the matrix YCY^{-1} presented in the table 6.7 is not correct as well as the value of relative departure from normality. This shows that using the random vector b as the right hand side in this case caused that the matrix YCY^{-1} has different departure from normality than the matrix A_S. This observation illustrates the sensitivity of the problem.

Table 6.6: Results of the minimization of the departure from normality of the matrix A_S, $b = e_n$

Starting point of the minimization process	The measure of the departure from normality	Initial departure from normality	Minimum
$R_{initial}$	Absolute	$3.54 \cdot 10^{-1}$	$3.10 \cdot 10^{-1}$
$R_{initial}$	Relative	$2.89 \cdot 10^{-2}$	$2.48 \cdot 10^{-2}$
Identity	Absolute	$1.82 \cdot 10^{6}$	$1.02 \cdot 10^{2}$
Identity	Relative	1.32	$2.77 \cdot 10^{-1}$

Table 6.7: Results of the minimization of the departure from normality of the matrix A_S, b, is random vector

Starting point of the minimization process	The measure of the departure from normality	Initial departure from normality	Minimum
$R_{initial}$	Absolute	$1.95 \cdot 10^{5}$	$7.06 \cdot 10^{3}$
$R_{initial}$	Relative	1.4141	1.4094
Identity	Absolute	$2.27 \cdot 10^{31}$	$6.81 \cdot 10^{21}$
Identity	Relative	1.41421	1.41421

6.3.4 Minimization of the departure from normality of matrices used as a numericall example of restarted GMRES preconditiond by deflation

The deflation is a preconditioning technique for restarted GMRES. Preconditioning restarted GMRES by deflation aims to accelerate convergence by manipulation with eigenvalues of the system matrix. Paper [23] describes among others examples of the matrices and the right hand sides for which preconditioning by deflation significantly accelerates convergence of restarted GMRES. We expect that sets $\mathcal{S}_2(f, \lambda_1, \ldots, \lambda_2)$ corresponding to these matrices and right hand sides contain nearly normal matrices, because deflation accelerates the convergence by manipulation with eigenvalues. We attempted to find matrices included in sets $\mathcal{S}_2(f, \lambda_1, \ldots, \lambda_2)$ corresponding to the matrices A_{def1} and A_{def2} and the right hand side $b = (1, \ldots, 1)^T$ used as an example in [23] with minimal departure from normality. These matrices are defined as $A_{def1} = S_1 D S_1^{-1}$ and $A_{def2} = S_2 D S_2^{-1}$,

where $D = \mathrm{diag}\{1, 2, \ldots, N\}$ and

$$S_1 = \begin{pmatrix} 1 & 0.9 & & \\ & \ddots & \ddots & \\ & & \ddots & 0.9 \\ & & & 1 \end{pmatrix}, \quad S_2 = \begin{pmatrix} 1 & 1.1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1.1 \\ & & & 1 \end{pmatrix}.$$

The dimension of the matrices A_{def1} and A_{def2} used in [23] is 100×100. We decided to use for our purposes matrices $A_{def1}, A_{def2} \in \mathbb{F}^{10\times 10}$, because we were not able to effectively work with large matrices. Results of the minimization are summarized in tables 6.8 and 6.9.

Table 6.8: Results of the minimization of the departure from normality of the matrix A_{def1}

Starting point of the minimization process	The measure of the departure from normality	Initial departure from normality	Minimum
$R_{initial}$	Absolute	$2.73 \cdot 10^{1}$	7.37
$R_{initial}$	Relative	$6.69 \cdot 10^{-2}$	$1.87 \cdot 10^{-2}$
Identity	Absolute	$4.00 \cdot 10^{16}$	$9.31 \cdot 10^{4}$
Identity	Relative	1.41421	1.000000000

Table 6.9: Results of the minimization of the departure from normality of the matrix A_{def2}

Starting point of the minimization process	The measure of the departure from normality	Initial departure from normality	Minimum
$R_{initial}$	Absolute	$1.34 \cdot 10^{2}$	$1.04 \cdot 10^{1}$
$R_{initial}$	Relative	$2.77 \cdot 10^{-1}$	$2.64 \cdot 10^{-2}$
Identity	Absolute	$9.74 \cdot 10^{16}$	$3.68 \cdot 10^{4}$
Identity	Relative	1.41421	0.9999999998

Let us compare these results with the results from previous subsection. As we explained in previous subsection, we expected that the sets $\mathcal{S}_2(f, \lambda_1, \ldots, \lambda_2)$ corresponding to the matrices A_N and A_S and an arbitrary right hand sides b does not contain any nearly normal matrix with the right hand side b, because bound based on the eigenvalues does not predict GMRES(A_N,b) or GMRES(A_S,b) convergence. The result of the minimization of the relative departure from normality of the matrix A_N using random right hand side b is matrix which relative departure from normality is $2.56 \cdot 10^{-1}$ (see table 6.5). The result of the minimization of the relative departure from normality of the matrix A_S using right hand side e_N is matrix which relative departure from normality is $2.48 \cdot 10^{-2}$ (see table 6.6). We expected that the minimums of the relative departure from normality

corresponding to the matrices A_{def1}, A_{def2} and right hand side $b = (1, \ldots, 1)^T$ are much smaller than the minimums corresponding to the matrices A_N and A_S, because the preconditioning by deflation strongly affects the GMRES convergence of A_{def1}, A_{def2} by manipulation with the eigenvalues and consequently we expect that sets $\mathcal{S}_2(f, \lambda_1, \ldots, \lambda_2)$ corresponding to the matrices A_N and A_S contains nearly normal matrices. The minimum of the relative departure from normality that we found corresponding to the matrix A_{def1} resp. A_{def2} is $1.87 \cdot 10^{-2}$ resp. $2.64 \cdot 10^{-2}$; see tables 6.8 and 6.9. Comparing these results, we can see that minimum corresponding to the matrix A_N $2.56 \cdot 10^{-1}$ is approximately 10 times bigger than minimums corresponding to the matrices A_{def1} and A_{def2}. Minimum corresponding to the matrix A_S is approximately the same as the minimums corresponding to the matrices A_{def1} and A_{def2}. We were slightly disappointed with the closeness of these minimums. On the other hand, these results do not prove, that our expectations were not reasonable, because the minimization process was not able to find in general global minimums of the functions (6.2) and (6.3), as we showed in subsections 6.3.1 and 6.3.2.

Chapter 7

The norm of the error

The convergence of GMRES is often expressed in the terms of residuals. Another important characteristic of convergence is the norm of the error. In section 7.1 we present an unpublished result of Gérard Meurant [27], [28], [29] a relation (7.38), which is a formula for the norm of the GMRES error in step k. This formula shows the dependance of the norm of the k-th GMRES error on eigenvalues and convergence curve. We would like to recall, that by a convergence curve we mean a sequence of the norms of the GMRES residuals. Further we would like to introduce our upper estimate (7.45) of the norm of the k-th error of GMRES (section 7.2). We are also interested in the characterization of the eigenvalue distributions of matrices that give a small error norm in the k-th step (section 7.3). Throughout this chapter we use the same notation as in the previous chapters and we assume that the initial GMRES approximation of the solution of the problem of the form (1.1) is $x_0 = 0$.

