Charles University in Prague

Faculty of Mathematics and Physics

MASTER THESIS

Tomáš Gergelits

Analysis of Krylov subspace methods

Department of Numerical Mathematics

Supervisor of the master thesis:	prof. Ing. Zdeněk Strakoš, DrSc.
Study programme:	Mathematics
Specialization:	Computational mathematics

Prague 2013

I am grateful to Zdeněk Strakoš, my supervisor, for all the concern and lasting support during writing this master thesis, as well as for his patience and valuable discussions. He was always most helpful and encouraging.

I would like to thank Hanka Bílková for her prompt and efficient help with typographic layout of substantial part of the text.

I truly appreciate help of my colleagues in careful reading of this thesis. Their useful comments has deserved my huge gratitude. Finally I thank them for pleasant and friendly atmosphere.

Last but not least, special thanks go to my family for being supportive during my studies.

This work was supported by the ERC-CZ project LL1202, by the GACR grant 201/09/0917 and by the GAUK grant 695612.

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In .... date .... Tomáš Gergelits

Název práce: Analýza Krylovovských metod
Autor: Tomáš Gergelits
Katedra: Katedra numerické matematiky
Vedoucí diplomové práce: prof. Ing. Zdeněk Strakoš, DrSc.

Abstrakt: Po odvození metody sdružených gradientů (CG) a krátkém přehledu souvislostí s dalšími oblastmi matematiky se práce zaměřuje na konvergenční chování v přesné aritmetice i v aritmetice s konečnou přesnotí. Podrobně je popsán principiální rozdíl mezi CG a Čebyševovou semi-iterační metodou a je diskutována praktická využitelnost široce rozšířeného lineárního odhadu založeného na vlastnostech Čebyševových polynomů. Na příkladu odhadů rychlosti konvergence založených na složených polynomech je ukázána nutnost zahrnutí vlivu zaokrouhlovacích chyb do jakýchkoli úvah o rychlosti konvergence metody CG, které mají být využitelné v praktických výpočtech. Blízkost navzájem si odpovídajících CG aproximací vzniklých ve výpočtech v aritmetice s konečnou přesností a v přesné aritmetice je studována porovnáním jejich trajektorií. Práce je zakončena diskuzí problémů spojených s citlivostí Gauss-Christoffelovy kvadratury, jež s metodou CG úzce souvisí. Na poslední dvě témata může být navázáno v další práci.

Klíčová slova: Metoda sdružených gradientů, Čebyševova semi-iterační metoda, výpočty v konečné aritmetice, zpoždění konvergence, odhady rychlosti konvergence založené na složených polynomech, citlivost Gauss-Christoffelovy kvadratury

Title: Analysis of Krylov subspace methods
Author: Tomáš Gergelits
Department: Department of Numerical Mathematics
Supervisor: prof. Ing. Zdeněk Strakoš, DrSc.

Abstract: After the derivation of the Conjugate Gradient method (CG) and the short review of its relationship with other fields of mathematics, this thesis is focused on its convergence behaviour both in exact and finite precision arithmetic. Fundamental difference between the CG and the Chebyshev semi-iterative method is described in detail. Then we investigate the use of the widespread linear convergence bound based on Chebyshev polynomials. Through the example of the composite polynomial convergence bounds it is showed that the effects of rounding errors must be included in any consideration concerning the CG rate of convergence relevant to practical computations. Furthermore, the close correspondence between the trajectories of the CG approximations generated in finite precision and exact arithmetic is studied. The thesis is concluded with the discussion concerning the sensitivity of the closely related Gauss-Christoffel quadrature. The last two topics may motivate our further research.

Keywords: Conjugate Gradient Method, Chebyshev semi-iterative method, finite precision computations, delay of convergence, composite polynomial convergence bounds, sensitivity of the Gauss-Christoffel quadrature

Contents

Introduction

Krylov subspace methods for iteratively solving large and sparse linear algebraic systems and eigenvalue problems are widely used in the matrix computations and they are counted among the "Top 10 Algotihms" of the 20th century [5, 6]. They can be considered as the projection methods [28] where the approximate solution is found in the sequence of nested subspaces. These subspaces are build up using increasing power of the operator (matrix) with respect to the initial vector and thus the Krylov subspace methods are by their nature highly nonlinear. One of their main advantages is that the matrix does not need to be explicitly stored. Instead, only a function which performs matrix-vector multiplication is required. Their analysis and the understanding of their convergence behaviour is the subject of research of numerous mathematicians.

In this master thesis we restrict ourselves to Hermitian positive definite matrices and we study convergence behaviour of the conjugate gradient method (CG) introduced by Hestenes and Stiefel in 1952; see [13]. We believe that its deep understanding can help in the study of the other Krylov subspace methods (though in the non-normal case the matter is far more complicated).

Substantial part of this master thesis has recently been published in the journal Numerical Algorithms as the original paper

- Gergelits T., Strakoš Z., *Composite convergence bounds based on Chebyshev polynomials and finite precision conjugate gradient computations,* Numerical Algorithms (2013), available online at `http://dx.doi.org/10.1007/s11075-013-9713-z`.

The exposition in the master thesis is subordinate also to other topics not covered in the paper. We enclose the paper as an attachment in Appendix which represents the inherent part of the master thesis. The topics covered in the paper are briefly outlined in Section 1.4 and in Chapter 4; for the full exposition we refer to the enclosed paper. The integration of the paper to this master thesis and the use of its content for the academical purposes is in agreement with the policy of the Springer, the copyright holder. The final publication is available at `link.springer.com`.

The master thesis is organized as follows. Chapter 1 briefly reviews, after derivation of the CG method through the minimization of the quadratic functional, the known relationships of the CG method with the Lanczos algorithm, Vorobyev's moment problem, (simplified) Stieltjes moment problem, the Gauss-Christoffel quadrature and the orthogonal polynomials. The chapter continues with the review of the convergence properties of the CG method in exact arithmetic. It furthermore discuss the practical relevance of the widespread linear convergence bound based on the Chebyshev polynomials and describe the fundamental difference between the CG method and the Chebyshev semi-iterative (CSI) method. In Chapter 2 we describe theoretical results of Paige [23, 24, 25, 26] and Greenbaum [10] which allow to understand and mathematically describe the finite precision behaviour of the CG method. Loss of orthogonality caused by rounding errors in finite precision CG computations results in a delay of convergence and the convergence *rate* may be substantially different in finite precision and exact computations. As we have found and demonstrated in this master thesis, the *trajectories* of the exact and the computed approximations are close to each other. This correspondence is studied in Chapter 3. Chapter 4 summarizes the investigation of the composite polynomial convergence bounds in finite precision CG computations (details are given in Appendix). In Chapter 5 we investigate the behaviour of the closely related Gauss-Christoffel quadrature for the distribution functions with clustered points of increase.

We are well aware of the fact that the Krylov subspace methods are seldom used without preconditioning. We can consider the studied linear systems as the preconditioned system and the presented results are thus applicable to the preconditioned CG method. An interested reader is referred to [28, Chapters 9–14] or the survey [2].

1. The CG method

Method of Conjugate Gradients (CG) is an iterative method for solving systems of linear equations

$$Ax = b \tag{1.1}$$

with Hermitian and positive definite (HPD) matrix $A \in \mathbb{C}^{N\times N}$ and the right-hand side $b \in \mathbb{R}^N$. The CG method is optimal in a sense that it minimizes the energy norm of the error over the given Krylov subspace.

We would like to emphasize here the importance of normality of the matrix A. It ensures that the spectral decomposition represent the superposition of A onto one-dimensional subspace which are orthogonal to each other, and it insures that the convergence behaviour of the CG method is fully determined by the eigenvalues of A and the right-hand side b.

1.1 Derivation of the CG method

The CG method can be derived using many different approaches. The approach we use in this thesis is based on the well known equivalence between solving the linear system

$$Ax = b \tag{1.2}$$

where $A \in \mathbb{C}^{N\times N}$ is Hermitian positive definite (HPD) matrix and $b \in \mathbb{C}^N$ is a right-hand side vector, and the minimization of the quadratic functional

$$F(z) = \frac{1}{2}z^*Az - z^*b. \tag{1.3}$$

Our exposition follows [19, Section 2.5.3].

Considering an approximation x_k to the solution x of (1.2) (i.e., the minimum of the functional (1.3)) gives the equality

$$F(x_k) = \frac{1}{2}(x - x_k)^*A(x - x_k) - \frac{1}{2}x^*Ax = \frac{1}{2}\left\|x - x_k\right\|_A^2 - \frac{1}{2}\left\|x\right\|_A^2 \tag{1.4}$$

where $\left\|z\right\|_A := (z^*Az)^{1/2}$ is the energy norm (also called A-norm) and we see that the minimalization of the functional $F(z)$ over some subspace of $\mathbb{C}^N$ is the same as the minimalization of $\left\|x - z\right\|_A$ over the same subspace $\mathbb{C}^N$. Thus the energy norm is the natural measure of distance of the approximation x_k to the solution x.

Let x_0 be an initial approximation and let the sequence of approximations x_k be constructed by the simple recurrence

$$x_k = x_{k-1} + \alpha_{k-1}p_{k-1}, \quad k = 1, 2, \ldots \tag{1.5}$$

where p_k is carefully chosen direction vector and the coefficient α_{k-1} is settled to minimize the functional $F(z)$ (i.e., $\left\|x - z\right\|_A$) along the line $x_{k-1} + \alpha p_{k-1}$. Simple algebraic manipulation gives

$$\left\|x - x_k\right\|_A^2 = \left\|x - x_{k-1}\right\|_A^2 - 2\alpha(x - x_{k-1})^*Ap_{k-1} + \alpha^2 p_{k-1}^*Ap_{k-1}$$

and thus the minimum is attained for

$$\alpha_{k-1} = \frac{p_{k-1}^* r_{k-1}}{p_{k-1}^* A p_{k-1}}, \tag{1.6}$$

where $r_k = A(x - x_k)$ is the residual vector which, using (1.5), can be computed iteratively

$$r_k = r_{k-1} - \alpha_{k-1} A p_{k-1} \quad k = 1, 2, \ldots. \tag{1.7}$$

An immediate consequence of the choice of the parameter α_{k-1} is the orthogonality between the residual and the direction vector, i.e.,

$$p_{k-1}^* r_k = p_{k-1}^* (r_{k-1} - \alpha_{k-1} A p_{k-1}) = 0. \tag{1.8}$$

It remains to choose the direction vectors p_k. The simplest choice for the initial vector p_0 is $p_0 \equiv r_0$. For the choice $p_k \equiv r_k$ we get the method of the steepest descent ($r_k = -\nabla F(x_k)$), its convergence is guaranteed but can be very poor. The main reason for the poor convergence is that in every step we use the information only from the last iteration and the choice of α_{k-1} gives only one-dimensional minimization in each step. In order to minimize over subspaces of larger dimension, the direction vector p_k must combine information from several iteration steps. The simplest possibility is to add an information about the previous direction vector p_{k-1} and to compute

$$p_k = r_k + \beta_k p_{k-1}. \tag{1.9}$$

The iteration process can stop only if $p_k = 0$ or $\alpha_k = 0$. Independently on the choice of the parameter β_k, the orthogonality between p_{k-1} and r_k (see (1.8)) gives

$$p_k^* r_k = r_k^* r_k = \|r_k\|^2 \tag{1.10}$$

and in both cases of possible breakdown we get $r_k = 0$ and $Ax_k = b$.

In order to motivate the choice of β_k below, we first notice that the use of (1.5) and (1.6) gives

$$x - x_k = x - x_{k-1} - \frac{p_{k-1}^* A(x - x_{k-1})}{p_{k-1}^* A p_{k-1}} p_{k-1}. \tag{1.11}$$

and thus the error $x - x_k$ can be viewed as the A-orthogonalization of $x - x_{k-1}$ against the direction vector p_{k-1}. In other words,

$$x - x_{k-1} = x - x_k + \alpha_{k-1} p_{k-1} \tag{1.12}$$

can be viewed as the A-orthogonal decomposition of $x - x_{k-1}$. The Pythagorean theorem then gives

$$\|x - x_{k-1}\|_A^2 = \|x - x_k\|_A^2 + \alpha_{k-1}^2 \|p_{k-1}\|_A^2 . \tag{1.13}$$

The repetitive use of (1.12) and (1.13) gives

$$x - x_0 = \sum_{j=1}^{k} \alpha_{j-1} p_{j-1} + (x - x_k) \tag{1.14}$$

and

$$\|x - x_0\|_A^2 = \|x - x_k\|_A^2 + \sum_{j=0}^{k-1} \alpha_j^2 \|p_j\|_A^2 \,. \tag{1.15}$$

The expansion (1.14) and the identity (1.15) holds for arbitrary choice of the direction vectors $p_j,\ j = 0, \ldots, k-1$.

Now *assume* that the direction vectors p_j, $j = 0, \ldots, k-1$ are A-orthogonal. Then

$$x - x_k = x - x_0 - \sum_{j=1}^{k} \alpha_{j-1} p_{j-1} \tag{1.16}$$

represents the A-orthogonal decomposition of the initial error $x - x_0$. Consequently, $\|x - x_k\|_A$ is minimal over all possible approximations x_k in the k-th dimensional subspace generated by the direction vectors $p_0, \ldots, p_{k-1}$, i.e.,

$$\|x - x_k\|_A = \min_{y \in x_0 + \mathrm{span}\{p_0, \ldots, p_{k-1}\}} \|x - y\|_A \,. \tag{1.17}$$

Moreover, the A-orthogonality of the direction vectors implies that $p_N = 0$ and thus the iteration process finds the exact solution in at most N steps.

We have only one undetermined coefficient β_k and thus we can prescribe only the local A-orthogonality between the subsequent direction vectors

$$p_{k-1}^* A p_k = 0 \tag{1.18}$$

which gives

$$\beta_k = -\frac{p_{k-1}^* A r_k}{p_{k-1}^* A p_{k-1}}. \tag{1.19}$$

and the algorithm is fully determined. It can be shown by induction (see, e.g., [13, Theorem 5:1]) that

$$p_i^* A p_j = 0 \quad \text{and} \quad r_i^* r_j = 0 \quad \text{for} \quad j \neq i. \tag{1.20}$$

Thus we see, that the described choice of the coefficients β_k motivated by the *local* A-orthogonality guarantees the *global* A-orthogonality of the direction vectors, the *global* orthogonality of the residual vectors and thus the minimalization property (1.17) is guaranteed. Finally, using $p_{k-1}^* r_{k-1} = r_{k-1}^* r_{k-1}$ and

$$-A p_{k-1} = \alpha_{k-1}^{-1} (r_k - r_{k-1}) = \frac{p_{k-1}^* A p_{k-1}}{p_{k-1}^* r_{k-1}} (r_k - r_{k-1}),$$

we get

$$\alpha_{k-1} = \frac{r_k^* r_k}{p_{k-1}^* A p_{k-1}} \quad \text{and} \quad \beta_k = \frac{r_k^* r_k}{r_{k-1}^* r_{k-1}} \tag{1.21}$$

Combining the equations (1.5), (1.9) and (1.21) gives the standard implementation of the CG method; see Algorithm I.

1.2 The CG method in context

In this section we give a description of the relationship of the CG method with the Lanczos algorithm and with the Vorobyev's moment problem. Furthermore, we briefly review interconnections of the CG method with the simplified Stieltjes moment problem, the Gauss-Christoffel quadrature and the orthogonal polynomials. We believe these relationships of great importance for proper understanding of the behaviour of the CG method. The exposition loosely follows [30] and [19, Section 3.7]. We refer to the recent monograph [19] for a detail exposition of these topics which also contains many historical comments and an extensive list of references.

From the definition of r_j and p_j and from their orthogonality properties, it easily follows that the residual vectors and the direction vectors form the orthogonal (resp. A-orthogonal) basis of the Krylov subspace

$$\mathcal{K}_k(A, r_0) \equiv \operatorname{span}(r_0, Ar_0, \dots, A^{k-1}r_0) \tag{1.22}$$

associated with the matrix A and the vector r_0, i.e.,

$$\begin{aligned} \mathcal{K}_k(A, r_0) &= \operatorname{span}(r_0, \dots, r_{k-1}) \\ &= \operatorname{span}(p_0, \dots, p_{k-1}). \end{aligned} \tag{1.23}$$

Consequently, the CG approximations are uniquely determined by the relations

$$x_k \in x_0 + \mathcal{K}_k(A, r_0), \quad r_k \perp \mathcal{K}_k(A, r_0), \quad k = 1, 2, \dots. \tag{1.24}$$

and we see that the CG method is the projection method where both search and constraints space are the k-th Krylov subspace. Throughout this section we assume that the dimension of $\mathcal{K}_k(A, r_0)$ is equal to k, i.e., that the CG algorithm has not stopped before and (for the simplicity of the exposition) that the eigenvalues of A are distinct.

The Lanczos algorithm (see Algorithm II) introduced in [17] constructs in the exact arithmetic an orthonormal basis $\{v_1, \dots, v_k\}$ of the k-th Krylov subspace associated with the HPD matrix A and the initial vector r_0. In matrix notation, the Lanczos algorithm can be expressed as

$$AV_k = V_k T_k + \delta_{k+1} v_{k+1} e_k^T \tag{1.25}$$

Algorithm I The CG method

```
input $A$, $b$, $x_0$
$r_0 := b - Ax_0$
$p_0 := r_0$
for $k = 1, 2, \dots$
    $\alpha_{k-1} := \frac{r_{k-1}^* r_{k-1}}{p_{k-1}^* A p_{k-1}}$
    $x_k := x_{k-1} + \alpha_{k-1} p_{k-1}$
    $r_k := r_{k-1} - \alpha_{k-1} A p_{k-1}$
    $\beta_k := \frac{r_k^* r_k}{r_{k-1}^* r_{k-1}}$
    $p_k := r_k + \beta_k p_{k-1}$
end
```

Algorithm II The Lanczos algorithm

```
input $A$, $r_0$
$v_0 := 0$
$v_1 := r_0/\|r_0\|$
$\delta_1 := 0$
for $k = 1, 2, \ldots$
    $w := Av_k - \delta_k v_{k-1}$
    $\gamma_k := v_k^* w$
    $w := w - \gamma_k v_k$
    $\delta_{k+1} := \|w\|$
    $v_{k+1} := w/\delta_{k+1}$
end
```

where V_k is the column matrix of the Lanczos vectors v_j, $j = 1, \ldots, k$, and where

$$T_k \equiv \begin{bmatrix} \gamma_1 & \delta_2 & & & \\ \delta_2 & \gamma_2 & \delta_3 & & \\ & \ddots & \ddots & \ddots & \\ & & \delta_{k-1} & \gamma_{k-1} & \delta_k \\ & & & \delta_k & \gamma_k \end{bmatrix} \tag{1.26}$$

is the Jacobi matrix (i.e., the symmetric tridiagonal matrix with positive subdiagonal entries) which contains the coefficients of the Lanczos recurrence. With the use of the orthogonality condition

$$0 = V_k^* r_k = V_k^*(r_0 - AV_k y_k) = \|r_0\| e_1 - V_k^* A V_k y_k = \|r_0\| e_1 - T_k y_k$$

we see that the CG approximation x_k can be in exact arithmetic equivalently computed as

$$x_k = x_0 + V_k y_k, \quad T_k y_k = \|r_0\| e_1. \tag{1.27}$$

The eigenvalues of T_k are called Ritz values and they are computed from T_k by the Lanczos method to approximate a few dominant eigenvalues of A. We enclose the correspondence between the CG method and the Lanczos algorithm by the identities among the computed vectors and the recurrence coefficients:

$$\begin{aligned} v_{k+1} &= (-1)^k \frac{r_k}{\|r_k\|}, \\ \gamma_k &= \frac{1}{\alpha_{k-1}} + \frac{\beta_{k-1}}{\alpha_{k-2}}, \quad \beta_0 \equiv 0, \ \alpha_{-1} \equiv 1, \\ \delta_{k+1} &= \frac{\sqrt{\beta_k}}{\alpha_{k-1}}, \qquad\qquad k = 1, 2, \ldots. \end{aligned} \tag{1.28}$$

Consider the $N \times N$ Hermitian matrix A as an operator on $\mathbb{C}^N$. The mapping

$$V_k V_k^* : \mathbb{C}^N \to \mathcal{K}_k(A, r_0) \tag{1.29}$$

then represents an orthogonal projector onto Krylov subspace $\mathcal{K}_k(A, r_0)$ and the operator

$$A_k \equiv V_k V_k^* A V_k V_k^* \tag{1.30}$$

is the restriction and orthogonal projection of the operator A onto $\mathcal{K}_k(A, r_0)$. The matrix representation of this operator A_k with respect to the basis V_k is then given by the tridiagonal matrix T_k. Using the Vorobyev's operator formulation of the problem of moments, we will interpret the tridiagonal matrix T_k as a result of the moment matching model reduction.

We see that the action of the operator A_k defined on $\mathcal{K}_k(A, r_0)$ is identical to the action of the operator A restricted to $\mathcal{K}_k(A, r_0)$ and thus A_k is the solution of the Vorobyev's moment problem:

Given A and r_0, determine A_k such that

$$\begin{aligned} A_k r_0 &= A r_0 \\ A_k^2 r_0 &= A^2 r_0 \\ &\vdots \\ A_k^{k-1} r_0 &= A^{k-1} r_0 \\ A_k^k r_0 &= (V_k V_k^*) A^k r_0. \end{aligned} \tag{1.31}$$

Now we will see that the operator A_k matches the first $2k$ moments, i.e., it holds

$$r_0^* A_k^s r_0 = r_0^* A^s r_0, \quad s = 0, 1, \dots, 2k-1, \tag{1.32}$$

equivalently,

$$e_1^* T_k^s e_1 = v_1^* A^s v_1, \quad s = 0, 1, \dots, 2k-1. \tag{1.33}$$

For $s = 1, \dots, n-1$, the statement follows from (1.31) and it is trivial for $s = 0$. Since $V_k V_k^*$ is a projector, multiplication of the last row of (1.31) by $V_k V_k^*$ implies

$$V_k V_k^* (A_k^k r_0 - A^k r_0) = 0 \tag{1.34}$$

and thus the vector $(A_k^k r_0 - A^k r_0)$ is orthogonal to all basis vectors $r_0, \dots, A^{k-1} r_0$. The symmetry of A and A_k and the use of (1.31) then gives

$$r_0^T A^j (A^k r_0 - A_k^k r_0) = 0, \quad j = 0, \dots, k-1 \tag{1.35}$$

which gives the result.

The relationship with the simplified Stieltjes moment problem and the Gauss-Christoffel quadrature now easily follows from the use of the spectral decompositions

$$T_k = Z_k \Theta_k Z_k^*, \quad A = Q \Lambda Q^*, \quad \text{where} \tag{1.36}$$

$Z_k = [z_1^{(k)}, \dots, z_k^{(k)}]$ and $Q = [q_1, \dots, q_N]$ are the orthogonal matrices of eigenvectors of T_k and A and where $\Theta_k = \operatorname{diag}(\theta_1^{(k)}, \dots, \theta_k^{(k)})$ and $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_N)$ are the diagonal matrices with the eigenvalues $0 < \xi < \theta_1^{(k)} < \dots < \theta_k^{(k)} < \zeta$ and $0 < \xi < \lambda_1 < \dots < \lambda_N < \zeta$. Substitution of the spectral decompositions (1.36) to the identity (1.33) gives

$$\sum_{j=1}^{k} \omega_j^{(k)} \left\{ \theta_j^{(k)} \right\}^s = \sum_{j=1}^{N} \omega_j \lambda_j^s, \quad s = 0, 1, \dots, 2k-1 \tag{1.37}$$

where $\omega_j^{(k)} = (z_j^{(k)}, e_1)^2$ and $\omega_j = (q_j, v_1)^2$. These sums can be considered as the Riemann-Stieltjes integrals with respect to the piecewise constant non-decreasing distribution functions

$$\omega^{(k)}(\lambda) = \begin{cases} 0 & \text{for} \quad \xi \leq \lambda < \theta_1^{(k)} \\ \sum_{j=1}^{i} \omega_j^{(k)} & \text{for} \quad \theta_i^{(k)} \leq \lambda < \theta_{i+1}^{(k)}, \quad i = 1, \dots, k-1 \\ \sum_{j=1}^{k} \omega_j^{(k)} = 1 & \text{for} \quad \theta_k^{(k)} \leq \lambda < \zeta \end{cases} \tag{1.38}$$

and

$$\omega(\lambda) = \begin{cases} 0 & \text{for} \quad \xi \leq \lambda < \lambda_1 \\ \sum_{j=1}^{i} \omega_j & \text{for} \quad \lambda_i \leq \lambda < \lambda_{i+1}, \quad i = 1, \dots, N-1 \\ \sum_{j=1}^{N} \omega_j = 1 & \text{for} \quad \lambda_N \leq \lambda < \zeta \end{cases} \tag{1.39}$$

and (1.37) then takes the form

$$\int_{\xi}^{\zeta} \lambda^s \, d\omega^{(k)}(\lambda) = \int_{\xi}^{\zeta} \lambda^s \, d\omega(\lambda), \quad s = 0, 1, \dots, 2k-1. \tag{1.40}$$

We see that the left sides of the identities (1.40) and (1.37) represent the solution of the simplified Stieltjes moment problem, i.e.,

$$\int_{\xi}^{\zeta} f(\lambda) \, d\omega^{(k)}(\lambda) = \sum_{j=1}^{k} \omega_j^{(k)} f(\theta_j^{(k)}) = e_1^* f(T_k) e_1 \tag{1.41}$$

represents the k-th Gauss-Chistoffel quadrature of the Riemann-Stieltjes integral

$$\int_{\xi}^{\zeta} f(\lambda) \, d\omega(\lambda) = \sum_{j=1}^{N} \omega_j f(\lambda_j) = v_1^* f(A) v_1. \tag{1.42}$$

Thus the CG method determines the sequence of distribution functions $\omega^{(j)}(\lambda)$, $j = 1, 2, \dots$ which approximate in the optimal way the original distribution function $\omega(\lambda)$. Their weights and nodes are equal to the squared first components of the associated normalized eigenvectors and to the eigenvalues of the Jacobi matrix T_k generated in the first k steps of the Lanczos process applied to the matrix A and the initial vector r_0.

The relation (1.24) implies that the error of the k-th CG approximation can be written in terms of the polynomial with constant term equal to one, i.e.,

$$x - x_k = \varphi_k^{CG}(A)(x - x_0), \quad \varphi_k(0) = 1. \tag{1.43}$$

We point out that the CG polynomials $\varphi_j^{CG}(\lambda)$, $j = 0, 1, \dots, k-1$ form a sequence of orthogonal polynomials with respect to both scalar products induced by the distribution functions $\omega(\lambda)$ and $\omega^{(k)}(\lambda)$. The CG polynomial $\varphi_k^{CG}(\lambda)$ represents the $(k+1)$-st orthogonal polynomial with respect to $\omega(\lambda)$ and its zeros are the nodes $\theta_j^{(k)}$, $j = 1, \dots, k$ of $\omega^{(k)}(\lambda)$, i.e.,

$$\varphi_k^{CG}(\lambda) = (-1)^k \frac{(\lambda - \theta_1^{(k)}) \dots (\lambda - \theta_k^{(k)})}{\theta_1^{(k)} \dots \theta_k^{(k)}}. \tag{1.44}$$

We have found interesting identity (see also [19, Theorem 3.4.11], it holds that

$$\int_{\xi}^{\zeta} \frac{(\varphi_k^{CG}(\lambda))^2}{\lambda}\, d\omega(\lambda) = \int_{\xi}^{\zeta} \frac{\varphi_k^{CG}(\lambda)}{\lambda}\, d\omega(\lambda), \quad \text{i.e.,} \tag{1.45}$$

$$\sum_{i=1}^{N} \frac{(\varphi_k^{CG}(\lambda_i)^2}{\lambda_i}\omega_i = \sum_{i=1}^{N} \frac{\varphi_k^{CG}(\lambda_i)}{\lambda_i}\omega_i. \tag{1.46}$$

Proof.

$$\int_{\xi}^{\zeta} \frac{(\varphi_k^{CG}(\lambda))^2 - \varphi_k^{CG}(\lambda)}{\lambda}\, d\omega(\lambda) = \int_{\xi}^{\zeta} \frac{\varphi_k^{CG}(\lambda)(\varphi_k^{CG}(\lambda)-1)}{\lambda}\, d\omega(\lambda) \tag{1.47}$$

$$= \int_{\xi}^{\zeta} \varphi_k^{CG}(\lambda)\frac{(\varphi_k^{CG}(\lambda)-1)}{\lambda}\, d\omega(\lambda). \tag{1.48}$$

Since $\varphi_k^{CG}(0) = 1$, we know that $\nu(\lambda) = \frac{(\varphi_k^{CG}(\lambda)-1)}{\lambda}$ is the polynomial of degree $k-1$. Since the k-th CG polynomial is orthogonal with respect to distribution function $\omega(\lambda)$ to any polynomial of lower degree, we get

$$\int_{\xi}^{\zeta} \varphi_k^{CG}(\lambda)\nu(\lambda)\, d\omega(\lambda) = 0. \tag{1.49}$$

□

Similarly, it holds that

$$\int_{\xi}^{\zeta} \frac{(\varphi_j^{CG}(\lambda))^2}{\lambda}\, d\omega^{(k)}(\lambda) = \int_{\xi}^{\zeta} \frac{\varphi_j^{CG}(\lambda)}{\lambda}\, d\omega^{(k)}(\lambda), \quad j = 0, \dots, k-1. \tag{1.50}$$

1.3 Energy norm, its bounds and estimates

In this section we present several equivalent formulations of the energy norm of the CG error, in order to illustrate its nonlinear behaviour and its correspondence to convergence of Ritz values. We enclose this section by the identity which is one of the conner stones of reliable a-posteriori error estimates.

From the (1.17), (1.23) and (1.43) it follows that

$$\begin{aligned} \|x - x_k\|_A^2 &= \min_{y \in x_0 + \mathcal{K}(A, r_0)} \|x - y\|_A^2 = \min_{\substack{\varphi(0)=1 \\ \deg(\varphi) \le k}} \|\varphi(A)(x - x_0)\|_A^2 \\ &= \min_{\substack{\varphi(0)=1 \\ \deg(\varphi) \le k}} \|r_0\|^2 \sum_{j=1}^{N} \frac{\varphi^2(\lambda_j)}{\lambda_j}\omega_j = \|r_0\|^2 \sum_{j=1}^{N} \frac{(\varphi_k^{CG}(\lambda_j)^2}{\lambda_j}\omega_j, \end{aligned} \tag{1.51}$$

i.e., we see that the CG convergence depends on eigenvalues of A and on the size of projections of initial residual to the eigenvectors. From (1.51) we also get

$$\frac{\|x - x_k\|_A}{\|x - x_0\|_A} \le \min_{\substack{\varphi(0)=1 \\ \deg(\varphi) \le k}} \max_{i=1,\dots,n} |\varphi(\lambda_i)|. \tag{1.52}$$

In [9, pp. 253-254] it has been proved that

$$E^k(\lambda^{-1}) = \frac{\|x - x_k\|_A^2}{\|r_0\|^2}, \tag{1.53}$$

where

$$E^k(\lambda^{-1}) = \int_\xi^\zeta \lambda^{-1}\, d\omega(\lambda) - \int_\xi^\zeta \lambda^{-1}\, d\omega^{(k)}(\lambda) \tag{1.54}$$

is the k-th error of the Gauss-Christoffel quadrature. From (1.54) and (1.51) it follows that

$$\frac{\|x - x_k\|_A^2}{\|r_0\|^2} = \sum_{j=1}^{N} \prod_{l=1}^{k} \left(\frac{1}{\lambda_j^{1/2}} - \frac{\lambda_j^{1/2}}{\theta_l^{(k)}} \right)^2 \omega_j. \tag{1.55}$$

1.4 Comparison of CG with the Chebyshev semi-iterative (CSI) method

In Section B.2 of the enclosed paper Appendix B we carefully reveal the fundamental difference between the convergence properties of the Chebyshev semi-iterative (CSI) method and the CG method. Here we summarize the obtained results and refer the reader to Section B.2 for details.