7.1 The dependence of the k-th norm of the error on eigenvalues and convergence curve of $GMRES(A, b)$

Let K be the matrix with column $b, Ab, \ldots, A^{N-1}b$. Since characteristic polynomial annihilates matrix A, the formula

$$B = AK = KC \tag{7.1}$$

holds. Putting this together with $B = WR$ and $Y = RC^{-1}$ (subsection 4.3) gives

$$K = WY. \tag{7.2}$$

Let $K = VU$ be the QR factorization of matrix K. Using this QR factorization and the relation (7.2) yields

$$K^*K = Y^*W^*WY = Y^*Y = U^*V^*VU = U^*U \tag{7.3}$$

and

$$Y = W^*VU. \tag{7.4}$$

Since $Y^*Y = U^*U$ and U^* is lower triangular, then U^*U is the Cholesky factorization of Y^*Y. Let us recall the block decomposition of the matrix Y

$$Y = \begin{pmatrix} h & R \\ & 0 \end{pmatrix} \tag{7.5}$$

and denote the first $N-1$ components of vector h as $\hat{h} = (\eta_1, \ldots, \eta_{N-1})^T$.

$$Y = \begin{pmatrix} \hat{h} & R \\ \eta_N & 0 \end{pmatrix} \text{ and } Y^*Y = \begin{pmatrix} \|h\|^2 & \hat{h}^*R \\ R^*\hat{h} & R^*R \end{pmatrix} \tag{7.6}$$

Let 0_i be the i dimensional zero column vector. Further denoting the Cholesky factor

$$L = U^* = \begin{pmatrix} l_1 & 0_{N-1}^T \\ l & \hat{L} \end{pmatrix}, \tag{7.7}$$

Expressions (7.3), (7.6) and (7.7) yield

$$l_1 = \|h\|, \quad l = \frac{R^*\hat{h}}{\|h\|}, \tag{7.8}$$

$$\hat{L}\hat{L}^* = R^*R - \frac{1}{\|h\|^2} R^*\hat{h}\hat{h}^*R. \tag{7.9}$$

Moreover

$$U^{-1} = \begin{pmatrix} \frac{1}{l_1} & -\frac{l^*\hat{L}^{-*}}{l_1} \\ 0_{N-1} & \hat{L}^{-*} \end{pmatrix}. \tag{7.10}$$

The Arnoldi algorithm (see section 2.1) applied to the matrix A gives $H = V^*AV$. Using the decompositions $H = V^*AV$, $A = WYCY^{-1}W^*$ (see (4.24)) and the relation (7.4) yields

$$H = UCU^{-1}. \tag{7.11}$$

Let us partition the matrix $CU^{-1} \in \mathbb{F}^{N\times N}$ as

$$CU^{-1} = \begin{pmatrix} 0_{N-1}^T & \beta_0 \\ \tilde{L}^{-*} & \beta \end{pmatrix}, \tag{7.12}$$

where $\tilde{L}^{-*} \in \mathbb{F}^{(N-1)\times(N-1)}$ is a lower triangular matrix and $\beta \in \mathbb{F}^{N-1}$. The straightforward multiplication gives

$$H = \begin{pmatrix} l^*\tilde{L}^{-*} & \beta_0 l_1 + l^*\beta \\ \hat{L}^*\tilde{L}^{-*} & \hat{L}^*\beta \end{pmatrix}. \tag{7.13}$$

It is necessary to introduce some more notation. Let $h_{i,j}$ be the entry of the matrix H on the position $(.)_{i,j}$. Let

$$H_k = \begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & \cdots & h_{1,k} \\ h_{2,1} & h_{2,2} & & & \vdots \\ & h_{3,2} & \ddots & & \vdots \\ & & \ddots & \ddots & \vdots \\ & & & h_{k,k-1} & h_{k,k} \end{pmatrix}, \tag{7.14}$$

$H_k^{(e)}$ be the extended matrix

$$H_k^{(e)} = \begin{pmatrix} H_k \\ h_{k+1,k} e_k^T \end{pmatrix}. \tag{7.15}$$

and let $\tilde{H}_{k-1}$ and (h^{k-1}) be the submatrices defined by

$$\tilde{H}_{k-1} = \begin{pmatrix} h_{2,1} & h_{2,2} & \cdots & \cdots & h_{2,k-1} \\ & h_{3,2} & & & \vdots \\ & & \ddots & & \vdots \\ & & & h_{k-1,k-2} & h_{k-1,k-1} \\ & & & & h_{k,k-1} \end{pmatrix}, \tag{7.16}$$

$$(h^{k-1})^* = (h_{1,1}\ h_{1,2}\ \cdots\ h_{1,k-1}). \tag{7.17}$$

Let the matrices $\hat{L}_k \in \mathbb{F}^{k\times k}$, $\tilde{L}_k \in \mathbb{F}^{k\times k}$ and $R_k \in \mathbb{F}^{k\times k}$ be the matrices obtained by deleting the last $N-k$ columns and the last $N-k$ rows of the matrices $\hat{L} \in \mathbb{F}^{N\times N}$, $\tilde{L} \in \mathbb{F}^{N\times N}$ and $R \in \mathbb{F}^{N\times N}$ respectively. Let the vectors l_k and $\hat{h}_k$ be the vectors of the first k components of vectors l and $\hat{h}$ respectively.

Based on previous definitions and formulas (7.8) and (7.13), the vector h^k and the matrix $\tilde{H}_k$ can be written in the form

$$(h^k)^* = l_k^* \tilde{L}_k^{-*}, \quad h^k = \tilde{L}_k^{-1} \frac{R_k^* \hat{h}_k}{\|h\|} \tag{7.18}$$

and

$$\tilde{H}_k = \hat{L}_k^* \tilde{L}_k^{-*}. \tag{7.19}$$

Furthermore let V_k be the matrix whose columns are the first k columns of matrix V. The k-th iterate of GMRES x_k is the linear combination of these columns, i. e. x_k can be written in the form

$$x_k = V_k z_k. \tag{7.20}$$

The vector z_k is obtained by solving the least squares problem

$$z_k = \operatorname{argmin}\{\|z_k\| : \|b - Az_k\| = \min_{y \in x_0 + \mathcal{K}_k} \|b - Ay\|\}. \tag{7.21}$$

It is well known, that solution of such a problem can be expressed using the normal equations, i. e.

$$(H_k^{(e)})^* H_k^{(e)} z^k = \|r^0\| (H_k^{(e)})^* e^1. \tag{7.22}$$

The problem is not solved numerically this way, nevertheless to express z_k in terms of normal equations is advantageous for analytic formulation of the k-th error, as a reader may see below. Using (7.16) and (7.17), the extended matrix can be written in the form

$$H_k^{(e)} = \begin{pmatrix} (h^k)^* \\ \tilde{H}_k \end{pmatrix}. \tag{7.23}$$

Formulas (7.22) and (7.23) gives

$$z^k = \|r^0\|(\tilde{H}_k^*\tilde{H}_k + h^k(h^k)^*)^{-1}h^k. \tag{7.24}$$

The Sherman-Morrison formula applied to $(\tilde{H}_k^*\tilde{H}_k + h^k(h^k)^*)^{-1}$ gives

$$(\tilde{H}_k^*\tilde{H}_k + h^k(h^k)^*)^{-1} = \tilde{H}_k^{-1}\tilde{H}_k^{-*} + \frac{\|\tilde{H}_k^{-*}h^k\|^2}{1+\|\tilde{H}_k^{-*}h^k\|^2}\tilde{H}_k^{-1}\tilde{H}_k^{-*} \tag{7.25}$$

and consequently

$$z^k = \|r_0\|\frac{1}{1+\|\tilde{H}_k^{-*}h^k\|^2}\tilde{H}_k^{-1}\tilde{H}_k^{-*}h^k. \tag{7.26}$$

Using the result from [4]

$$\|r^k\|^2 = \frac{\|r_0\|}{1+\|\tilde{H}_k^{-1}h^k\|^2} \tag{7.27}$$

as well as equalities (7.18), (7.19) and obvious equality

$$\|r^k\|^2 = \|h\|^2 - \|\hat{h}_k\|^2 \tag{7.28}$$

gives together with (7.26)

$$z^k = \frac{\|h\|^2 - \|\hat{h}^k\|^2}{\|h\|^2}\tilde{L}_k^*(\hat{L}_k^{-*}\hat{L}_k^{-1})R_k^*\hat{h}^k. \tag{7.29}$$

Further using the previous notations and formula (7.9)

$$\hat{L}_k\hat{L}_k^* = (\hat{L}\hat{L}^*)_k = R_k^*R_k - \frac{1}{\|h\|^2}R_k^*\hat{h}_k\hat{h}_k^*R_k, \tag{7.30}$$

where $(\hat{L}\hat{L}^*)_k \in \mathbb{F}^{k\times k}$ is obtained by deleting the last $N-k$ columns and the last $N-k$ rows of the matrix $(\hat{L}\hat{L}^*) \in \mathbb{F}^{N\times N}$ and the inverse the matrix $\hat{L}_k\hat{L}_k^*$ is

$$\hat{L}_k^{-*}\hat{L}_k^{-1} = R_k^{-1}\left(I - \frac{1}{\|h\|^2}\hat{h}_k\hat{h}_k^*\right)^{-1}R_k^{-*}. \tag{7.31}$$