- The Chebyshev polynomials are orthogonal with respect to certain continuous and discrete inner products which contain, apart from the extremal eigenvalues λ_1 and λ_N, no further information about the data A, b and r_0. In comparison, the CG polynomials are orthogonal with respect to discrete inner product which is fully determined by the data A and r_0.

- The CG norm of the error is determined by the discrete minimization problem, i.e.,

  $$\|x - x_k^{CG}\|_A^2 = \|r_0\|^2 \sum_{j=1}^{N} \frac{(\varphi_k^{CG}(\lambda_j)^2}{\lambda_j} \omega_j.$$

  The CSI norm of the error is tightly bounded by the minimization problem over the whole interval $[\lambda_1, \lambda_N]$, i.e.,

  $$\|x - x_k^{CSI}\|_A^2 \le \max_{\lambda \in [\lambda_1, \lambda_N]} \left|\phi_k^{CSI}(\lambda)\right|$$

- The linear bound

  $$2 \left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \right)^k$$

  is *relevant to the CSI method* and only as a consequence is valid also for the CG method.

2. The CG method in finite precision arithmetic

A detail analysis of the finite precision convergence behaviour of the CG method and of the closely related Lanczos algorithm is present in [20], where numerous numerical experiments are performed and where also the technically complicated issues are considered. In this chapter we want to review the most important results which allow to understand and even mathematically describe the behaviour of finite precision CG computations. Our exposition is based on [19, Section 5.9] and on the nicely written survey paper [21].

In finite precision computations, both Lanczos vectors from the Lanczos algorithm and the residual vectors from the CG method loose (usually very quickly) their orthogonality. As a consequence of the loss of orthogonality in the Lanczos algorithm, the elements of the computed Jacobi matrix T_k may differ by several orders of their exact arithmetic counterparts. Moreover, the multiple Ritz approximations to single original eigenvalues appear and, consequently, the convergence of Ritz values to other eigenvalues is delayed. The consequence of the loss of orthogonality in the CG algorithm is illustrated in Figure 2.1, which depicts the relative energy norm of the error of the CG algorithm applied to a matrix A given by **Spectrum 1-Q** (see Appendix A) with the parameters

$$N = 25, \quad \lambda_1 = 0.1, \quad \lambda_N = 1000, \quad \rho = 0.6,$$

right-hand side b of ones and the zero initial approximation. The dash-dotted line corresponds to exact CG computation[1] and the solid line to finite precision

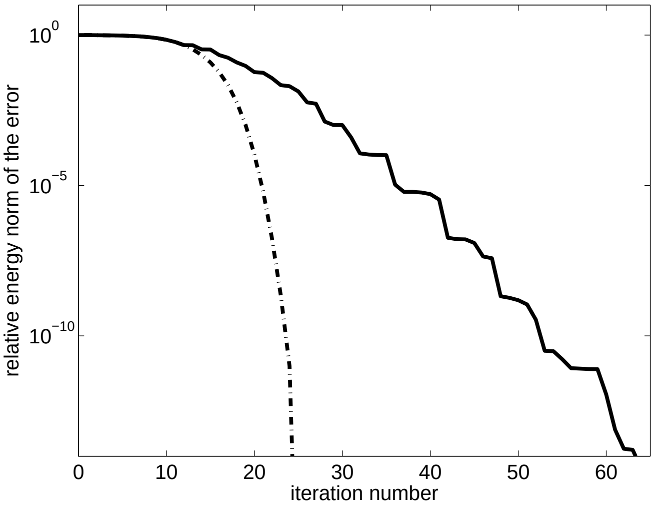


Figure 2.1: Delay of convergence in CG computations due to rounding errors. The convergence curve of finite precision CG computations (solid line) differs both qualitatively (see the characteristic staircase behaviour) and quantitatively from the curve of exact CG computations (dash-dotted line).

[1] The exact arithmetic is simulated by full double reorthogonalization of the computed residuals; see [26, p. 252], [11].

CG computation. As expected, the error of the exact CG vanishes at iteration 25. However, in finite precision computations, 25 iterations are not sufficient to reach even the modest decrease of the error and far more iterations are needed to obtain a small error. We see that finite precision computations require more iterations than its exact arithmetic counterpart to reach the same level of accuracy, i.e., the convergence of the CG approximate solution is delayed.

In exact arithmetic, the CG convergence behaviour depends on the convergence of the Ritz values to the eigenvalues of A; see (1.55). Intuitively, we may expect the same in finite precision computations. Indeed, the work of Anne Greenbaum [10], reviewed below in Section 2.1, reveals that the appearance of the multiple Ritz approximations cause a delay of convergence in finite precision CG computations.

The analysis of Greenbaum is heavily based on the results of Paige [23, 24, 25, 26]. He presented rigorous mathematical analysis of rounding errors in the Lanczos algorithm which allows to understand the numerical behaviour of the Lanczos method. Despite the common wisdom of that time, he clarified that the Lanczos method can be used as a reliable and efficient numerical tool for computing accurate approximations of dominant eigenvalues of the matrix A. Moreover, he revealed that the algorithm behaves numerically like the Lanczos algorithm with full reorthogonalization until a very close eigenvalue approximation is found. The most celebrated result of him is, that the loss of orthogonality follows a rigorous pattern. He has proved that the loss of orthogonality can occur only in the direction corresponding to converged Ritz value.

2.1 Backward-like analysis by Greenbaum

Exposition in this section follows a part of Section B.4 of the paper enclosed in Appendix B. Shortly speaking, Greenbaum has proved that

> the finite precision Lanczos computation for a matrix A and a given starting vector v produces in steps 1 through k the same Jacobi matrix T_k as the exact Lanczos computation for some particular larger matrix $\widehat{A}(k)$ and some particular starting vector $\widehat{v}(k)$ while the eigenvalues of $\widehat{A}(k)$ all lie *within tiny intervals around the eigenvalues of A*. The size as well as (all) individual entries of $\widehat{A}(k)$ and $\widehat{v}(k)$ depend on the rounding errors in the steps 1 through k.

Moreover, it has been shown, that the relationship between the CG method and the Lanczos algorithm holds, with small inaccuracy, also in finite precision. Thus an analogous statement is valid, with a small inaccuracy specified in [10], also for the behaviour of finite precision CG computations. We would like to emphasize that $\widehat{A}(k)$ *is not given by a slight perturbation of the matrix A*. The matrix $\widehat{A}(k)$ is typically much larger than A; see the illustration in Figure 2.2.

As stated above, the matrix $\widehat{A}(k)$ and the vector $\widehat{v}(k)$ depend on the iteration step k. The numerical experiments performed in [11] and the reasoning about the delay in finite precision CG computations suggests that the particular matrix $\widehat{A}(k)$ constructed for the k steps of the given finite precision CG computation can be replaced (with an acceptable inaccuracy) by a matrix $\widehat{A}$ having sufficiently many

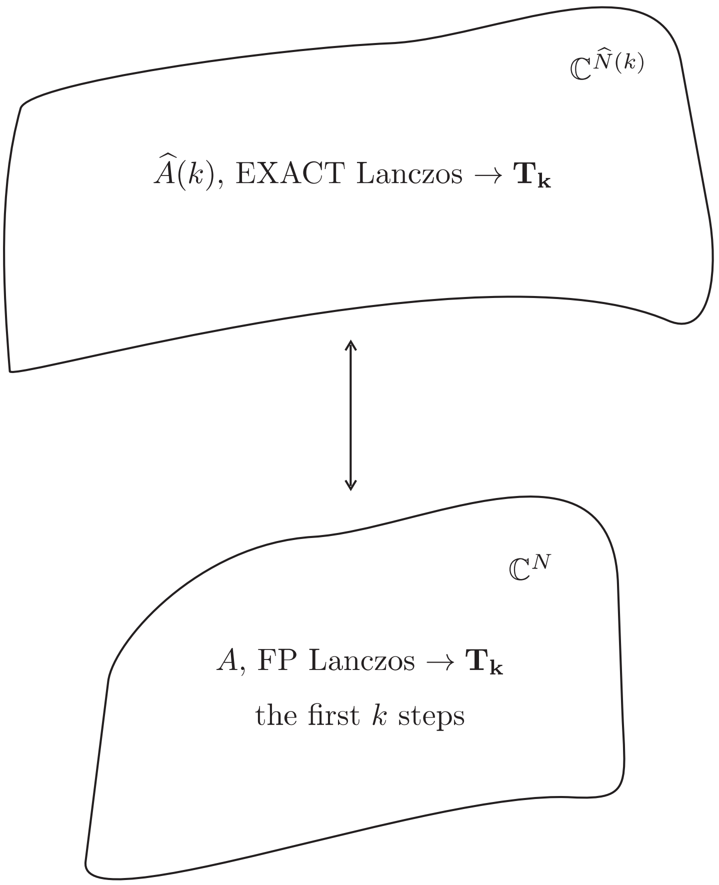

$\mathbb{C}^{\widehat{N}(k)}$
$\widehat{A}(k)$, EXACT Lanczos → $\mathbf{T_k}$
$\mathbb{C}^N$
A, FP Lanczos → $\mathbf{T_k}$
the first k steps

Figure 2.2: For any $k = 1, 2, \ldots$, the first k steps of the finite precision Lanczos computation for $A \in \mathbb{C}^{N\times N}$ can be analyzed as the first k steps of the exact Lanczos for the (possibly much larger) matrix $\widehat{A}(k) \in \mathbb{C}^{\widehat{N}(k)\times\widehat{N}(k)}$ which generates the same $k \times k$ Jacobi matrix T_k. The matrix $\widehat{A}(k)$ depends on k.

eigenvalues in *tight clusters around each eigenvalue of* A; see also the detailed argumentation in [21] and [19, Section 5.9]. The appropriate starting vector associated with $\widehat{A}$ can be constructed from A and b independently of k. We illustrate this on the example from the beginning of this chapter and in Figure 2.3 we show the results of exact CG computations applied to the matrix where each eigenvalue of A was replaced by five eigenvalues clustered uniformly in the interval of width $\Delta = 2{\cdot}10^{-13}$ and where associated right hand side $\widehat{b}$ is obtained from b by decomposition of each individual entry into 5 equal parts such that $\|b\|^2 = \|\widehat{b}\|^2$; see [11].

2.2 Maximal attainable accuracy

The approximate solutions in iterative computation can not, in general, reach arbitrary accuracy and it has no sense to continue computations after the maximal attainable accuracy has been reached. The value of the maximal attainable accuracy can strongly depend on the algorithmic realization and the mathematically equivalent algorithms can behave differently in finite precision computations. An example is given in [3], where four different algorithms for solving the system of

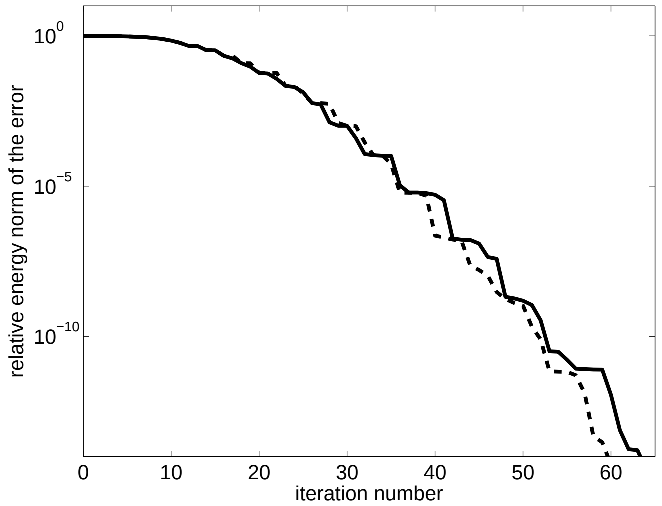


Figure 2.3: The convergence behaviour of finite precision CG computations (solid line) applied to the matrix with single eigenvalues corresponds with the behaviour of exact CG computations (dashed line) applied to the matrix with clusters of eigenvalues.

the normal equations $A^TAx = A^Tb$ are compared and in [12], where the accuracy of three-term recurrences and two-term recurrences associated with Krylov subspace solvers is examined. In particular, considering the CG method, two two-term recurrences should be preferred to the three-term recurrence.

Studying the difference of true and iterative residual, some bounds of the maximal attainable accuracy can be obtained; see [20, Chapter 6], the survey [14] and [21, Section 5.4]. However, in most practical applications, the computations are stopped before the maximal attainable accuracy is reached. In the numerical experiments in this thesis we assume that the iteration is stopped before the maximal attainable accuracy is reached.

3. Tracking the trajectory of finite precision CG computations

The CG method determines in exact arithmetic an orthogonal basis of the Krylov subspace $\mathcal{K}_k(A, r_0)$ given by the residuals r_j, $j = 0, 1, \ldots, k-1$. However, in finite precision CG computations the orthogonality among the computed residual vectors is (usually quickly) lost and they often become even (numerically) linearly dependent. Consequently, the computed residual vectors may, at the step k, span a subspace of a dimension smaller than k. This *rank-deficiency* of the computed Krylov subspace bases then causes a *delay* of convergence of finite precision computations, which can be defined as the difference between the number of iterations required to attain a prescribed accuracy in finite precision computations and the number of iterations required to attain the same accuracy in exact arithmetic.

To illustrate this behaviour, we reproduce and extend the experiment from [19, Figure 5.17] and compare the convergence curves of the energy norm of the errors of exact[1] and finite precision (FP) CG computations applied to a matrix A given by **Spectrum 1-Q** (see Appendix A) with the parameters

$$N = 25, \quad \lambda_1 = 0.1, \quad \lambda_N = 100, \quad \rho = 0.65$$

and a right-hand side b of ones. In Figure 3.1 we observe the characteristic staircase-like behaviour of FP CG computations (dash-dotted line with circles) and, in comparison with exact CG computations (solid line with circles), a substantial delay of convergence which is caused by the loss of orthogonality among

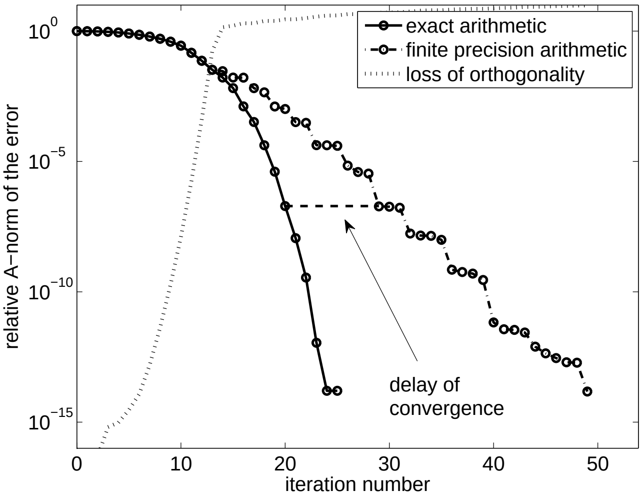


Figure 3.1: Convergence curve of finite precision CG computation (dash-dotted line with circles) is delayed in comparison with the convergence curve for exact CG computation (solid line with circles). The orthogonality among the computed residual vectors (dotted line), measured by $\|I - V_k^* V_k\|_F$ where columns of V_k are Lanczos vectors $v_j = r_j/\|r_j\|$, is lost after a few iterations.

[1] The exact arithmetic is hereinafter simulated by full double reorthogonalization of the computed residuals; see [26, p. 252], [11].

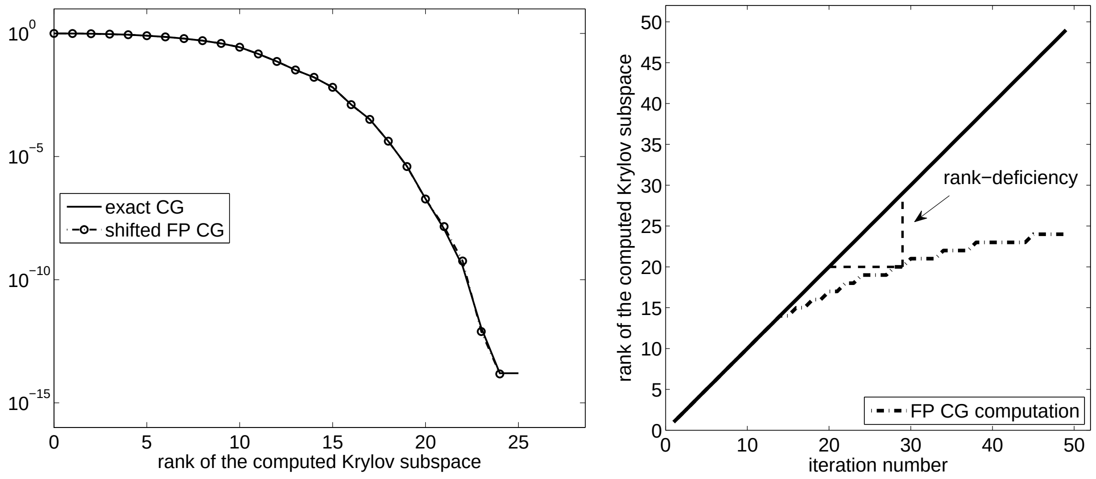


Figure 3.2: Left: Convergence curve of FP CG computation is shifted horizontally correspondingly to the rank-deficiency of the computed Krylov subspace (dash-dotted line with circles) and compared with the convergence curve of exact CG computations (solid line). Visually, the curves coincide to each other. Right: The delay of convergence is equal to the rank-deficiency of the computed Krylov subspace, i.e., to the quantity $k - \text{rank}(\mathcal{K}_k(A, r_0))$, which can be seen as the vertical difference between the solid and dash-dotted lines.

the computed residual vectors (dotted line) and subsequent rank-deficiency of the computed Krylov subspace.

The correspondence between the delay of convergence and the rank-deficiency of the computed Krylov subspace is illustrated in the left part of Figure 3.2 where the exact CG convergence curve is compared with the curve for finite precision CG computation which is shifted by the (numerical) rank-deficiency of the computed Krylov subspace (dash-dotted line with circles). More specifically, the curve is composed of the points

$$\left(\text{rank}(\mathcal{K}_k(A, r_0));\ \frac{\|x - x_k\|_A}{\|x - x_0\|_A}\right), \quad k = 1, 2, \ldots \tag{3.1}$$

(where x_k denotes the k-th approximation from finite precision CG computation) such that for every value $\bar{k} = 1, 2, \ldots$ it plots the point corresponding to the latest iteration k where

$$\bar{k} = \text{rank}(\mathcal{K}_k(A, r_0)) \tag{3.2}$$

and thus it reflects the merit of the CG method, i.e., the minimalization of A norm of the error over the given subspace. Throughout this chapter, $\text{rank}(\mathcal{K}_k(A, r_0))$ is given by the number of singular values of the matrix $[v_1, v_2, \ldots, v_k]$ (computed Lanczos vectors) which are greater or equal to 10^{-1}, i.e., we apply the MATLAB command `rank` with the threshold 10^{-1} to the matrix $[v_1, v_2, \ldots, v_k]$. Setting a different threshold value would not change the main point, but the observed correspondence would not be so perfect. The computed rank of the Krylov subspace $\mathcal{K}_k(A, r_0)$ is plotted in the right part of Figure 3.2 by the dash-dotted line and the rank-deficiency $k - \text{rank}(\mathcal{K}_k(A, r_0))$, which gives the delay of convergence, can be seen as the vertical difference between the dash-dotted and solid lines.

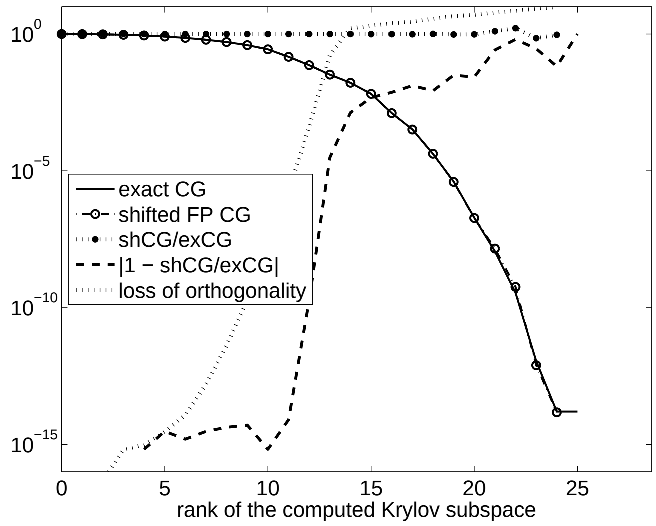


Figure 3.3: The line with large dots plots the ratio of energy norms of the error of shifted FP CG (dash-dotted line with circles) and exact CG (solid line). Its variance from the ideal value one is plotted by dashed line and determines the quality of the correspondence of both curves. The correspondingly shifted curve of loss of orthogonality in finite precision CG computations is plotted as dotted line.

The quality of the tight correspondence between the two curves from the left part of Figure 3.2 is illustrated in Figure 3.3. The line with dots plots the ratio of both curves, i.e., it is generated by points

$$\left(\bar{k} = \operatorname{rank}(\mathcal{K}_k(A, r_0));\ \frac{\|x - x_k\|_A}{\|x - \bar{x}_{\bar{k}}\|_A}\right), \quad \bar{k} = 1, 2, \ldots \tag{3.3}$$

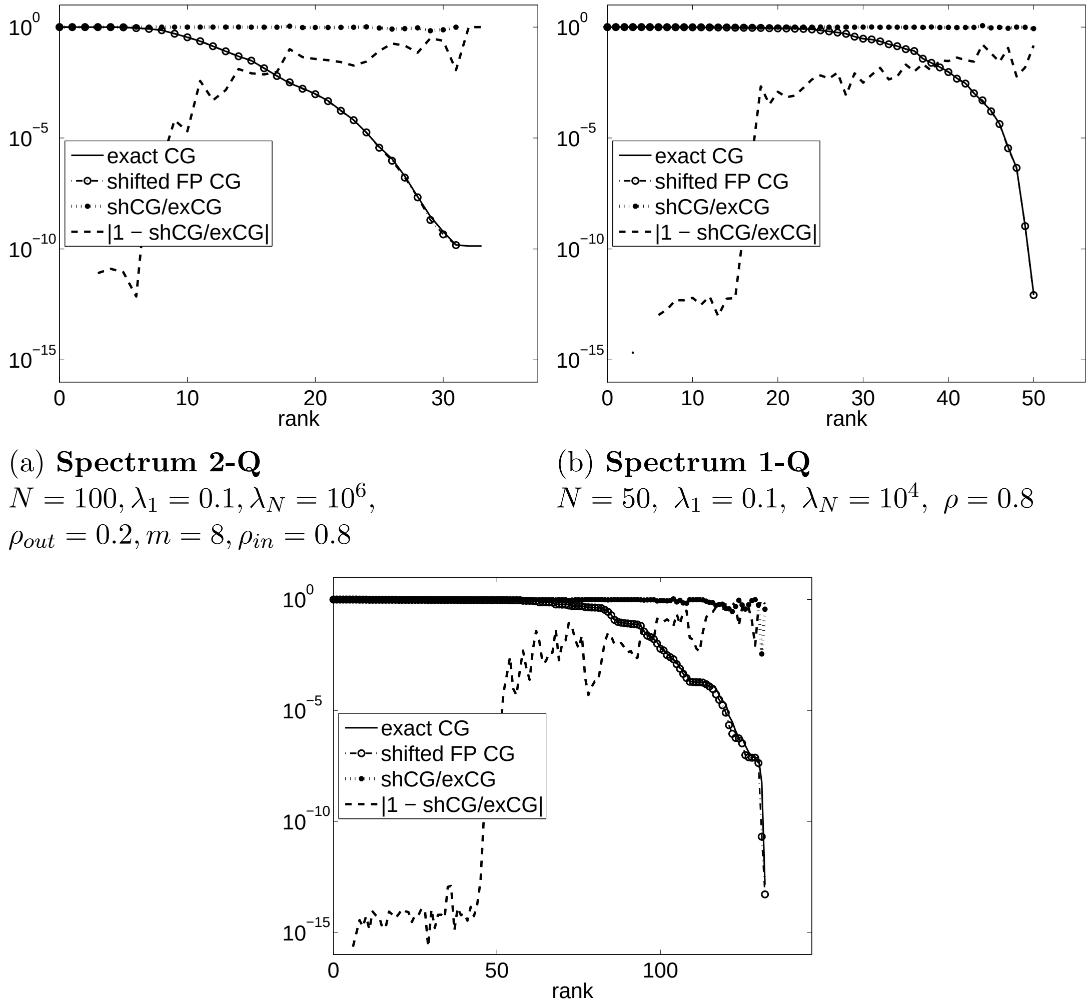
where $\bar{x}_{\bar{k}}$ is the $\bar{k}$-th exact CG approximation. Ideally, their ratio would be one and thus both curves would be identical. This would mean that, in terms of the energy norm, finite precision CG computations follow exactly the trajectory of exact CG computations and the movement along the trajectory is delayed by the rank-deficiency of the computed Krylov subspace. We see that, till the very end of computation, the line with dots is indeed very close to the ideal value one. The level of perturbation from this value, i.e., the curve

$$\left(\bar{k} = \operatorname{rank}(\mathcal{K}_k(A, r_0));\ \left|1 - \frac{\|x - x_k\|_A}{\|x - \bar{x}_{\bar{k}}\|_A}\right|\right), \quad \bar{k} = 1, 2, \ldots. \tag{3.4}$$

(dashed line) gives us a detail insight into the quality of correspondence between the curves for exact CG (denoted as exCG) and shifted FP CG computations (shCG). We see that the perturbations are of lower order, in other words,

$$\left|\frac{\|x - \bar{x}_{\bar{k}}\|_A - \|x - x_k\|_A}{\|x - \bar{x}_{\bar{k}}\|_A}\right| \ll 1 \tag{3.5}$$

holds throughout the computation, i.e., the difference between both curves is of lower order in comparison with the actual size of error. Thus we can consider the trajectory of energy norm of finite precision CG computations being enclosed



(a) **Spectrum 2-Q**
$N = 100, \lambda_1 = 0.1, \lambda_N = 10^6,$
$\rho_{out} = 0.2, m = 8, \rho_{in} = 0.8$

(b) **Spectrum 1-Q**
$N = 50,\ \lambda_1 = 0.1,\ \lambda_N = 10^4,\ \rho = 0.8$

(c) **Bcsstk04** (default setting)

Figure 3.4: The observed phenomenon is illustrated for various input data. Dotted line plots the ratio of energy norms of shifted FP CG (dash-dotted line with circles) and exact CG (solid line). The variance from the ideal value one is plotted by dashed line.

in a narrow "tunnel" around the trajectory of energy norm of exact CG computations. The narrowness of the tunnel comes from (3.5), i.e., if (3.5) holds, the width of the tunnel is small even in comparison with the corresponding size of the error and thus the energy norm of finite precision CG computations follows indeed very closely the trajectory of energy norm of exact CG computations. At the early stage, the dashed line in Figure 3.3 (i.e., the curve (3.4)) remains on the level of machine precision. This corresponds to the first stage of computations where the rounding errors cause no substantial loss of orthogonality and the computed Krylov subspace has full rank. The correspondingly shifted curve for loss of orthogonality in finite precision CG computations, i.e., the curve generated by points

$$(\text{rank}(\mathcal{K}_k(A, r_0));\ \|I - V_k^* V_k\|_F)\,, \quad k = 1, 2, \ldots. \tag{3.6}$$

is plotted in Figure 3.3 as a dotted line.

In order to demonstrate the phenomenon observed above on more examples, we show in Figure 3.4 analogous plots for various input data. The first matrix is given by the problem **Bcsstk04** from the database `MatrixMarket` and the spectra of the other two were adopted from the paper enclosed in Appendix; see Figure 3.4 for particular settings. The right-hand side b is always a vector of ones. The message from all these examples is still the same:

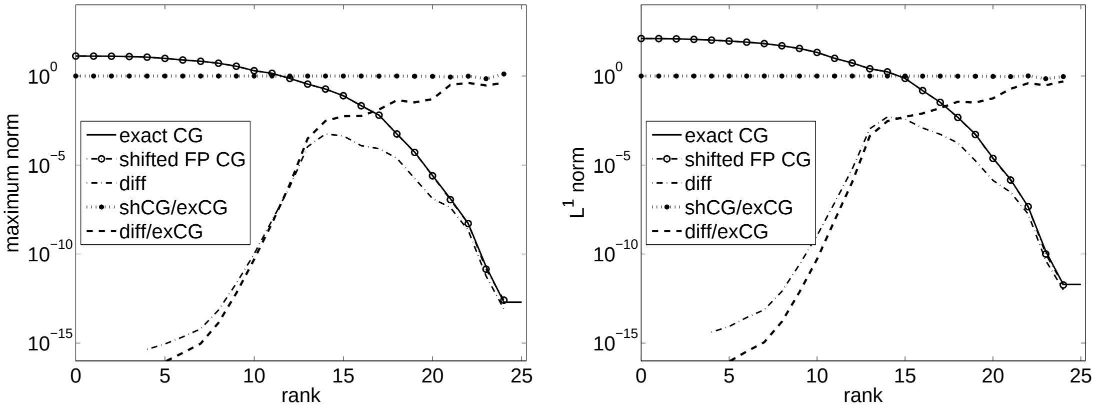
> The trajectory of energy norm of the error from finite precision CG computations follows tightly the one from exact CG computations but the process is delayed by the rank-deficiency of the computed Krylov subspace.

We would like to point out that the comparison presented in this section is somewhat different from the concept in Section 2.1 where finite precision CG computations are associated with exact CG computations for different matrix $\widehat{A}$ of specific structure and substantially larger size than the size of A. That concept enables us to associate the k-th error $\|x - x_k\|_A$ of finite precision CG computations with the k-th error $\|\widehat{x} - \widehat{x}_k\|_{\widehat{A}}$ of exact CG computations for $\widehat{A}$ but the approximation vectors x_k and $\widehat{x}_k$ live in completely different spaces; see the illustration in Figure 2.2. Conversely, the comparison made in this section associates the k-th error $\|x - x_k\|_A$ of finite precision CG computations with the $\bar{k}$-th error $\|x - \bar{x}_{\bar{k}}\|_A$ of exact CG computations for the same linear system where $\bar{k} = \mathrm{rank}(\mathcal{K}_k(A, r_0))$ and the rank-deficiency $k - \mathrm{rank}(\mathcal{K}_k(A, r_0))$ of the computed Krylov subspace is the delay of convergence. We would like to emphasize that since the vectors x_k and $\bar{x}_{\bar{k}}$ belong to the same space of dimension N, we can compare the approximation vectors themselves.

3.1 Correspondence of approximation vectors

The observation about the close correspondence of the trajectories of FP and exact CG computations from the previous section was formulated in terms of the energy norm. However, since the approximation vectors x_k and $\bar{x}_{\bar{k}}$ belong to the same vector space, we can study the correspondence of the trajectories of FP and exact CG computations in terms of closeness of the approximation vectors x_k and $\bar{x}_{\bar{k}}$ themselves.

In Figure 3.5, using the input data from the beginning of the previous section, we study the difference $x_k - \bar{x}_{\bar{k}}$ (dash-dotted line) in comparison with the error vectors $x - x_k$ (dash-dotted line with circles) and $x - \bar{x}_{\bar{k}}$ (solid line) using L^∞ (maximum) norm on the left and L^1 norm on the right. Similarly as in the left part of Figure 3.2, the curve of the error or finite precision CG computations shifted back by the level of rank-deficiency is visually indistinguishable from the curve of the error of exact CG computations and thus their ratio (line with large dots) remains throughout the computation process very close to the value one. On the other hand, we can see that the norm $\|x_k - \bar{x}_{\bar{k}}\|_\infty$ (resp. $\|x_k - \bar{x}_{\bar{k}}\|_1$) remains throughout the computation process below the norm of the errors $\|x - \bar{x}_{\bar{k}}\|_\infty$ and $\|x - x_k\|_\infty$ (resp. $\|x - \bar{x}_{\bar{k}}\|_1$ and $\|x - x_k\|_1$), i.e., we observe that the approximations x_k and $\bar{x}_{\bar{k}}$ are closer to each other than to the solution x. This is in more detail illustrated by the dashed line which plots the ratio of the norm of the difference



Spectrum 1-Q ($N = 25,\ \lambda_1 = 0.1,\ \lambda_N = 100,\ \rho = 0.65$)

Figure 3.5: Comparison in L^∞ (maximum) norm on the left, resp. in L^1 norm on the right. The difference of the approximations $x_k - \bar{x}_{\bar{k}}$ (dash-dotted line) remains throughout computation smaller than the corresponding errors $x - x_k$ and $x - \bar{x}_{\bar{k}}$ of FP CG (dash-dotted line with circles) and exact CG (solid line). The closeness of x_k and $\bar{x}_{\bar{k}}$ in comparison with the actual size of the error $x - \bar{x}_{\bar{k}}$ is illustrated by the ratio of their norms (dashed line). The ratio of norms of the errors (line with large dots) is throughout computation very close to the value one.