Applying the Sherman-Morrison formula, one can obtain

$$\hat{L}_k^{-*}\hat{L}_k^{-1}R_k^*\hat{h}_k = R_k^{-1}\left(I + \frac{1}{\|h\|^2 - \|\hat{h}_k\|}\hat{h}_k\hat{h}_k^*\right)\hat{h}_k \tag{7.32}$$

$$= \frac{\|h\|^2}{\|h\|^2 - \|\hat{h}_k\|^2}R_k^{-1}\hat{h}_k \tag{7.33}$$

and therefore

$$z^k = \tilde{L}_k^* R_k^{-1} \hat{h}_k \tag{7.34}$$

and consequently

$$x^k = WYU^{-1} \begin{pmatrix} \tilde{L}_k^* R_k^{-1} \hat{h}_k \\ 0_{N-k} \end{pmatrix}. \tag{7.35}$$

The expression

$$x^k = WYC^{-1} \begin{pmatrix} 0 \\ \tilde{L}_k^* R_k^{-1} \hat{h}_k \\ 0_{N-(k+1)} \end{pmatrix} \tag{7.36}$$

is obtained by noticing that $YU^{-1} = YC^{-1}CU^{-1}$ and

$$CU^{-1} \begin{pmatrix} \tilde{L}_k^* R_k^{-1} \hat{h}_k \\ 0_{N-k} \end{pmatrix} = \begin{pmatrix} 0 \\ \tilde{L}_k^* R_k^{-1} \hat{h}_k \\ 0_{N-(k+1)} \end{pmatrix}. \tag{7.37}$$

Finally the formula for the norm of the k-th error of GMRES is

$$\|\epsilon^k\| = \left\| YC^{-1} \begin{pmatrix} 1 \\ -R_k^{-1} \hat{h}_k \\ 0_{N-k+1} \end{pmatrix} \right\|, \tag{7.38}$$

because the exact solution can be expressed in following way

$$x = A^{-1}b = (WYCY^{-1}W^*)^{-1}b = WYC^{-1}Y^{-1}W^*b = WYC^{-1}e^1, \tag{7.39}$$

where e^1 is the first column of $N \times N$ dimensional identity matrix.

7.2 The minimization of an upper bound of the norm of the error of GMRES in step k

In this section, we use result (7.38) from previous section to derive an upper bound (7.45) for the norm of k-th GMRES error and minimize this bound. Notice that $x = WYC^{-1}e^1$ is the exact solution of the linear system (see (7.39)). If we change the eigenvalues and thus C, we obtain another matrix (but we do not change the right hand side) and we do not change the residual convergence curve as well as the $(n-1)$-st iterates. However, this changes the exact solution x and, of course, the error.

The question is: How to choose spectrum to make $\|\epsilon^k\|$ as small as possible? We were not able to minimize the norm of the error itself, but we were able to minimize the upper estimate (7.45). We obtained our estimate (7.45) in the following way:

At first, it is suitable to write formula (7.38) in the form

$$\|\epsilon^k\| = \left\| YC^{-1}e^1 - YC^{-1} \begin{pmatrix} 0 \\ R_k^{-1} \hat{h}_k \\ 0_{N-k+1} \end{pmatrix} \right\|. \tag{7.40}$$

We consider characteristic polynomial to be of the form $q(z) = z^N + \sum_{i=0}^{N-1} \alpha_i z^{i1}$, so the inverse of companion matrix is of the form

$$C^{-1} = \begin{pmatrix} -\frac{\alpha_1}{\alpha_0} & 1 & 0 & \cdots & 0 \\ -\frac{\alpha_2}{\alpha_0} & 0 & 1 & & . \\ \vdots & . & & \ddots & . \\ -\frac{\alpha_{N-1}}{\alpha_0} & . & & & 1 \\ -\frac{1}{\alpha_0} & 0 & . & . & 0 \end{pmatrix} \tag{7.41}$$

and consequently

$$\|\epsilon^k\| = \left\| Y \begin{pmatrix} -\frac{\alpha_1}{\alpha_0} \\ -\frac{\alpha_2}{\alpha_0} \\ \vdots \\ -\frac{1}{\alpha_0} \end{pmatrix} - Y \begin{pmatrix} R_k^{-1}\hat{h}_k \\ 0_{N-k} \end{pmatrix} \right\|, \tag{7.42}$$

$$\|\epsilon^k\| = \left\| Y \left(\begin{pmatrix} -\frac{\alpha_1}{\alpha_0} \\ -\frac{\alpha_2}{\alpha_0} \\ \vdots \\ -\frac{1}{\alpha_0} \end{pmatrix} - \begin{pmatrix} R_k^{-1}\hat{h}_k \\ 0_{N-k} \end{pmatrix} \right) \right\|. \tag{7.43}$$

Let us denote

$$f_k = \begin{pmatrix} -\frac{\alpha_1}{\alpha_0} \\ -\frac{\alpha_2}{\alpha_0} \\ \vdots \\ -\frac{1}{\alpha_0} \end{pmatrix} - \begin{pmatrix} R_k^{-1}\hat{h}_k \\ 0_{N-k} \end{pmatrix}, \tag{7.44}$$

then we can write

$$\|\epsilon^k\| = \|Y f_k\| \leq \|Y\| \|f_k\|. \tag{7.45}$$

Our goal is to minimize the norm of f_k, i. e. we are trying to minimize elements of f_k (the best would be to set them all to the zero), by modifying the spectrum. Modifying the spectrum is equivalent to modifying the coefficients α_i, $i = 0, \ldots, N-1$. It is easy to zero out the first k elements of f_k. The choice of $\alpha_1, \ldots, \alpha_k$ is determined by the set of equations

$$\begin{aligned} -\frac{\alpha_1}{\alpha_0} &= \left(R_k^{-1}\hat{h}^k \right)_1 \\ &\vdots \\ -\frac{\alpha_k}{\alpha_0} &= \left(R_k^{-1}\hat{h}^k \right)_k . \end{aligned} \tag{7.46}$$

The last element of f_k is $-1/\alpha_0$, so we choose α_0 as big as possible to set $-1/\alpha_0$ as small as possible. We zero out the rest of the elements of f_k by setting the coefficients $\alpha_{k+1}, \ldots, \alpha_{N-1}$ to zero

$$\alpha_i = 0 \quad i = k+1, \ldots, N-1. \tag{7.47}$$

[1] We used different convention of the signs of alphas in this chapter. We considered characteristic polynomial to be of the form $q(z) = z^N - \sum_{i=0}^{N-1} \alpha_i z^i$ in previous chapters.

Consequently, the polynomial that makes the norm of the firs error $\|\epsilon_1\|$ small is of the form $p(\lambda) = \lambda^n + \alpha_1\lambda + \alpha_0$. The polynomial that makes the norm of the second error $\|\epsilon_2\|$ small is of the form for $\|\epsilon_2\|$ is of the form $p(\lambda) = \lambda^n + \alpha_2\lambda^2 + \alpha_1\lambda + \alpha_0$, etc. We will denote polynomials satisfying equation (7.46) and (7.47) for given k as $p_{\epsilon k}$ in the next section.

7.3 Characterization of the spectrum

We attempted to characterize the distribution of the spectrum of the companion matrix C that makes the k-th GMRES error of the matrix $A = WYCY^{-1}W^*$ (see (4.24)) with right hand side b small, i. e. we attempted to characterize the the roots of the polynomials $p_{\epsilon k}$. We were not able to find precise characterization of the spectrum of the companion matrix C that makes the k-th GMRES error of the general matrix A and the right hand side b small. We did the at lest the following numerical experiment that provides some insight on the distribution of such a spectrum.