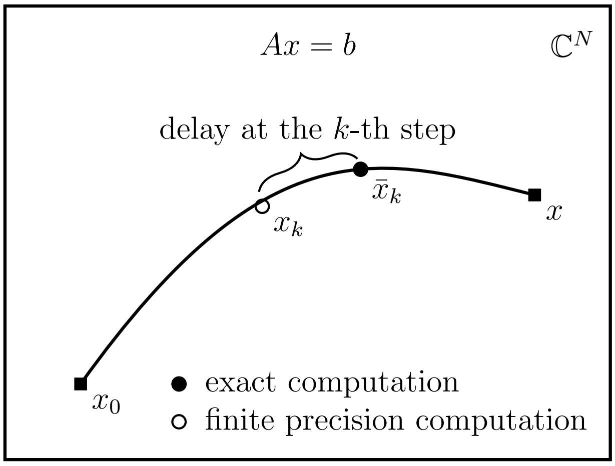


Figure 3.6: Finite precision CG computation tightly follows the trajectory of exact CG computations with the delay which is given by the rank-deficiency of the computed Krylov subspace.

between approximations and the norm of the corresponding error of exact CG computations. We see that

$$\frac{\|x_k - \bar{x}_{\bar{k}}\|_\infty}{\|x - \bar{x}_{\bar{k}}\|_\infty} \ll 1 \quad \text{resp.} \quad \frac{\|x_k - \bar{x}_{\bar{k}}\|_1}{\|x - \bar{x}_{\bar{k}}\|_1} \ll 1 \tag{3.7}$$

holds throughout the computation process, i.e., the distance between approximations is of lower order in comparison with the actual level of error.

Thus, based on the results of our numerical experiments (see also Figure 3.8 below), we can formulate a similar observation as in the previous section, but in terms of approximation vectors themselves:

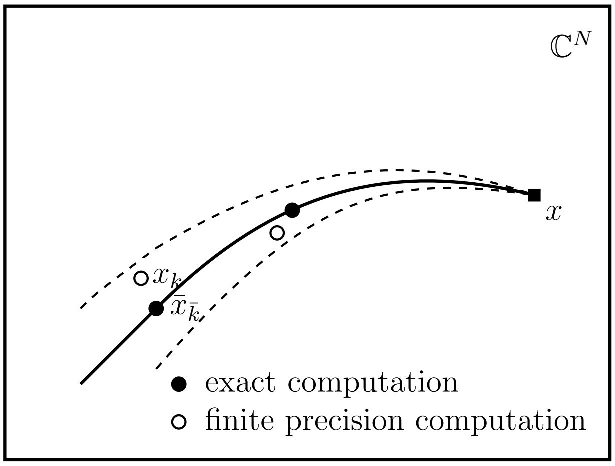


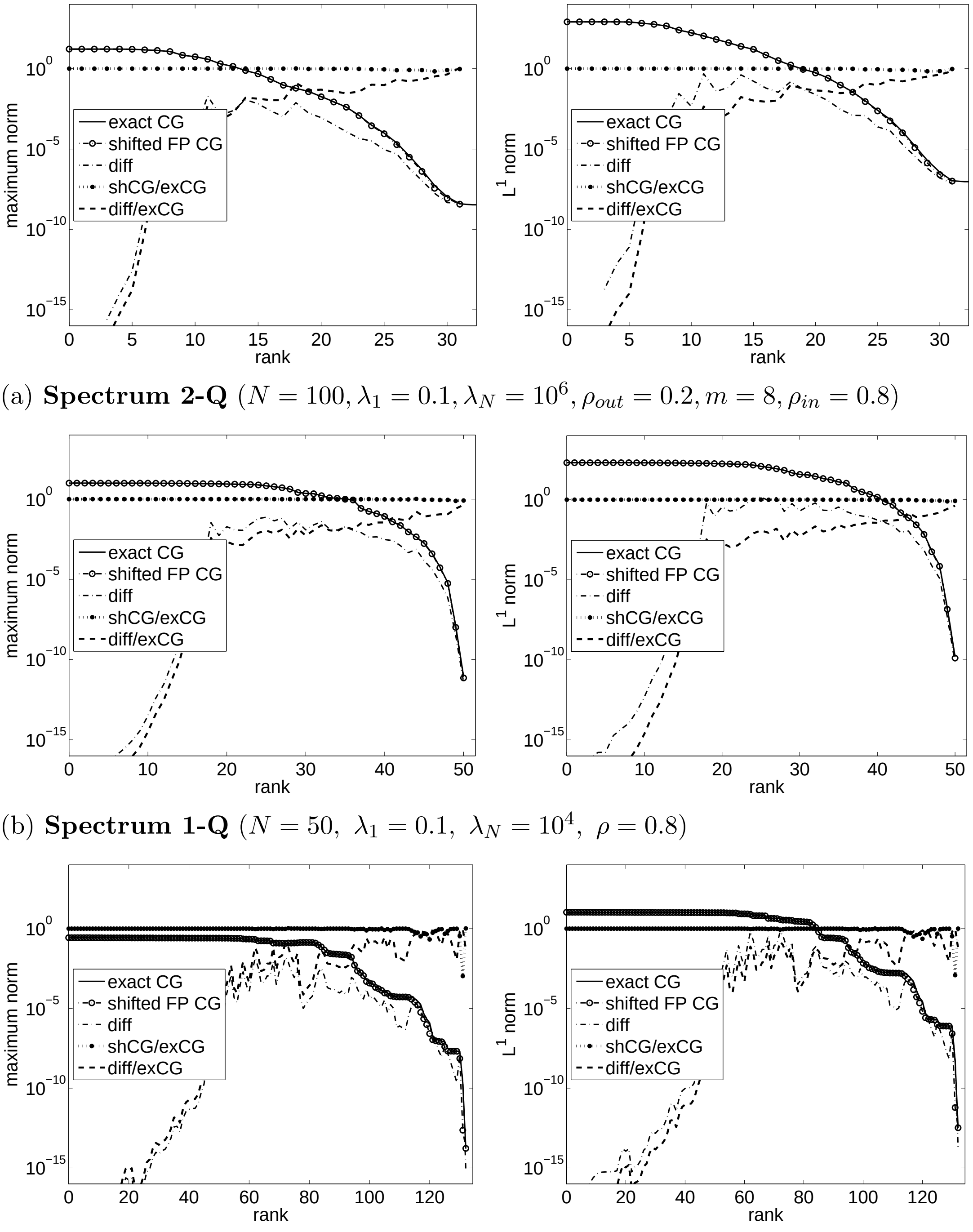
Figure 3.7: The trajectory of finite precision CG computation is enclosed in the narrow tunnel around the trajectory of exact CG computation. The tunnel is tight in the sense that its width is throughout the computation very small in comparison with the actual level of error.

> The trajectory of approximation vectors generated by the CG method in finite precision arithmetic applied to linear system $Ax = b$ is enclosed in a narrow "tunnel" around the trajectory of approximation vectors from the CG method in exact arithmetic applied to the same system. The progress of finite precision CG computations through this tunnel is delayed by the rank-deficiency of the computed Krylov subspace; for illustration see Figure 3.6. The tunnel is narrow in the sense that its width (i.e., the distance between approximations x_k and $\bar{x}_{\bar{k}}$) is very small in comparison with its "length" (i.e., with the actual level of error $x - \bar{x}_{\bar{k}}$) throughout the whole process of computation; for illustration see Figure 3.7.

As in the previous section, in order to illustrate the ideas presented above on more examples, we show in Figure 3.8 analogous plots using the same set of input data as above (see Figure 3.8(a)–(c) for particular settings). The results are very similar to those in Figure 3.5 and thus support the observation formulated above.

3.2 Correspondence of Krylov subspaces

In experiments above we have observed that finite precision CG computations follow closely the trajectory of exact CG computations, we have seen that the approximation vector x_k generated in the k-th iteration of the CG method in finite precision corresponds to the $\bar{k}$-th approximation $\bar{x}_{\bar{k}}$ of the exact CG method where $\bar{k} = \text{rank}(\mathcal{K}_k(A, r_0))$. The CG approximations are formed within the sequence of nested Krylov subspaces and thus the close relation between x_k and $\bar{x}_{\bar{k}}$ suggests that also the whole subspaces $\mathcal{K}_k(A, r_0)$ and $\overline{\mathcal{K}}_{\bar{k}}(A, r_0) = [\bar{v}_1, \ldots, \bar{v}_{\bar{k}}]$, where the latter is the $\bar{k}$-th Krylov subspace generated by the CG algorithm in exact arithmetic and $\bar{v}_{\bar{j}}$, $1, \ldots, \bar{k}$ are the corresponding exact Lanczos vectors, should be in some sense close to each other. We study the closeness of those sub-



(a) **Spectrum 2-Q** ($N = 100, \lambda_1 = 0.1, \lambda_N = 10^6, \rho_{out} = 0.2, m = 8, \rho_{in} = 0.8$)

(b) **Spectrum 1-Q** ($N = 50,\ \lambda_1 = 0.1,\ \lambda_N = 10^4,\ \rho = 0.8$)

(c) **Bcsstk04** (default setting)

Figure 3.8: The phenomenon observed above is illustrated on our set of different input data. Line with large dots plots the ratio of norms of the errors of shifted FP CG (dash-dotted line with circles) and exact CG (solid line). The norm of difference of approximation vectors x_k, $\bar{x}_{\bar{k}}$ is plotted as dash-dotted line and its ratio with the norm of the error of exact CG computations is plotted as dashed line.

spaces by comparing their (numerical) rank with the (numerical) rank of subspace of their union $\mathcal{K}_k(A, r_0) \cup \overline{\mathcal{K}}_{\bar{k}}(A, r_0)$. The equality

$$\operatorname{rank}(\mathcal{K}_k(A, r_0)) = \operatorname{rank}(\mathcal{K}_k(A, r_0) \cup \overline{\mathcal{K}}_{\bar{k}}(A, r_0)) \qquad (3.8)$$

Table 3.1: The Krylov subspace generated in finite precision CG computations, the Krylov subspace of corresponding dimension associated with exact CG computations and the subspace spanned by their union are compared through their ranks. The fact that the union has the same rank as the individual subspaces implies that subspace $\mathcal{K}_k(A, r_0)$ and its counterpart from exact CG span the same space. The comparison is given for four different input data used in the previous sections; see the captions (a)–(d) and Appendix A for particular settings.

k	5	23	41	59	77	95	113	131	149
$\bar{k} = \text{rank}(\mathcal{K}_k(A, r_0))$	5	13	17	20	24	26	28	30	32
$\text{rank}(\overline{\mathcal{K}}_{\bar{k}}(A, r_0))$	5	13	17	20	24	26	28	30	32
$\text{rank}(\mathcal{K}_k(A, r_0) \cup \overline{\mathcal{K}}_{\bar{k}}(A, r_0))$	5	13	17	20	24	26	28	30	32

(a) **Spectrum 2-Q** ($N = 100, \lambda_1 = 0.1, \lambda_N = 10^6, \rho_{out} = 0.2, m = 8, \rho_{in} = 0.8$)

k	15	51	87	123	159	195	231	267	293
$\bar{k} = \text{rank}(\mathcal{K}_k(A, r_0))$	15	32	39	43	46	48	49	50	50
$\text{rank}(\overline{\mathcal{K}}_{\bar{k}}(A, r_0))$	15	32	39	43	46	48	49	50	50
$\text{rank}(\mathcal{K}_k(A, r_0) \cup \overline{\mathcal{K}}_{\bar{k}}(A, r_0))$	15	32	39	43	46	48	49	50	50

(b) **Spectrum 1-Q** ($N = 50,\ \lambda_1 = 0.1,\ \lambda_N = 10^4,\ \rho = 0.8$)

k	30	113	196	279	362	445	528	611	664
$\bar{k} = \text{rank}(\mathcal{K}_k(A, r_0))$	30	82	93	104	112	119	126	131	132
$\text{rank}(\overline{\mathcal{K}}_{\bar{k}}(A, r_0))$	30	82	93	104	112	119	126	131	132
$\text{rank}(\mathcal{K}_k(A, r_0) \cup \overline{\mathcal{K}}_{\bar{k}}(A, r_0))$	30	82	93	104	113	120	127	131	132

(c) **Bcsstk04** (default setting)

k	9	14	19	24	29	34	39	44	50
$\bar{k} = \text{rank}(\mathcal{K}_k(A, r_0))$	9	14	16	19	20	22	23	23	24
$\text{rank}(\overline{\mathcal{K}}_{\bar{k}}(A, r_0))$	9	14	16	19	20	22	23	23	24
$\text{rank}(\mathcal{K}_k(A, r_0) \cup \overline{\mathcal{K}}_{\bar{k}}(A, r_0))$	9	14	16	19	20	22	23	23	24

(d) **Spectrum 1-Q** ($N = 25, \lambda_1 = 0.1, \lambda_N = 100, \rho = 0.65$)

would mean that finite precision CG computation generates in its k-th iteration (numerically) the same subspace as which is spanned by the $\bar{k}$-th Krylov subspace from exact CG computations.

In Table 3.1, we summarize the results of this comparison for the four different input data from the previous sections; see Table 3.1(a)–3.1(d) and Appendix A for particular settings. The (numerical) rank of $\mathcal{K}_k(A, r_0) \cup \overline{\mathcal{K}}_{\bar{k}}(A, r_0)$ is given as in the previous sections, i.e., we apply the Matlab command `rank` with the threshold 10^{-1} to the matrix of Lanczos vectors $[v_1, \ldots, v_k, \bar{v}_1, \ldots, \bar{v}_{\bar{k}}]$. We observe that the agreement between the computed Krylov subspace $\mathcal{K}_k(A, r_0)$

and its exact arithmetic counterpart $\overline{\mathcal{K}}_{\bar{k}}(A, r_0)$ is indeed nearly perfect. The computed ranks coincide exactly throughout the computations (with the small inaccuracy in the problem **Bsstk04** from the `MatrixMarket`; see Table 3.1(c)), In other words, we observe that the sequence of subspaces being built-up in finite precision CG computations is numerically basically the same as the one which is built in exact CG computations. However, due to rounding errors, the process may be substantially delayed. In this sense, the Krylov subspace seems to be stable to the effect of rounding errors produced by the CG (Lanczos) algorithm in finite precision arithmetic.

To our best knowledge, the relation between the k-th Krylov subspace computed by the CG (Lanczos) algorithm in finite precision arithmetic and the Krylov subspace of the corresponding dimension $\bar{k}$ generated by exact CG and associated with the same data A and r_0 has not been addressed in literature and its rigorous theoretical analysis may represent the topic of our further research. The related problem of sensitivity of Krylov subspace to small perturbations was studied in several papers; see, e.g., [4, 16] or [27]. Using different approaches[2] and techniques, both papers [4, 27] study and measure the distance between subspaces $\overline{\mathcal{K}}_j(A, v)$ and $\overline{\mathcal{K}}_j(A + \Delta A, v)$, where ΔA goes through the class of perturbations small enough to ensure that $\overline{\mathcal{K}}_j(A+\Delta A, v)$ has full column rank. The sensitivity of the Krylov subspace to this class of perturbations can be measured by the condition number of Krylov subspace (see [4, Definition 3]), both papers contain expressions for this condition number and study the possibilities of its computation or estimation.