We computed for the matrix $A_{conv-diff}$ and the right hand side b, which are described in detail in the sections 6.2 and 6.3, the spectrum that makes the k-th GMRES error small and then we attempted to find some characteristic property of this spectrum. Such a spectrum corresponds to the polynomials $p_{\epsilon k}$ as we explained above. We chose the dimension of the matrix to be 9×9 and the scalar parameter $\nu = 10^{-3}$; see section 6.3. This matrix is motivated by the discretization of convection-diffusion problem. The coefficients $\alpha_1, \ldots, \alpha_{N-1}$ of the polynomials $p_{\epsilon k}$ corresponding to the matrix $A_{conv-diff}$ and the right hand side b are determined by the sets of equations (7.46) and (7.47) and the free parameter α_0. As we mentioned in previous section, large α_0 causes small GMRES error. The signs of the coefficients $\alpha_1, \ldots, \alpha_{N-1}$ depend on the sign of the free parameter α_0. We decided to eliminate the influence of the sign of the parameter α_0 by solving the equations

$$\begin{array}{ll} \left|\frac{\alpha_i}{\alpha_0}\right| = |R_k^{-1}\hat{h}_k| & i = 1, \ldots, k \\ \alpha_i = 0 & i = k+1, \ldots, N-1 \end{array} \tag{7.48}$$

instead of the equations (7.46) and (7.47) for parameters $\alpha_0 = \pm 1$ and $\alpha_0 = \pm 1000$. The set of the solutions of the equations (7.46) and (7.47) (i. e. the coefficients of the polynomials $p_{\epsilon k}$) is a subset of the solutions of the equations (7.48). We obtained by solving the equations (7.48) for $k = 1, \ldots, 4$, parameters $\alpha_0 = \pm 1$ and $\alpha_0 = \pm 1000$ and the matrix and the right hand side mentioned above 120 polynomials. We rounded the coefficients of the polynomials to the integer for simplicity of exposition. The sum of the polynomials solving the equations (7.46) and (7.47) is only 16 for parameters $\alpha_0 = \pm 1$ and $\alpha_0 = \pm 1000$ and $k = 1, \ldots, 4$. It is interesting that all 120 polynomials solving (7.48) including those 16 polynomials $p_{\epsilon k}$ solving (7.46) and (7.47) has the same qualitative property that $N - k$ roots are approximately on a circle and the rest is close to zero. We could expect such a result, because of the form of the polynomials $p_{\epsilon k}$. We can consider these polynomials as perturbed polynomials of the form

$$p(\lambda) = \lambda^N + \alpha_0. \tag{7.49}$$

It is well known that polynomials of the form (7.49) have roots placed on a circle, i. e. all roots of such polynomial has absolute value equal to $|\alpha_0|^{1/N}$.

We illustrate these results on several graphs and tables. The equations (7.48) are for $k = 1$ satisfied for example by polynomials $p_1(x) = x^9 + 3x + 1$ and $p_2(x) = x^9 - 3x + 1$. Figure 7.1 resp. 7.2 shows the roots of polynomials p_1 resp. p_2. We can see that in both cases one root is close to zero and the rest is approximately on a circle. The equations (7.48) are for $k = 2$ satisfied for example by polynomials $p_3 = x^9 + 23x^2 + 9x + 1$ and $p_4 = x^9 - 23x^2 + 9x - 1$. Its roots are depicted in the figures 7.3 and 7.4. We observe again, that k roots are close to zero and the rest is approximately on a circle. We chose for $k = 3$ as an example polynomials $p_5 = x^9 + 168x^3 + 94x^2 + 17x + 1$ and $p_6 = x^9 + 168000x^3 + 94000x^2 + 17000x + 1000$. Figures 7.5 and 7.5 show that the change of α_0 from 1 to 1000 does not change qualitatively the eigenvalue distribution at all, except for the absolute value of the eigenvalues. For the sake of completeness we introduce figure 7.7 showing the roots of polynomial $p_7 = x^9 + 1578x^4 + 1065x^3 - 260x^2 + 27x + 1$ satisfying (7.48) for $k = 4$. To make this more transparent, we show the roots and the absolute values of the roots of polynomials $p_1, \ldots, p_7$ in the tables 7.1,. . .,7.4.

In the end of this chapter, we would like to emphasize the fact, that we are able to zero out the first $(N-1)$ elements of f_k by proper choice of the coefficients $\alpha_1, \ldots, \alpha_{N-1}$. On the other hand, we are not able to zero out the last element of f_k. Considering the first $(N-1)$ elements of f_k equal to zero, the norm of the k-th error ϵ_k is equal to the norm of the last column of the matrix Y multiplied by $1/\alpha_0$. In other words, the coefficients $\alpha_1, \ldots, \alpha_{N-1}$ of the polynomial $p_{\epsilon k}$ are generally different in every step k and zero out the firs $(N-1)$ elements of f_k. The minimal error ϵ_k depends in every step k only on the last column of the matrix Y, resp. submatrix R and the choice of the free parameter α_0, i. e. the minimal error is the same in every step k if and only if we don't change the free parameter α_0.

Figure 7.1: The roots of the polynomial p_1

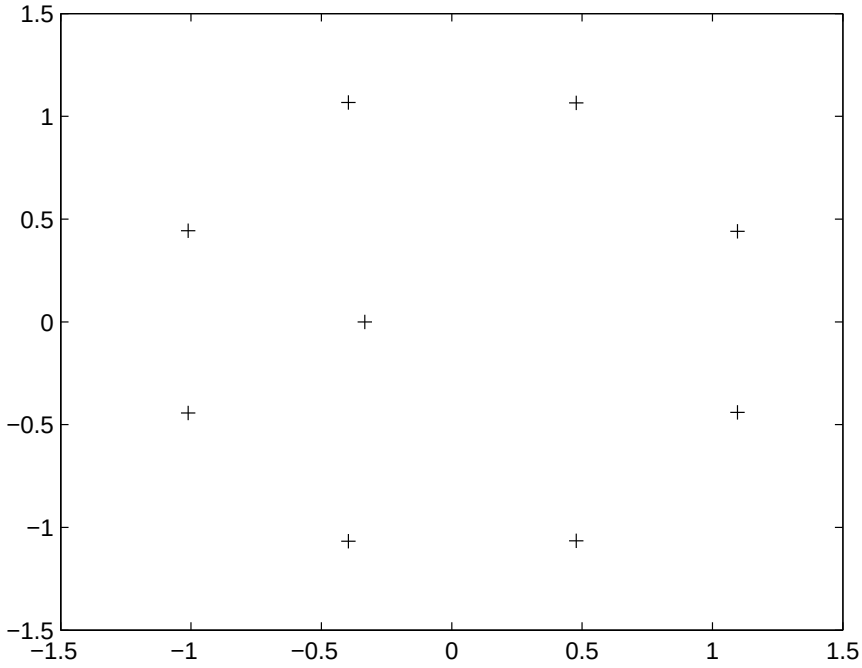


Figure 7.2: The roots of the polynomial p_2

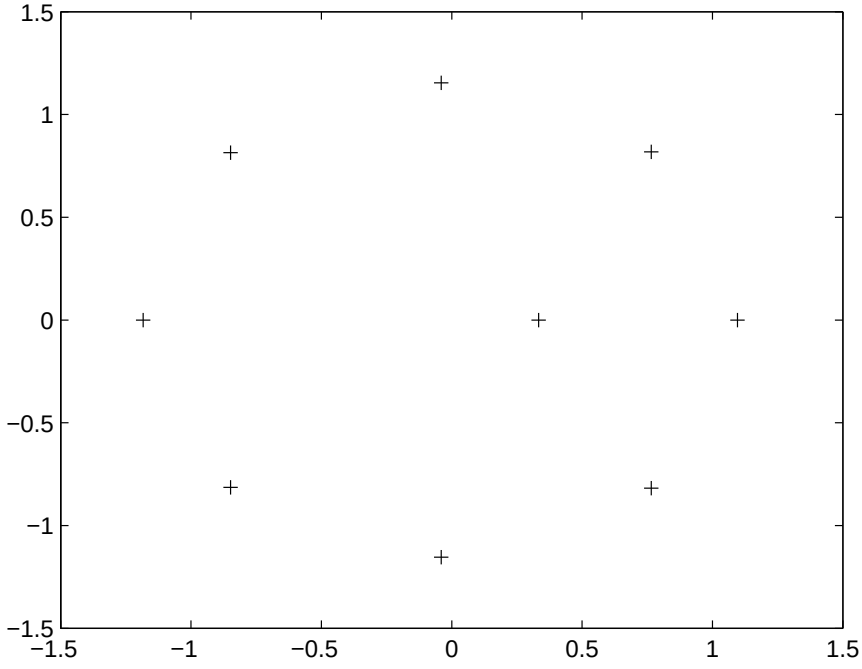


Figure 7.3: The roots of the polynomial p_3

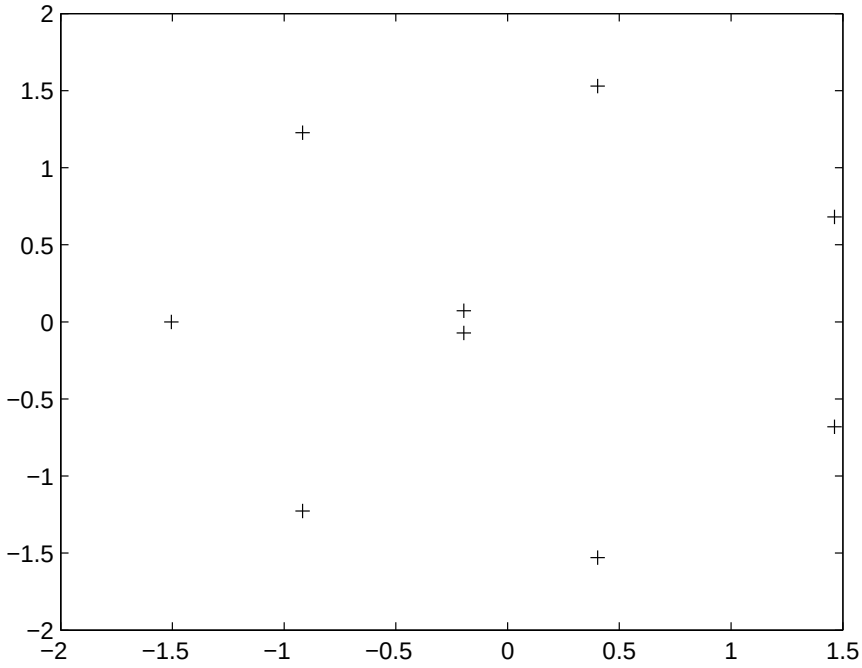


Figure 7.4: The roots of the polynomial p_4

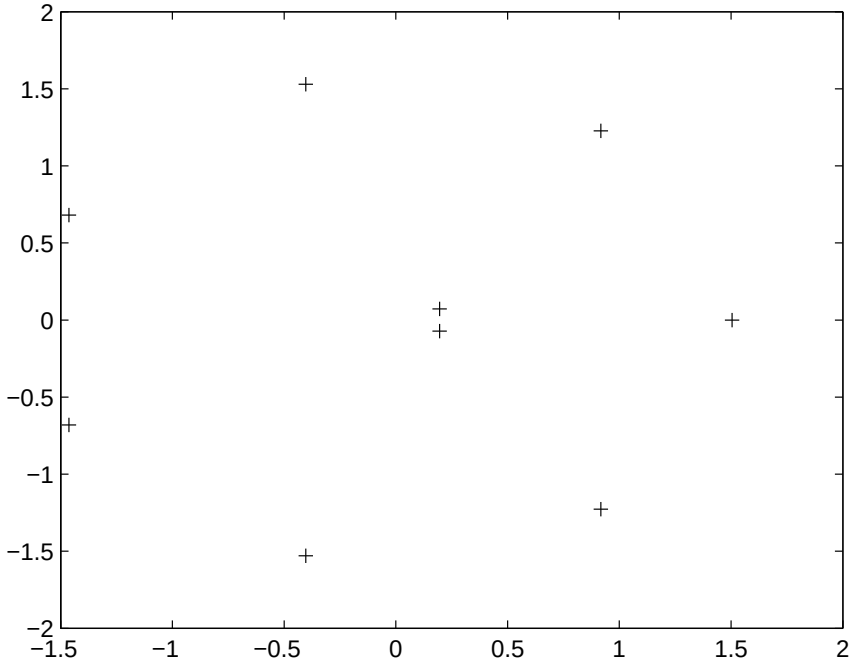


Figure 7.5: The roots of the polynomial p_5

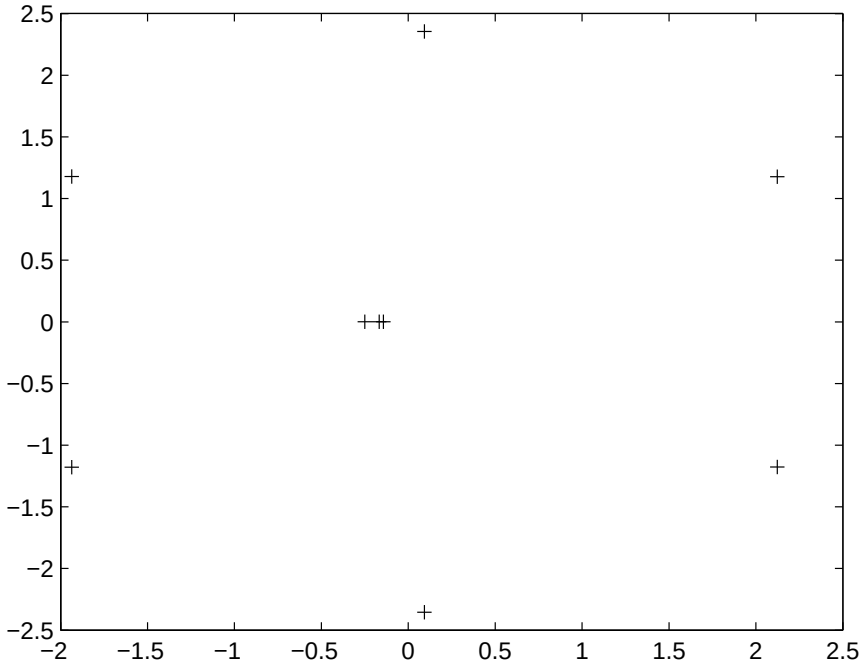


Figure 7.6: The roots of the polynomial p_6

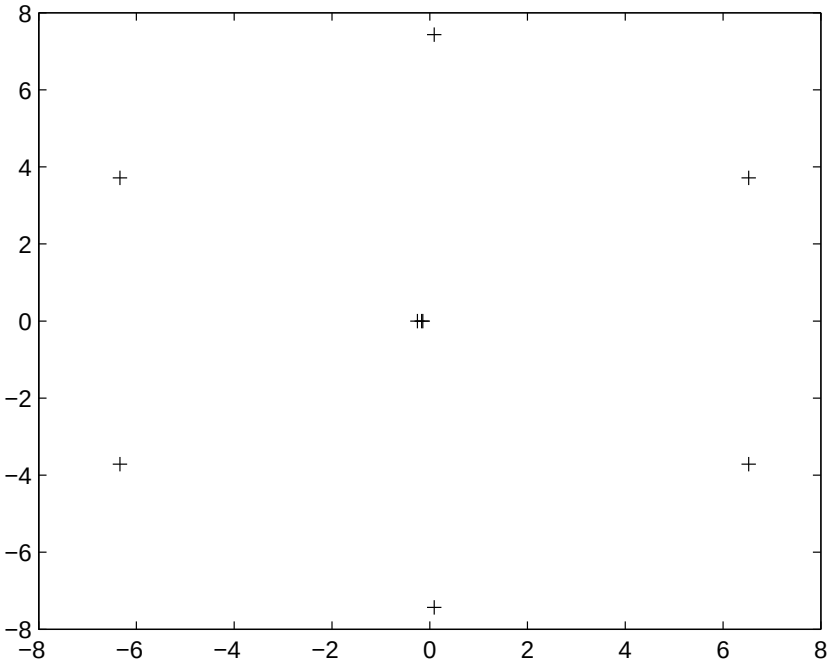


Figure 7.7: The roots of the polynomial p_7

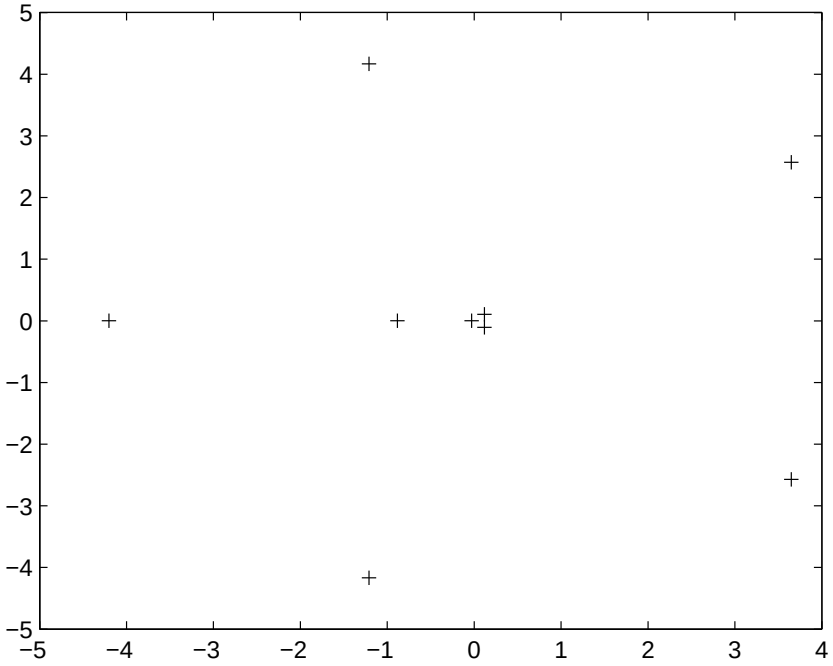


Table 7.1: The roots of the polynomials p_1 and p_2

$p_1 = x^9 + 3x + 1$		$p_2 = x^9 - 3x + 1$	
Roots	Abs. value of the roots	Roots	Abs. value of the roots
$1.0962 + 0.4408i$	1.1815	-1.1833	1.1833
$1.0962 - 0.4408i$	1.1815	$-0.8483 + 0.8147i$	1.1761
$0.4773 + 1.0649i$	1.1670	$-0.8483 - 0.8147i$	1.1761
$0.4773 - 1.0649i$	1.1670	$-0.0401 + 1.1536i$	1.1543
$-0.3963 + 1.0670i$	1.1382	$-0.0401 - 1.1536i$	1.1543
$-0.3963 - 1.0670i$	1.1382	$0.7652 + 0.8179i$	1.1201
$-1.0106 + 0.4434i$	1.1036	$0.7652 - 0.8179i$	1.1201
$-1.0106 - 0.4434i$	1.1036	1.0964	1.0964
-0.3333	0.3333	0.3334	0.3334

Table 7.2: The roots of the polynomials p_3 and p_4

$p_4 = x^9 + 23x^2 + 3x + 1$		$p_4 = x^9 - 23x^2 + 9x - 1$	
Roots	Abs. value of the roots	Roots	Abs. value of the roots
$1.4626 + 0.6805i$	1.6132	$-1.4626 + 0.6805i$	1.6132
$1.4626 - 0.6805i$	1.6132	$-1.4626 - 0.6805i$	1.6132
$0.4029 + 1.5295i$	1.5817	$-0.4029 + 1.5295i$	1.5817
$0.4029 - 1.5295i$	1.5817	$-0.4029 - 1.5295i$	1.5817
$-0.9175 + 1.2272i$	1.5323	$0.9175 + 1.2272i$	1.5323
$-0.9175 - 1.2272i$	1.5323	$0.9175 - 1.2272i$	1.5323
-1.5047	1.5047	1.5047	1.5047
$-0.1957 + 0.0721i$	0.2085	$0.1957 + 0.0721i$	0.2085
$-0.1957 - 0.0721i$	0.2085	$0.1957 - 0.0721i$	0.2085

Table 7.3: The roots of the polynomials p_5 and p_6

$p_5 = x^9 + 168x^3 + 94x^2 + 17x + 1$	
Roots	Abs. value of the roots
$2.1229 + 1.1769i$	2.4273
$2.1229 - 1.1769i$	2.4273
$0.0926 + 2.3547i$	2.3565
$0.0926 - 2.3547i$	2.3565
$-1.9357 + 1.1781i$	2.2660
$-1.9357 - 1.1781i$	2.2660
-0.2500	0.2500
-0.1667	0.1667
-0.1429	0.1429
$p_6 = x^9 + 168000x^3 + 94000x^2 + 17000x + 1000$	
Roots	Abs. value of the roots
$6.5247 + 3.7150i$	7.5082
$6.5247 - 3.7150i$	7.5082
$0.0932 + 7.4300i$	7.4306
$0.0932 - 7.4300i$	7.4306
$-6.3382 + 3.7151i$	7.3467
$-6.3382 - 3.7151i$	7.3467
-0.2500	0.2500
-0.1667	0.1667
-0.1429	0.1429

Table 7.4: The roots of the polynomial p_7

$p_7 = x^9 + 1578x^4 + 1065x^3 - 260x^2 + 27x + 1$	
Roots	Abs. value of the roots
$3.6484 + 2.5725i$	4.4642
$3.6484 - 2.5725i$	4.4642
$-1.2100 + 4.1685i$	4.3406
$-1.2100 - 4.1685i$	4.3406