In general, the difference between the Krylov subspaces $\overline{\mathcal{K}}_j(A + \Delta A, v)$ and $\overline{\mathcal{K}}_j(A + \Delta A, v)$ can grow exponentially in dependence on the perturbation matrix ΔA and thus the condition number of Krylov subspace may be large. Having the computed Krylov subspace $\mathcal{K}_k(A, r_0)$ with the numerical rank $\bar{k} = \mathrm{rank}(\mathcal{K}_k(A, r_0))$, consider its subspace $\underline{\mathcal{K}_k(A, r_0)}$ of the mathematical rank $\bar{k}$ cleansed from the influence of the smallest $k - \bar{k}$ singular values and suppose there exists a perturbation matrix $\Delta_{CG}A$ such that $\underline{\mathcal{K}_k(A, r_0)} = \overline{\mathcal{K}}_{\bar{k}}(A+\Delta_{CG}A, r_0)$. The stability of the Krylov subspace to the rounding errors produced by finite precision CG computations observed above implies that such $\overline{\mathcal{K}}_{\bar{k}}(A+\Delta_{CG}A, r_0)$ would be very close to the Krylov subspace $\overline{\mathcal{K}}_{\bar{k}}(A, r_0)$ from exact CG computations and their distance would be, substantially smaller than suggested by the classical results for sensitivity of Krylov subspaces to class of small perturbations of the coefficient matrix. The existence and the structure of such perturbation matrix may be subject of our further research.

[2] Sensitivity results of [4] could be suitable for the methods based on the Arnoldi algorithm and the results of [27] could be convenient for the methods based on short recurrences. In the symmetric case, the results of both approaches coincide.

4. Composite polynomial convergence bounds in finite precision CG computations

The content of this chapter can be understood as a kind of extended abstract of the paper enclosed in Appendix B. We outline the main ideas and refer the reader to Appendix B.

4.1 Outline of the enclosed paper

In Section 1.4 we have revealed that the linear convergence bounds based on the Chebyshev polynomials have, apart from very special situations, a little in common with the practical rate of the CG method. The CG method is nonlinear (see, e.g., (1.33) and (1.55)) and its convergence tends to accelerate during the iteration process. In other words, it exhibits the so called superlinear convergence. Thus the linear Chebyshev bounds are typically highly pessimistic. In order to describe the superlinear convergence, Axelsson [1] and Jennings [15] considered in presence of m large outlying eigenvalues the composite polynomial

$$q_m(\lambda)\chi_{k-m}(\lambda)/\chi_{k-m}(0), \tag{4.1}$$

where $\chi_{k-m}(\lambda)$ denotes the Chebyshev polynomial of degree $k-m$ shifted to the interval $[\lambda_1, \lambda_{N-m}]$ and where the polynomial $q_m(\lambda)$ has the roots at the outlying eigenvalues $\lambda_{N-m+1}, \ldots, \lambda_N$. This results in the bound

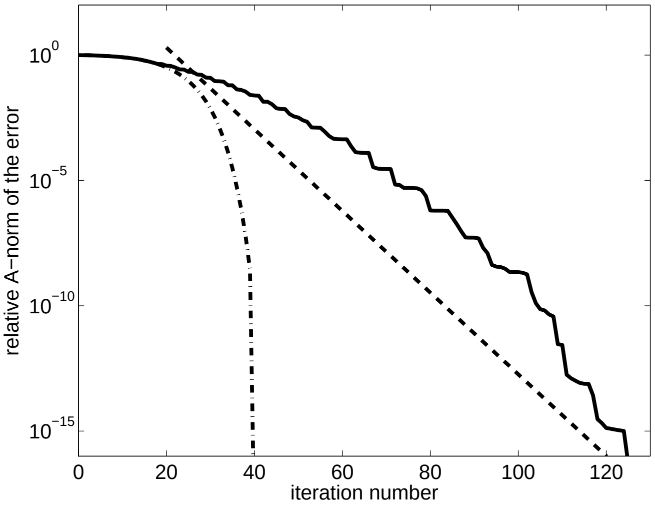


Figure 4.1: Rounding errors can cause a substantial delay of convergence in finite precision CG computations (solid line) in comparison to their exact precision counterpart (dash-dotted line). A composite polynomial bound (dashed line) fails to describe the finite precision CG behaviour *quantitatively* (the slope given by the bound is not descriptive) and *qualitatively* (the staircase-like shape of the convergence curve).

$$\frac{\|x - x_k\|_A}{\|x - x_0\|_A} \leq 2\left(\frac{\sqrt{\kappa_m(A)} - 1}{\sqrt{\kappa_m(A)} + 1}\right)^{k-m}, \quad k = m, m+1, \ldots, \tag{4.2}$$

where $\kappa_m(A) \equiv \lambda_{N-m}/\lambda_1$ is the so-called *effective condition number*. This quantity is typically substantially smaller than the condition number $\kappa(A) \equiv \lambda_N/\lambda_1$ which indicates a possibly faster convergence after m initial iterations.

However, as we have seen in the previous chapters, finite precision CG computations can be substantially delayed. Moreover, such delays are pronounced, in particular, in the presence of large outlying eigenvalues. Since the polynomial convergence bounds (4.2) were derived assuming exact arithmetic, it is by no means clear whether the composite polynomial bounds and the conclusions based on them are valid in finite precision CG computations. Figure 4.1 indeed shows that the composite polynomial convergence bound can fail to describe the finite precision CG.

Although the difficulty has been repeatedly noticed (see [15, 31, 10, 19]), persisting misunderstandings reappear in literature. The paper enclosed in Appendix B explains in detail that the composite polynomial bounds (4.2) must inevitably fail and that the composite polynomial (4.1) is not relevant for finite precision CG computations.

5. On some points of Gauss-Christoffel quadrature

In this chapter we will discuss several specific questions connected with the sensitivity of the Gauss-Christoffel quadrature approximation of the Riemann-Stieltjes integral with respect to small changes of the distribution function. Due to the correspondence between the CG method and the Gauss-Christoffel quadrature (see Section 1.2), this issue is closely related to the main focus of the thesis.

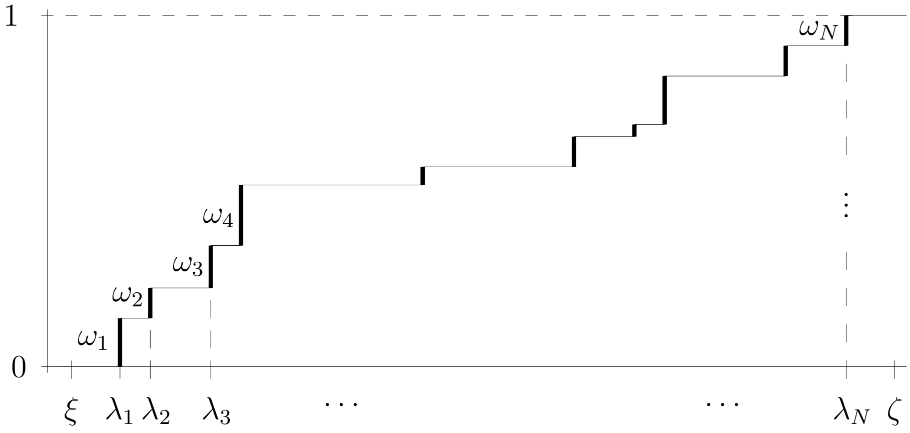


Figure 5.1: Illustration of the distribution function $\omega(\lambda)$.

Consider a piecewise constant nondecreasing distribution function $\omega(\lambda)$ with N points of increase $0 < \xi < \lambda_1 < \ldots < \lambda_N < \zeta$ and $\omega_1, \ldots, \omega_N$ the corresponding weights, i.e.,

$$\omega(\lambda) = \begin{cases} 0 & \text{for} \quad \xi \le \lambda < \lambda_1 \\ \sum_{j=1}^{i} \omega_j & \text{for} \quad \lambda_i \le \lambda < \lambda_{i+1}, \quad i = 1, \ldots, N-1 \\ \sum_{j=1}^{N} \omega_j = 1 & \text{for} \quad \lambda_N \le \lambda < \zeta \end{cases} \tag{5.1}$$

(see the illustration in Figure 5.1). Given a sufficiently small parameter δ, we will consider modified distribution function $\omega_\delta(\lambda)$ where each single point of increase λ_i is replaced by a cluster of s points of increase $\lambda_{i,1}, \ldots, \lambda_{i,s}$ which are uniformly distributed in the interval $[\lambda_i - \delta, \lambda_i + \delta]$ and where the corresponding weight ω_i is decomposed into s equal parts $\omega_{i,l}$, $l = 1, \ldots, s$; see the illustration in Figure 5.2. The modification parameter δ is set small enough to ensure that from the "bird's eye view" the distribution function $\omega_\delta(\lambda)$ matches with $\omega(\lambda)$. In other words, we demand preservation of interval which contains all points of increase and a clear separation of individual clusters, i.e.,

$$\delta \ll \lambda_1 \quad \text{and} \quad \delta \ll \frac{\lambda_{j+1} - \lambda_j}{2}, \; j = 1, \ldots, N-1. \tag{5.2}$$

Please note that for $\delta = 0$ the distribution functions coincide, i.e., $\omega(\lambda) = \omega_0(\lambda)$. As we will see, the important thing is that this modification changes the size of the support (i.e., the size of the set of all points of increase; see [8, p. 3]) of the original distribution function.

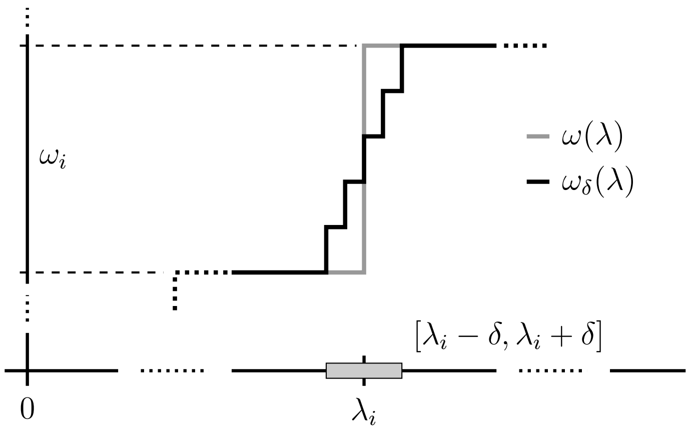

ω_i
— $\omega(\lambda)$
— $\omega_\delta(\lambda)$
$[\lambda_i - \delta, \lambda_i + \delta]$
0
λ_i

Figure 5.2: Illustration of the modified distribution function $\omega_\delta(\lambda)$.

In this chapter we are not concerned about the numerical aspects such as the effect of rounding errors or the existence of maximal attainable accuracy. We want to illustrate some mathematical phenomenons connected with the Gauss-Christoffel quadrature. Thus the numerical experiments below were performed using Matlab Symbolic Toolbox and using multiple precision arithmetic with 400 digits. Due to this, it is meaningful to run the experiments even with very small values of the parameter δ (such as 10^{-30}, 10^{-60}, $\ldots$). The value of the Riemann-Stieltjes integral $\int_\xi^\zeta f(\lambda)\, d\omega_\delta(\lambda)$ for $f(\lambda) = \lambda^{-1}$ is computed as the squared energy norm of the solution $x_\delta = A_\delta^{-1} b$, obtained via the Matlab backslash operator, where

$$\begin{aligned} A_\delta &= \operatorname{diag}(\lambda_{1,1}, \ldots, \lambda_{1,s}, \lambda_{2,1}, \ldots, \lambda_{N,s}) \\ b &= (\sqrt{\omega_{1,1}}, \ldots, \sqrt{\omega_{1,s}}, \sqrt{\omega_{2,1}}, \ldots, \sqrt{\omega_{N,s}}). \end{aligned} \tag{5.3}$$

The error of the k-point Gauss-Christoffel quadrature approximation of the integral is computed as the squared energy norm of the error in the k-th iteration of the exact[1] CG method (see (1.53) in Section 1.3) applied to the linear system (5.3) with the initial guess $x_0 \equiv 0$. The k-point Gauss-Christoffel approximation is (when needed) computed as the difference between the integral and the k-th error.

In the numerical experiments below we use for different values of the parameter δ the two following test problems with the distribution functions $\omega_\delta(\lambda)$:

Test 1; the distribution function $\omega_\delta^{(1)}(\lambda)$:

The original function $\omega^{(1)}(\lambda)$: Weights $\omega_i = 1/N$; points of increase λ_i given by **Spectrum 1**$(N = 25, \lambda_1 = 0.1, \lambda_N = 100, \rho = 0.5)$; see Appendix A.

The modified distribution function: $s = 2$, i.e., $\omega_\delta^{(1)}(\lambda)$ has two points of increase $\lambda_{i,1} = \lambda_i - \delta$ and $\lambda_{i,2} = \lambda_i + \delta$ instead of each point λ_i of $\omega^{(1)}(\lambda)$.

[1] As before, the exact arithmetic is simulated by double full reorthogonalization of the residual vectors; see [26, p. 252], [11].

Test 2; the distribution function $\omega_\delta^{(2)}(\lambda)$:

The original function $\omega^{(2)}(\lambda)$: Weights $\omega_i = 1/N$; points of increase λ_i given by **Spectrum 1**$(N = 20, \lambda_1 = 0.1, \lambda_N = 1000, \rho = 0.7)$; see Appendix A.

The modified distribution function: $s = 5$.

5.1 Sensitivity of Gauss-Christoffel quadrature

The problem of computing the Gauss-Christoffel quadrature has been studied by many researchers, the current state-of-the-art is well described in the monograph [8] by Gautschi and in the surveys [7, 18]. However, the question of sensitivity of the Gauss-Christoffel quadrature with respect to a small change of the associated distribution function has been, to our knowledge, firstly raised in the recent paper [22].

Given sufficiently smooth function $f(\lambda)$ and two distribution functions $\omega(\lambda)$, $\widehat{\omega}(\lambda)$ which are nondecreasing on a finite interval $[\xi, \zeta]$ and which are, in some sense, close to each other, the paper [22] investigates the k-point Gauss-Christoffel quadrature approximations

$$I_\omega^k = \sum_{j=1}^{k} \vartheta_j f(t_j) \quad \text{and} \quad I_{\widehat{\omega}}^k = \sum_{j=1}^{k} \widehat{\vartheta}_j f(\widehat{t}_j) \tag{5.4}$$

of the integrals

$$I_\omega \equiv \int_\xi^\zeta f(\lambda)\, d\omega(\lambda) \quad \text{and} \quad I_{\widehat{\omega}} \equiv \int_\xi^\zeta f(\lambda)\, d\widehat{\omega}(\lambda). \tag{5.5}$$

where $t_1, \dots, t_k$ (resp. $\widehat{t}_1, \dots, \widehat{t}_k$) and $\vartheta_1, \dots, \vartheta_k$ (resp. $\widehat{\vartheta}_1, \dots, \widehat{\vartheta}_k$) are the corresponding quadrature nodes and weights. Although it seems natural that the Gauss-Christoffel quadrature approximations (5.4) of the same degree should be close to each other, the paper reveals that it is not true in the case when the change of the distribution function affects the size of its support. It may happen that the difference between the Gauss-Christoffel quadrature approximations I_ω^k and $I_{\widehat{\omega}}^k$ of the same degree become, for several values k, much larger than the difference between the integrals I_ω and $I_{\widehat{\omega}}$. There are several different algorithms computing the Gauss-Christoffel quadrature and the paper [22] emphasizes that this sensitivity phenomenon is independent of the particular choice of the algorithm and that it is not caused by the effect of rounding errors. The paper also concludes that the sensitivity is observable for discrete (i.e., piecewise constant), continuous or even analytic distribution functions and for analytic integrands.