-4.2016	4.2016
-0.8829	0.8829
$0.1181 + 0.1065i$	0.1590
$0.1181 - 0.1065i$	0.1590
-0.0284	0.0284

Chapter 8

Conclusions

The basic principles of the most popular Krylov subspace methods (CG, MINRES and GMRES) are briefly explained. We have recalled some convergence results for these methods and summarized known characterizations of the matrices and the right hand sides generating the same Krylov residual spaces. The connections and the differences between the spectral decomposition $A = Q\Lambda Q^*$ (see (3.1)) and the decomposition of the form $A = WYCY^{-1}W^*$ (see (4.24)) are shown.

Let $\mathcal{S}_2(f, \{\lambda_1, \ldots, \lambda_N\})$ be the set of all pairs $\{A, b\}$ such that $\sigma(A) = \{\lambda_1, \ldots, \lambda_N\}$ and the sequence of the norms of residuals of GMRES applied to $\{A, b\}$ is f. We expect that if the convergence curve of GMRES applied to the nonnormal matrix A and the right hand side b seems to be determined by the eigenvalues of the matrix A, then the set $\mathcal{S}_2(f, \{\lambda_1, \ldots, \lambda_N\})$ corresponding to A and b contains a nearly normal matrix with the right hand side b. Several numerical experiments are done to examine this assertion. Those experiments showed that the computation of the decomposition of the form (4.24) and minimization of the functions

$$\|A^*A - AA^*\|_F = \|YCY^{-1}Y^{-*}C^*Y^* - Y^{-*}C^*Y^*YCY^{-1}\|_F$$

and

$$\frac{\|A^*A - AA^*\|_F}{\|A\|_F^2} = \frac{\|YCY^{-1}Y^{-*}C^*Y^* - Y^{-*}C^*Y^*YCY^{-1}\|_F}{\|YCY^{-1}\|_F^2}$$

via submatrix R of Y using the data obtained by computing the decomposition of the form (4.24) is an extremely sensitive problem.

An unpublished result of Gérard Meurant, the formula (7.38) for the norm of the k-th error of GMRES

$$\|\epsilon^k\| = \left\| YC^{-1} \begin{pmatrix} 1 \\ -R_k^{-1}\hat{h}_k \\ 0_{N-k+1} \end{pmatrix} \right\|.$$

and its derivation, is described in section 7.1. An upper estimate (7.45) of the k-th error of GMRES is derived. The estimate (7.45) is minimized via the companion matrix C; see section 7.2. The matrix C minimizing the k-th error of GMRES applied to the matrix $A = WYCY^{-1}W^*$ and the right hand side b ensures that the k-th error depends only on the last column of the matrix Y and the absolute term of the characteristic polynomial of the system matrix A. However, the matrix C

minimizing the k-th error is in general different in every step k. A numerical experiment illustrating the distribution of the spectrum of the matrix C minimizing the k-th error of the GMRES applied to the matrix $A = WYCY^{-1}W^*$ and the right hand side b is described in the section 7.3.