We illustrate this sensitivity in Figure 5.3 by reproducing the experiment from [22, Section 2]. We use the distribution function $\omega^{(1)}(\lambda)$ and the modified distribution function $\omega_\delta^{(1)}(\lambda)$ with $\delta = 10^{-8}$ and omit further in this section the superscript (1). We study the Gauss-Christoffel approximations I_ω^k and $I_{\omega_\delta}^k$ of the integrals I_ω and I_{ω_δ} for the smooth integrand $f(\lambda) = \lambda^{-1}$. In the top of Figure 5.3 we plot the errors $|E_\omega^k| = |I_\omega - I_\omega^k|$ (dashed line) and $|E_{\omega_\delta}^k| = |I_{\omega_\delta} - I_{\omega_\delta}^k|$ (solid line) and in the bottom we plot the difference between the integrals $|I_\omega - I_{\omega_\delta}|$

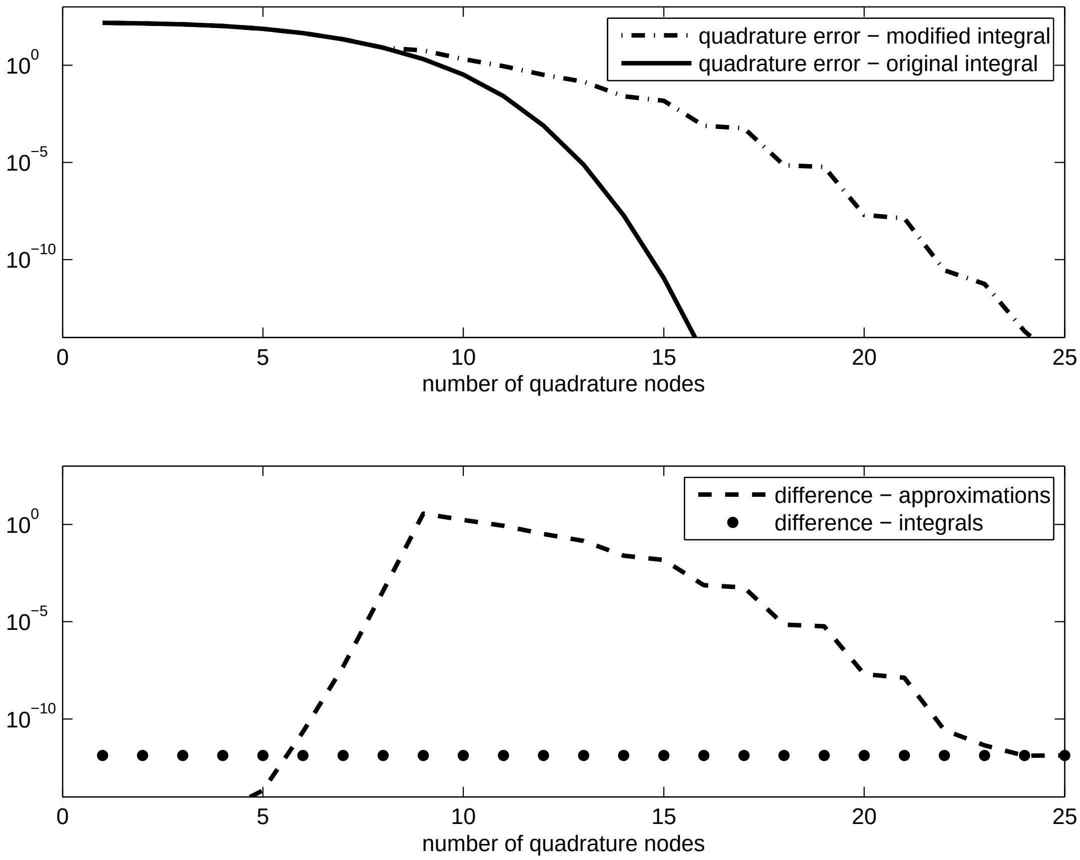


Figure 5.3: Illustration of the sensitivity of the Gauss-Christoffel quadrature to the small change of the distribution function. The top graph shows the errors of the k-th Gauss-Christoffel quadrature approximations for $f(\lambda) = \lambda^{-1}$ corresponding to the piecewise constant distribution function $\omega(\lambda)$ (solid line) and its modification $\omega_\delta(\lambda)$ with doubled points of increase (dashed line). The bottom graph shows the size of the difference between these k-point approximations (dash-dotted line) and the difference between the approximated integrals $\int_\xi^\zeta \lambda^{-1}\, d\omega(\lambda)$ and $\int_\xi^\zeta \lambda^{-1}\, d\omega_\delta(\lambda)$ (dots).

(dots) and the difference between the k-point Gauss-Christoffel approximations $|I_\omega^k - I_{\omega_\delta}^k|$ (dash-dotted line). Whereas the difference between the integrals I_ω and I_{ω_δ} is of order 10^{-12}, the Gauss-Christoffel approximations start to differ and the size of their difference reaches order 1 for $k = 9$. After that, the difference is dominated by the quadrature error $E_{\omega_\delta}^k$ as it follows from the identity

$$I_\omega^k - I_{\omega_\delta}^k = (I_\omega - I_{\omega_\delta}) - E_\omega^k + E_{\omega_\delta}^k, \tag{5.6}$$

where the first and the second term is of lower order for $k \geq 9$.

This dramatic change in the approximations of the integral can be linked to a sensitivity of the corresponding orthogonal polynomials. Although the distribution functions $\omega(\lambda)$ and $\omega_\delta(\lambda)$ may seem to be very close, the corresponding systems of orthogonal polynomials are quite different. Some of the zeros of the orthogonal polynomials (i.e., the quadrature nodes) corresponding to $\omega_\delta(\lambda)$ start to accumulate near the largest points of increase and thus, in comparison with the orthogonal polynomials corresponding to $\omega(\lambda)$, fewer zeros are located near the other points of increase; see [22] for details. This phenomenon is closely related to the fact that the presence of clustered eigenvalues affects the rate of convergence of the CG method; see the explanation given in [31, 32] and the discussion in [19, Section 5.6.5]. Consequently, the results of Greenbaum [10] reviewed in Section 2.1 indicate a close correspondence between the sensitivity of the Gauss-Christoffel quadrature and the convergence properties of the CG and Lanczos methods in finite precision arithmetic.

5.2 Discontinuity in Gauss-Christoffel quadrature

In this section we will study the behaviour of the Gauss-Christoffel quadrature and its error $E^k_{\omega_\delta}$ for different choices of the parameter δ. We are motivated by the following observation: For any $\delta \neq 0$, the distribution function $\omega_\delta(\lambda)$ has $N \cdot s$ distinct points of increase and thus the integral I_{ω_δ} is exactly computed just by the Ns-point Gauss-Christoffel quadrature and not before, i.e., $E^{N\cdot s}_{\omega_\delta} = 0$ and $E^k_{\omega_\delta} \neq 0$ for any $k < N \cdot s$. Similarly, the distribution function $\omega(\lambda) \equiv \omega_0(\lambda)$ has N finite points of increase and thus $E^N_{\omega_0} = 0$. Thus, if we define $K(\delta)$ as the number of nodes of the Gauss-Christoffel quadrature needed to compute exact value of the integral I_{ω_δ}, we get the following kind of discontinuity: It holds that

$$N \cdot s = \lim_{\delta \to 0} K(\delta) \neq K(0) = N. \tag{5.7}$$

We illustrate this phenomenon in Figure 5.4 where on the left (resp. on the right) we plot the curves of the error E_{ω_δ} corresponding to the distribution function $\omega^{(1)}_\delta(\lambda)$ with $N = 25, s = 2$ (resp. $\omega^{(2)}_\delta(\lambda)$ with $N = 20, s = 5$) for several different values of the parameter δ. In agreement with the theoretical observation (5.7), we see the necessity of 50 (resp. 100) quadrature points for exact computation of the integral for distribution functions $\omega_\delta(\lambda)$ with $\delta \neq 0$ compared to 25 (resp. 20) points for $\delta = 0$.

On the other hand, the quadrature error $E^k_{\omega_\delta}$ is for any $k \leq Ns$ a continuous function of parameter δ. More precisely, it holds that

$$\lim_{\delta \to 0} E^k_{\omega_\delta} = E^k_{\omega_0}, \quad k = 0, 1, \ldots, N \tag{5.8}$$

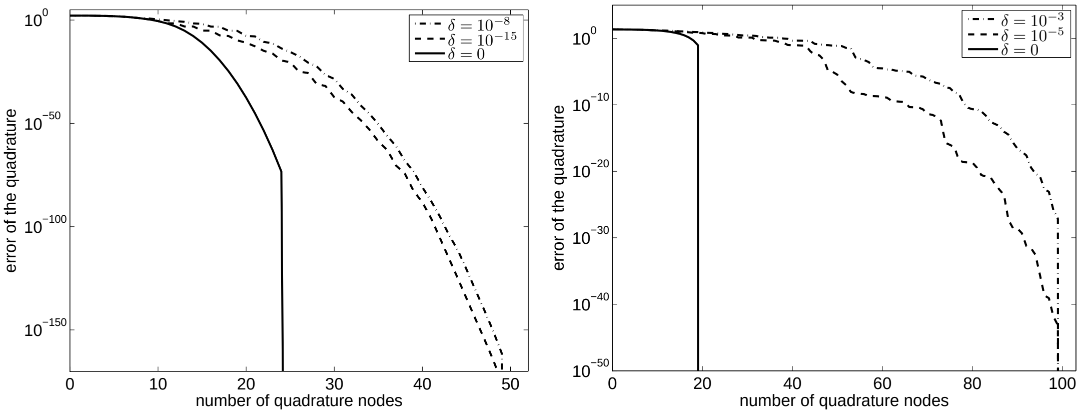
and, since $E^N_{\omega_0} = 0$ and the error $E^k_{\omega_\delta}$ is strictly decreasing in k,

$$\lim_{\delta \to 0} E^k_{\omega_\delta} = 0, \quad k = N, N+1, \ldots, N \cdot s. \tag{5.9}$$

In other words, the curve of the error E^k_ω, $k = 1, \ldots, N$ can be approximated to arbitrary precision by choosing sufficiently small parameter δ. More formally, for any $\epsilon > 0$, we can find δ such that $|E^k_\omega - E^k_{\omega_\delta}| < \epsilon$, $k = 1, \ldots, N$. The validity of (5.8) can be proved using the relation between the Gauss-Christoffel quadrature and the CG method. We know that

$$E^k_{\omega_\delta} = \|x_\delta - x_k(\delta)\|^2_{A_\delta} \tag{5.10}$$

where $x_\delta = A^{-1}_\delta b$, A_δ and b is from (5.3), and where $x_k(\delta)$ is the corresponding k-th CG approximation. The CG algorithm (see Algorithm I in page 7) gives



(a) **Test 1** — $\omega_\delta^{(1)}(\lambda)$ — $N = 25$, $s = 2$. (b) **Test 2** — $\omega_\delta^{(2)}(\lambda)$ — $N = 20$, $s = 5$.

Figure 5.4: Using two test problems, the convergence behaviour of the error $E^k_{\omega_\delta}$ of the k-point Gauss-Christoffel quadrature approximation is compared for the nonzero and zero parameter δ. The integral I_{ω_δ} is computed exactly by the Gauss-Christoffel quadrature iff its nodes are equal to the points of increase of $\omega_\delta(\lambda)$. Thus $E^N_{\omega_0} = 0$ while for any $\delta \neq 0$ the error remains nonzero till the Ns-point quadrature $I^{Ns}_{\omega_\delta}$.

that the formula for the approximation $x_k(\delta)$ is a composition of a finite number of continuous functions (denominators are strictly positive for $k \leq N$) and thus $x_k(\delta) \to x_k(0)$ which, together with the continuity with respect to δ of the scalar product $y^* A_\delta z$, gives (5.8).

This continuity is illustrated in Figure 5.5 on the two test problems with distribution functions $\omega_\delta^{(1)}(\lambda)$ in Figure 5.5(a) and $\omega_\delta^{(2)}(\lambda)$ in Figure 5.5(b). In the left part we see that for any given number of quadrature nodes k the error $E^k_{\omega_\delta}$ approach for decreasing values of δ to the value of the error E^k_ω (or to the value 0 for $k > N$). In the right figure we plot the error $E^k_{\omega_\delta}$ for several values of k as a function of δ and we observe that for sufficiently small values of δ the error $E^k_{\omega_\delta}$ drops to the level of the error E^k_ω.

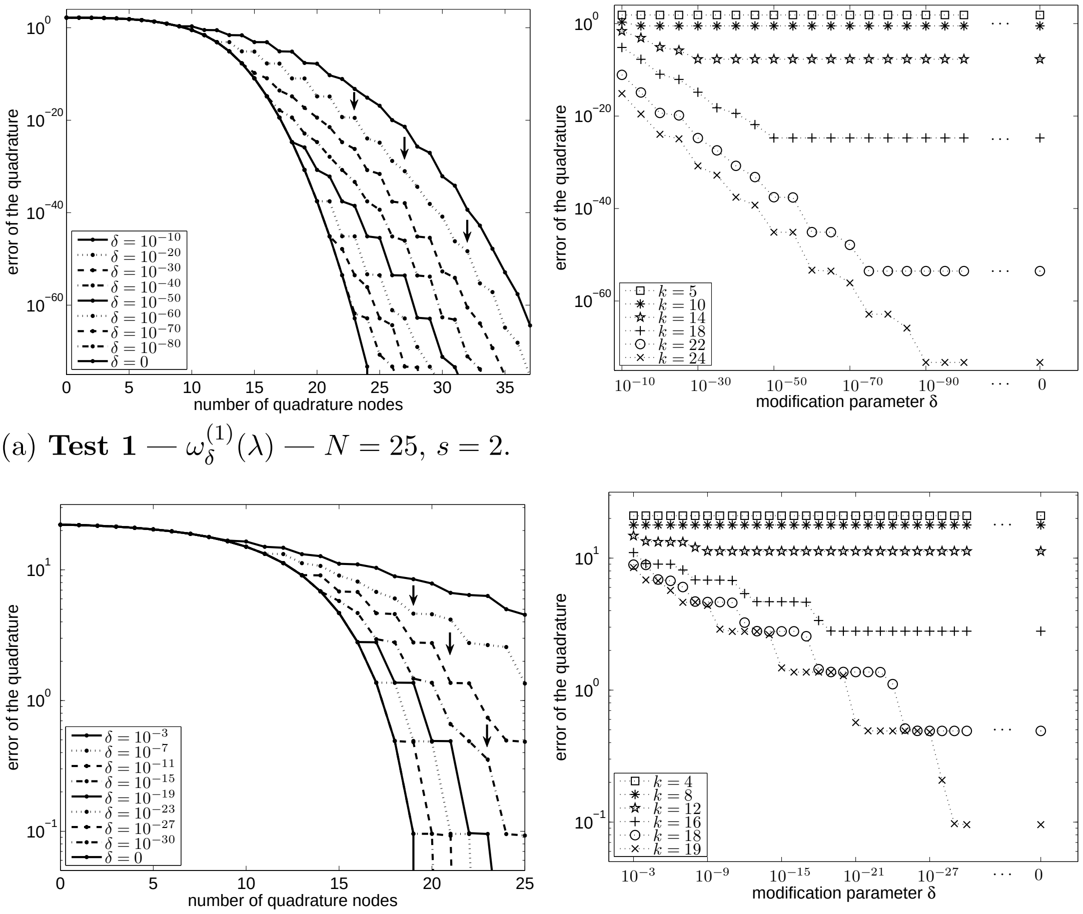
An immediate consequence of the continuity of the k-th error is the following observation: Let us define $K_\eta(\delta)$ as the number of quadrature nodes needed to decrease the quadrature error below the given level $\eta > 0$, i.e.,

$$K_\eta(\delta) = k \quad \text{iff} \quad E^k_{\omega_\delta} < \eta \leq E^{k-1}_{\omega_\delta}. \tag{5.11}$$

Then it holds

$$\lim_{\delta \to 0} K_\eta(\delta) = K_\eta(0), \quad \eta > 0, \tag{5.12}$$

cf., (5.7). In other words, for any level of accuracy $\eta > 0$ we can find sufficiently small parameter δ such that the integral I_{ω_δ} associated with the modified distribution function $\omega_\delta(\lambda)$ with clustered points of increase is approximated up to level of accuracy η by the same number of quadrature nodes as the integral I_ω associated with the distribution function $\omega(\lambda)$ with single points of increase. Please note that it is not in contradiction with the sensitivity phenomenon described in the previous section where the modification parameter δ was fixed. The sufficiently