Appendix

We present a code of the program `compapshouse` designed by Gérard Meurant in Matlab. `Compapshouse` computes the APS decomposition of the matrix A using right hand side b. Further, we present a code of the program `CompanionMatrixVPA` and `YinverseVPA`. Program `CompanionMatrixVPA` computes the companion matrix C using the spectrum of C and program `YinverseVPA` computes the inverse of the matrix Y; see (4.25). Programs `CompanionMatrixVPA` and `YinverseVPA` were designed by us.

```
function [W,Y,C,h,R]=compapshous(A,b);
%COMPAPSHOUS try to compute the decomposition from
%Arioli, Ptak and Strakos
%in a stable way using Householder transformations and
%not computing the Krylov matrix
%
% Author G. Meurant
% October 2009
%

n=size(A,1);

Ab=A*b;

% Reduce A*b to a multiple of e_1
% This gives R(1,1)
[v,beta,P]=house(Ab);

% Apply P to A
At=P*A*P';

% Reduce At to upper Hessenberg form
[Wt,HH]=hess(full(At));

W=Wt'*P;
W=W';

% h is W' b
h=W'*b;
% the components of h must be real and positive
% this works only if W is real!
G=speye(n,n);
for i=1:n
  if isreal(h(i))
    if h(i) < 0
      G(i,i)=-1;
      h(i)=-h(i);
    end
  else
```

```
    error('h is complex')
  end
end
% W is the unitary matrix we were looking for
W=W*G;
HH=W'*A*W;
HH=triu(HH,-1);

% Computation of R column by column

% We already know R(1,1) (except for the sign)
R=sparse(n-1,n-1);
R1=P*Ab;
R(1,1)=R1(1);
Rkk=R(1,1);
% First column of the inverse of R shifted down
p=[0; 1/R(1,1); zeros(n-3,1)];

for k=2:n-2
 hk=HH(1:n-1,k-1);
 Rk=Rkk*(hk-R*p);
 R(:,k)=Rk;
 % Compute the next column of the inverse of R
 r=R(1:k-1,k);
 Rkk=R(k,k);
 %if abs(Rkk) > 0
 if abs(Rkk) > 1e-16
  Ri=zeros(n-1,1);
  rho=1/Rkk;
  Ri(1:k-1)=-rho*(R(1:k-1,1:k-1)\r);
  Ri(k)=rho;
  % Shift down to obtain the next p
  p=zeros(n-1,1);
  p(2:k+1)=Ri(1:k);
 else
  % R is singular
  error('Compapshous: R is (almost) singular')
 end
end

% Last column
hn=HH(1:n-1,n-2);
Rn=Rkk*(hn-R*p);
R(:,n-1)=Rn;

% check the sign of R(1,1)
bAb=b'*Ab;
if sign(R(1,1)) ~= sign(bAb)
```

```
%if W(1,1) ~= P(1,1)
 R=-R;
end

% the matrix Y is constructed from h and R
Y=[h [R; zeros(1,n-1)]];

% We need to compute the last column of C = inv(Y) HH Y
Cn=Y\(HH*Y(:,n));
% alternative: put h at the end by permuting the columns
% and solve a triangular system
%YY=[[R; zeros(1,n-1)] h];
%y=YY\(HH*Y(:,n));
%Cn=[y(n); y(1:n-1)];

% Construct the companion matrix
C=[zeros(1,n-1) Cn(1); speye(n-1,n-1) Cn(2:end)];

function [v,beta,P]=house(x);
%HOUSE Householder transformation to zero components 2 to n
% from GVL

n=length(x);

sig=x(2:n)'*x(2:n);
v=[1; x(2:n)];

if sig == 0
 beta=0;
else
 mu=sqrt(x(1)^2+sig);
 if x(1) <= 0
  v(1)=x(1)-mu;
 else
  v(1)=-sig/(x(1)+mu);
 end
 beta=2*v(1)^2/(sig+v(1)^2);
 v=v/v(1);
end

P=speye(n,n)-beta*v*v';
```

```
function C=CompanionMatrixVPA(sigma,rozmer,mp)
%   CompanionMatrixVPA computes the companion matrix using
%   spectrum sigma
%   if mp=1 then CompanionMatrixVPA computes inverse of Y using
%   symbolic toolbox
%   rozmer is the dimmension of the the companion matrix
%   sigma are the eigenvalues of the matrix C

if mp==1
    C=vpa(zeros(rozmer,rozmer));
    p=poly(vpa(diag(sigma)));
    alfas=coeffs(p);
    C(:,rozmer)=-alfas(1:rozmer)';
else
    C=zeros(rozmer,rozmer);
    alfas=poly(sigma);
    for i=1:(rozmer)
         C(i,rozmer)=-alfas(rozmer+1-i+1);
    end
end

subdiagonal=ones(rozmer-1,1);
C=C+diag(subdiagonal,-1);
```

```
function Yinv=YinverseVPA(Y,mp)
%   YinverseVPA computes the inverse of Y
%   if mp=1 then YinverseVPA computes inverse of Y using
%   symbolic toolbox

n=length(Y);

if mp==1
    R=vpa(zeros(n-1,n-1));
    pom=vpa(zeros(1,n-1));
else
    R=zeros(n-1,n-1);
    pom=zeros(1,n-1);
end
hhat=pom';

for index1=1:(n-1)
    for index2=1:(n-1)
        R(index1,index2)=Y(index1,index2+1);
    end
end

for index=1:(n-1)
    hhat(index)=Y(index,1);
end

ethaN=Y(n,1);
Rinv=inv(R);
lastColumn=-Rinv*hhat/ethaN;

Yinv=[pom,1/ethaN; Rinv,lastColumn];
```

Bibliography

[1] M. Arioli. A stopping criterion for the conjugate gradient algorithms in a finite element method framework. *Numer. Math.*, 97(1):1–24, 2004.

[2] M. Arioli, E. Noulard, and A. Russo. Stopping criteria for iterative methods: applications to PDE's. *Calcolo*, 38(2):97–112, 2001.

[3] M. Arioli, V. Pták, and Z. Strakoš. Krylov sequences of maximal length and convergence of GMRES. *BIT*, 38(4):636–643, 1998.

[4] E. H. Ayachour. A fast implementation for GMRES method. *J. Comput. Appl. Math.*, 159(2):269–283, 2003.

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

[6] B. Beckermann and A. B. J. Kuijlaars. Superlinear convergence of conjugate gradients. *SIAM J. Numer. Anal.*, 39(1):300–329 (electronic), 2001.

[7] B. Beckermann and A. B. J. Kuijlaars. Superlinear CG convergence for special right-hand sides. *Electron. Trans. Numer. Anal.*, 14:1–19 (electronic), 2002. Orthogonal polynomials, approximation theory, and harmonic analysis (Inzel, 2000).

[8] B. A. Cipra. The Best of the 20th Century: Editors Name Top 10 Algorithms. *SIAM News*, 33(4), 2000.

[9] P. Concus, G. H. Golub, and D. P. O'Leary. A generalized conjugate gradient method for the numerical solution of elliptic partial differential equations. In *Sparse matrix computations (Proc. Sympos., Argonne Nat. Lab., Lemont, Ill., 1975)*, pages 309–332. Academic Press, New York, 1976.

[10] M. Crouzeix. Numerical range and functional calculus in Hilbert space. *J. Funct. Anal.*, 244(2):668–690, 2007.

[11] N. Dunford and J. T. Schwartz. *Linear operators. Part I.* Wiley Classics Library. John Wiley & Sons Inc., New York, 1988. General theory, With the assistance of William G. Bade and Robert G. Bartle, Reprint of the 1958 original, A Wiley-Interscience Publication.

[12] P. J. Eberlein. On measures of non-normality for matrices. *Amer. Math. Monthly*, 72:995–996, 1965.

[13] M. Eiermann. Fields of values and iterative methods, Presented at Oberwolfach meeting on Iterative Methods and Scientific Computing, Oberwolfach, Germany, April 1997.

[14] M. Eiermann. Fields of values and iterative methods. *Linear Algebra Appl.*, 180:167–197, 1993.

[15] M. Eiermann and O. G. Ernst. Geometric aspects of the theory of Krylov subspace methods. *Acta Numer.*, 10:251–312, 2001.

[16] H. C. Elman. *Iterative Methods for Large, Sparse, Nonsymmetric Systems of Linear Equations.* PhD thesis, 1982.

[17] H. C. Elman and A. Ramage. An analysis of smoothing effects of upwinding strategies for the convection-diffusion equation. *SIAM J. Numer. Anal.*, 40(1):254–281 (electronic), 2002.

[18] H. C. Elman and A. Ramage. A characterisation of oscillations in the discrete two-dimensional convection-diffusion equation. *Math. Comp.*, 72(241):263–288 (electronic), 2003.

[19] H. C. Elman, D. J. Silvester, and A. J. Wathen. Iterative methods for problems in computational fluid dynamics. In *Iterative methods in scientific computing (Hong Kong, 1995)*, pages 271–327. Springer, Singapore, 1997.

[20] L. Elsner and K. D. Ikramov. Normal matrices: an update. *Linear Algebra Appl.*, 285(1-3):291–303, 1998.

[21] L. Elsner and M. H. C. Paardekooper. On measures of nonnormality of matrices. *Linear Algebra Appl.*, 92:107–123, 1987.

[22] M. Embree. How descriptive are gmres convergence bounds? Technical Report 99/08, Oxford University Computing Laboratory Numerical Analysis, June 1999.

[23] J. Erhel, K. Burrage, and B. Pohl. Restarted GMRES preconditioned by deflation. *J. Comput. Appl. Math.*, 69(2):303–318, 1996.

[24] O. G. Ernst. Residual-minimizing Krylov subspace methods for stabilized discretizations of convection-diffusion equations. *SIAM J. Matrix Anal. Appl.*, 21(4):1079–1101 (electronic), 2000.

[25] V. Faber, A. Greenbaum, and D. E. Marshall. The polynomial numerical hulls of Jordan blocks and related matrices. *Linear Algebra Appl.*, 374:231–246, 2003.

[26] B. Fischer, A. Ramage, D. J. Silvester, and A. J. Wathen. On parameter choice and iterative convergence for stabilised discretisations of advection-diffusion problems. *Comput. Methods Appl. Mech. Engrg.*, 179(1-2):179–195, 1999.

[27] M. Gérard. Notes on gmres convergence (1) Private communication. 2009.

[28] M. Gérard. Notes on gmres convergence (5) Private communication. 2009.

[29] M. Gérard. Notes on gmres convergence (6) Private communication. 2009.

[30] A. Greenbaum. Generalizations of the field of values useful in the study of polynomial functions of a matrix. *Linear Algebra Appl.*, 347:233–249, 2002.

[31] A. Greenbaum. Card shuffling and the polynomial numerical hull of degree k. *SIAM J. Sci. Comput.*, 25(2):408–416 (electronic), 2003.

[32] A. Greenbaum. Some theoretical results derived from polynomial numerical hulls of Jordan blocks. *Electron. Trans. Numer. Anal.*, 18:81–90 (electronic), 2004.

[33] A. Greenbaum and L. Gurvits. Max-min properties of matrix factor norms. *SIAM J. Sci. Comput.*, 15(2):348–358, 1994.

[34] A. Greenbaum, V. Pták, and Z. Strakoš. Any nonincreasing convergence curve is possible for GMRES. *SIAM J. Matrix Anal. Appl.*, 17(3):465–469, 1996.

[35] A. Greenbaum and Z. Strakoš. Matrices that generate the same Krylov residual spaces. In *Recent advances in iterative methods*, volume 60 of *IMA Vol. Math. Appl.*, pages 95–118. Springer, New York, 1994.

[36] A. Greenbaum and L. N. Trefethen. GMRES/CR and Arnoldi/Lanczos as matrix approximation problems. *SIAM J. Sci. Comput.*, 15(2):359–368, 1994. Iterative methods in numerical linear algebra (Copper Mountain Resort, CO, 1992).

[37] R. Grone, C. R. Johnson, E. M. Sa, and H. Wolkowicz. Normal matrices. *Linear Algebra Appl.*, 87:213–225, 1987.

[38] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Research Nat. Bur. Standards*, 49:409–436 (1953), 1952.

[39] R. Hooke and T. A. Jeeves. Direct search solution of numerical and statistical problems. *J. Assoc. Comp. Mach.*, 8:212–221, 1961.

[40] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, 1990. Corrected reprint of the 1985 original.

[41] R. A. Horn and C. R. Johnson. *Topics in matrix analysis*. Cambridge University Press, Cambridge, 1994. Corrected reprint of the 1991 original.

[42] W. Joubert. A robust GMRES-based adaptive polynomial preconditioning algorithm for nonsymmetric linear systems. *SIAM J. Sci. Comput.*, 15(2):427–439, 1994.

[43] J. Liesen and Z. Strakoš. GMRES convergence analysis for a convection-diffusion model problem. *SIAM J. Sci. Comput.*, 26(6):1989–2009 (electronic), 2005.

[44] J. Liesen and Z. Strakoš. *Principles and Analysis of Krylov Subspace Methods (in preparation).*

[45] J. Liesen and P. Tichý. Convergence analysis of Krylov subspace methods. *GAMM Mitt. Ges. Angew. Math. Mech.*, 27(2):153–173 (2005), 2004.

[46] L. Lukšan, M. Tůma, J. Vlček, N. Ramešová, M. Šiška, J. Hartman, and C. Matonoha. UFO 2008 Interactive System for Universal Functional Optimization. Technical Report 1040, Institute of Computer Science, Academy of Sciences of the Czech Republic, December 2008.

[47] G. Meinardus. Über eine Verallgemeinerung einer Ungleichung von L. V. Kantorowitsch. *Numer. Math.*, 5:14–23, 1963.

[48] G. Meurant and Z. Strakoš. The Lanczos and conjugate gradient algorithms in finite precision arithmetic. *Acta Numerica*, 15:471–542, 2006.

[49] I. Moret. A note on the superlinear convergence of GMRES. *SIAM J. Numer. Anal.*, 34(2):513–516, 1997.

[50] R. B. Morgan. A restarted GMRES method augmented with eigenvectors. *SIAM J. Matrix Anal. Appl.*, 16(4):1154–1171, 1995.

[51] O. Nevanlinna. *Convergence of iterations for linear equations.* Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 1993.

[52] C. C. Paige and M. A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM J. Numer. Anal.*, 12(4):617–629, 1975.

[53] S. C. Reddy and L. N. Trefethen. Lax-stability of fully discrete spectral methods via stability regions and pseudo-eigenvalues. In *Spectral and high order methods for partial differential equations (Como, 1989)*, pages 147–164. North-Holland, Amsterdam, 1990.

[54] Y. Saad. *Iterative methods for sparse linear systems.* Society for Industrial and Applied Mathematics, Philadelphia, PA, second edition, 2003.

[55] Z. Strakoš. Convergence and numerical behaviour of the Krylov space methods. In *Algorithms for large scale linear algebraic systems (Gran Canaria, 1996)*, volume 508 of *NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci.*, pages 175–196. Kluwer Acad. Publ., Dordrecht, 1998.

[56] Z. Strakoš and J. Liesen. On numerical stability in large scale linear algebraic computations. *ZAMM Z. Angew. Math. Mech.*, 85(5):307–325, 2005.

[57] Z. Strakoš. *Theory of Convergence and Effects of Finite Precision Arithmetic in Krylov Subspace Methods.* PhD thesis, Academy of Sciences of the Czech Republic, 2001.

[58] K.-C. Toh. GMRES vs. ideal GMRES. *SIAM J. Matrix Anal. Appl.*, 18(1):30–36, 1997.

[59] L. N. Trefethen. Approximation theory and numerical linear algebra. In *Algorithms for approximation, II (Shrivenham, 1988)*, pages 336–360. Chapman and Hall, London, 1990.

[60] A. van der Sluis and H. A. van der Vorst. The rate of convergence of conjugate gradients. *Numer. Math.*, 48(5):543–560, 1986.

[61] H. A. van der Vorst. *Iterative Krylov methods for large linear systems*, volume 13 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2003.

[62] D. S. Watkins. *Fundamentals of matrix computations*. Pure and Applied Mathematics (New York). Wiley-Interscience [John Wiley & Sons], New York, 2002. Second editon.

[63] R. Winther. Some superlinear convergence results for the conjugate gradient method. *SIAM J. Numer. Anal.*, 17(1):14–17, 1980.

[64] J. Zítko. Generalization of convergence conditions for a restarted GMRES. *Numer. Linear Algebra Appl.*, 7(3):117–131, 2000.

[65] J. Zítko. Analysis of convergence of a restarted GMRES method augmented with eigenvectors. In *Numerical mathematics and advanced applications*, pages 989–996. Springer Italia, Milan, 2003.

[66] J. Zítko. Convergence conditions for a restarted GMRES method augmented with eigenspaces. *Numer. Linear Algebra Appl.*, 12(4):373–390, 2005.

[67] J. Zítko. Some remarks on the restarted and augmented GMRES method. *Electron. Trans. Numer. Anal.*, 31:221–227, 2008.