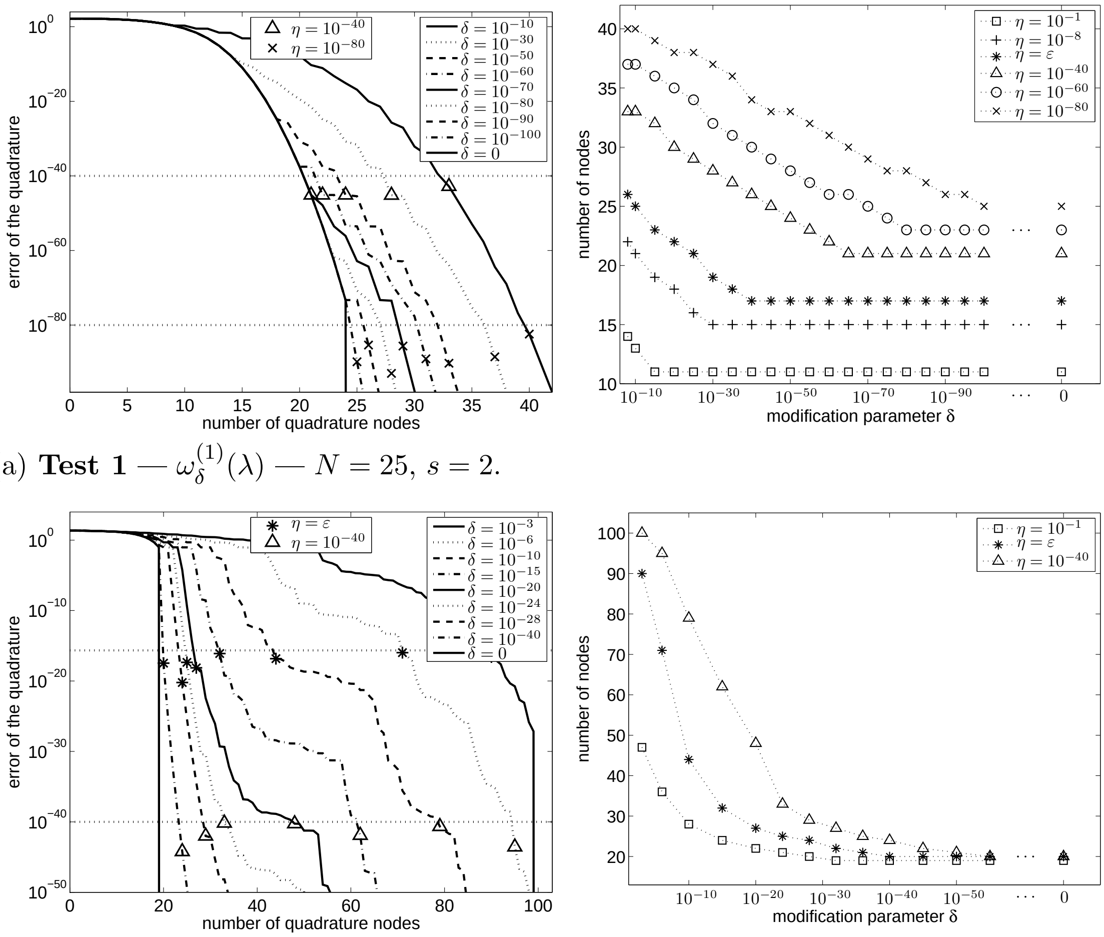
(a) **Test 1** — $\omega_\delta^{(1)}(\lambda)$ — $N = 25$, $s = 2$.

(b) **Test 2** — $\omega_\delta^{(2)}(\lambda)$ — $N = 20$, $s = 5$.

Figure 5.5: The illustration of the continuity of the Gauss-Christoffel quadrature error with respect to the parameter δ. Left: The error $E^k_{\omega_\delta}$ of the k-point Gauss-Christoffel quadrature approximations of the integral I_{ω_δ} converge for $k \leq N$ with decreasing δ to the error E^k_ω and, for $k > N$, to the value 0. Right: For several values of k, we plot the process of convergence of the quadrature error to the error E^k_ω as a function of the parameter δ.

small δ here may be extremely (asymptotically) small and with no relevance to practical computations.

The validity of (5.12) is illustrated in Figure 5.6(a) for the test problem with the distribution function $\omega_\delta^{(1)}(\lambda)$ and in Figure 5.6(b) for $\omega_\delta^{(2)}(\lambda)$. In the left part of Figure 5.6, we plot the curves of the quadrature error for several different parameters δ and, on each of these curves, we emphasize the number of quadrature nodes sufficient to decrease the error below given levels of accuracy $\eta = 10^{-80}$ (crosses), $\eta = 10^{-40}$ (triangles) and $\eta = \varepsilon$ (stars) where $\varepsilon = 2^{-52}$ is the machine precision unit in double precision arithmetic. We can see that the number of quadrature nodes needed to suppress the error below the given level of accuracy $\eta > 0$, i.e., the value $K_\eta(\delta)$, decreases with the parameter δ. This dependence is illustrated in the right part of Figure 5.6 where we plot $K_\eta(\delta)$ as a function of δ for several different levels of accuracy $\eta > 0$. In correspondence with (5.12) we see that $K_\eta(\delta) \to K_\eta(0)$ for any considered level $\eta > 0$ and that for sufficiently



(a) **Test 1** — $\omega_\delta^{(1)}(\lambda)$ — $N = 25$, $s = 2$.

(b) **Test 2** — $\omega_\delta^{(2)}(\lambda)$ — $N = 20$, $s = 5$.

Figure 5.6: On the left, we plot the quadrature errors associated with the distribution function $\omega_\delta(\lambda)$ for several different values δ and we emphasize (by stars, triangles or crosses) the moment of reaching the given level of accuracy $\eta > 0$. We see that the number of needed nodes is decreasing with the parameter δ. This is illustrated in the right part where we plot the function $K_\eta(\delta)$ as a function of δ for several different values of the level of accuracy $\eta > 0$ and we see that, for sufficiently small δ, it reaches the value $K_\eta(0)$.

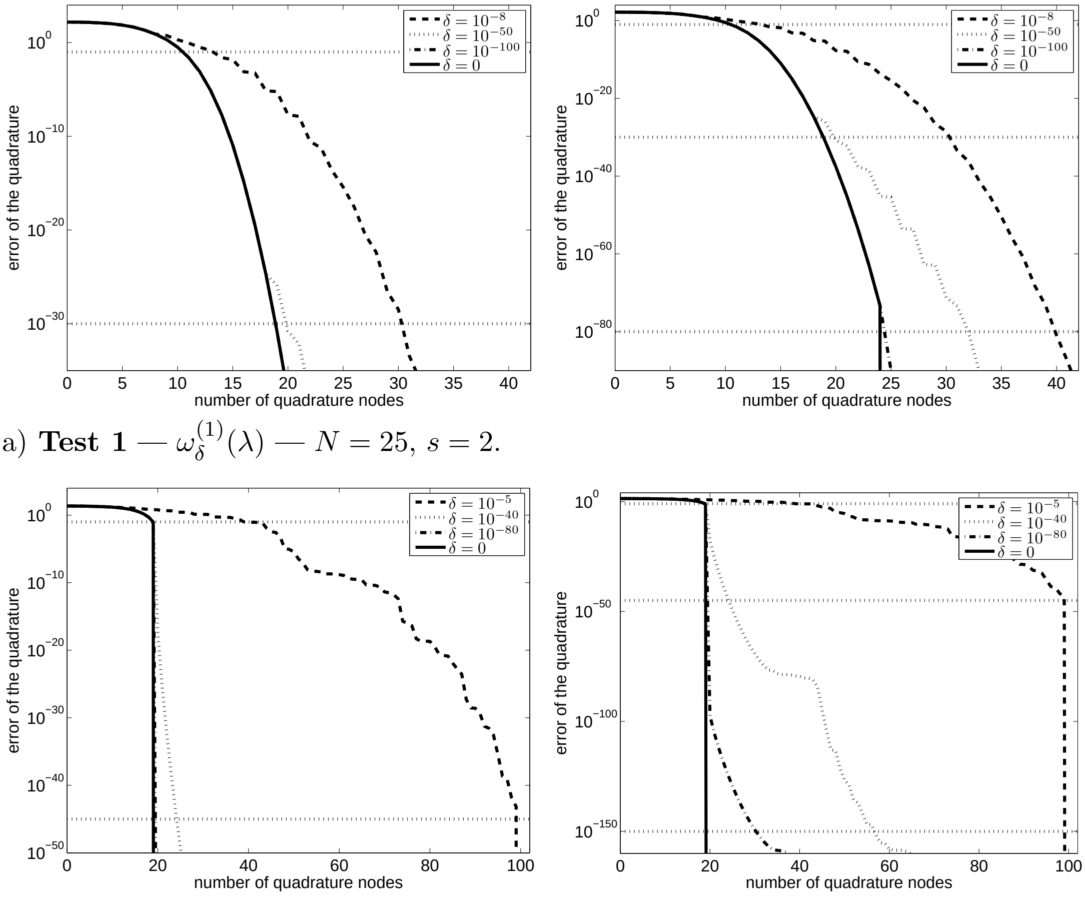
small δ the number of nodes sufficient to decrease the quadrature error $E^k_{\omega_\delta}$ below the level η is the same as for the quadrature error $E^k_{\omega_0}$.

The convergence behaviour of the error E_{ω_δ} of the Gauss-Christoffel quadrature approximations is conform with both statements (5.7) and (5.12), i.e., it holds

$$\lim_{\delta\to 0} K_\eta(\delta) = K_\eta(0),\ \eta > 0, \qquad \text{where}\ \ K_\eta(\delta) = k \ \ \text{iff}\ \ E^k_{\omega_\delta} < \eta \le E^{k-1}_{\omega_\delta},$$

$$N \cdot s = \lim_{\delta\to 0} K(\delta) \neq K(0) = N \qquad \text{where}\ \ K(\delta) = k \ \ \text{iff}\ \ E^k_{\omega_\delta} = 0.$$

This observation illustrates the trickiness and the possible weakness of the analysis of the convergence behaviour based on the asymptotic arguments. Given a particular parameter δ and a level of accuracy $\eta > 0$, it may be problematic



(a) **Test 1** — $\omega_\delta^{(1)}(\lambda)$ — $N = 25$, $s = 2$.

(b) **Test 2** — $\omega_\delta^{(2)}(\lambda)$ — $N = 20$, $s = 5$.

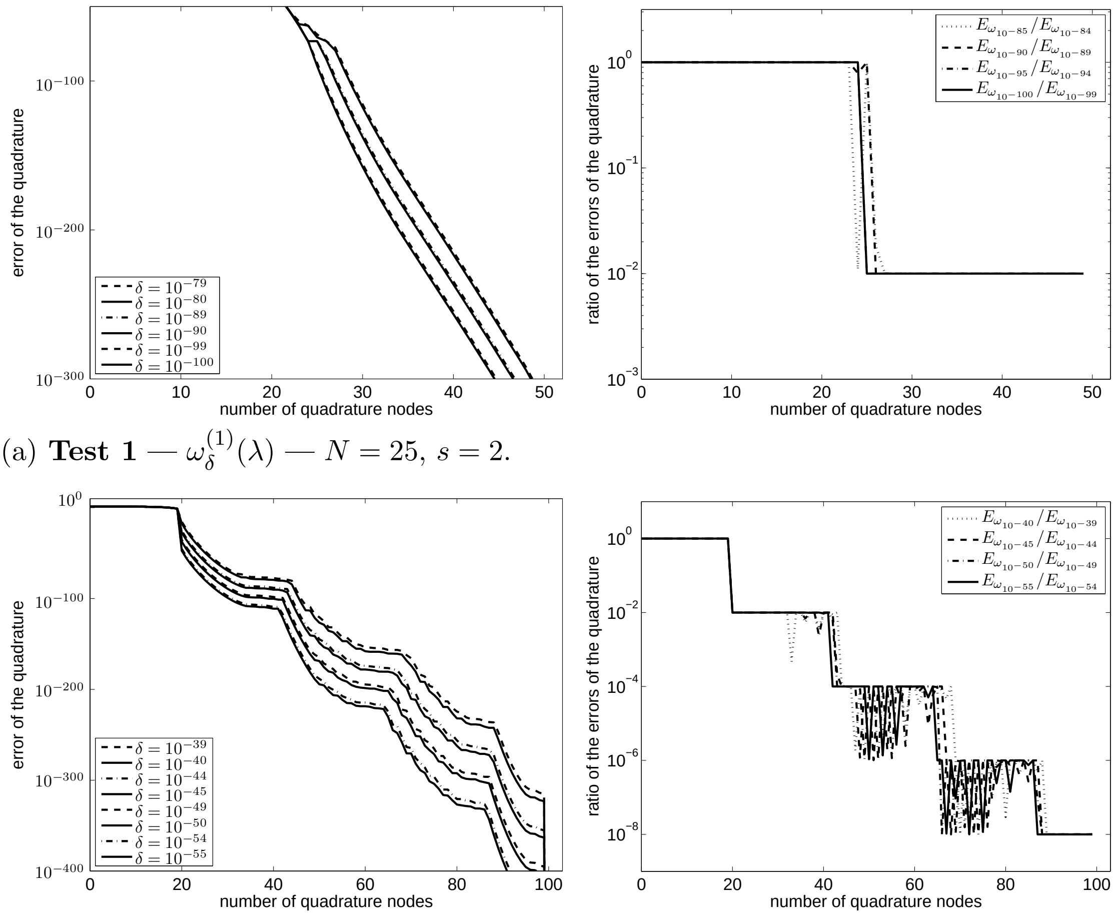
Figure 5.7: The illustration of the asymptotic convergence behaviour of the quadrature error. In dependence on the modification parameter δ and the given level of accuracy $\eta > 0$ (horizontal dotted lines), the number of quadrature nodes needed to decrease the quadrature error $E^k_{\omega_\delta}$ below the level η can be close to N as well as to $N \cdot s$.

to determine whether the number of nodes k needed to decrease the quadrature error $E^k_{\omega_\delta}$ below the level η is close to N or $N \cdot s$; see the illustration in Figure 5.7. In order to get more detail information about the asymptotic behaviour of the quadrature error for $\delta \to 0$, we study in Figure 5.8, on the example of our two test distribution functions $\omega_\delta^{(1)}(\lambda)$ and $\omega_\delta^{(2)}(\lambda)$, the rate of decrease of the error. We measure this rate by the ratio of the quadrature errors associated with two consecutive values of the parameter δ, i.e., in the right part of Figure 5.8(a) we plot four curves obtained by the points

$$\left(k;\ \frac{E^k_{\omega^{(1)}_{10^{-85}}}}{E^k_{\omega^{(1)}_{10^{-84}}}}, \frac{E^k_{\omega^{(1)}_{10^{-90}}}}{E^k_{\omega^{(1)}_{10^{-89}}}}, \frac{E^k_{\omega^{(1)}_{10^{-95}}}}{E^k_{\omega^{(1)}_{10^{-94}}}}, \frac{E^k_{\omega^{(1)}_{10^{-100}}}}{E^k_{\omega^{(1)}_{10^{-99}}}} \right), \quad k = 1, 2 \ldots, N \cdot s.$$

and in the right part of Figure 5.8(b) we plot four curves obtained by the points

$$\left(k;\ \frac{E^k_{\omega^{(2)}_{10^{-40}}}}{E^k_{\omega^{(2)}_{10^{-39}}}}, \frac{E^k_{\omega^{(2)}_{10^{-45}}}}{E^k_{\omega^{(2)}_{10^{-44}}}}, \frac{E^k_{\omega^{(2)}_{10^{-50}}}}{E^k_{\omega^{(2)}_{10^{-49}}}}, \frac{E^k_{\omega^{(2)}_{10^{-55}}}}{E^k_{\omega^{(2)}_{10^{-54}}}} \right), \quad k = 1, 2 \ldots, N \cdot s.$$



(a) **Test 1** — $\omega_\delta^{(1)}(\lambda)$ — $N = 25$, $s = 2$.

(b) **Test 2** — $\omega_\delta^{(2)}(\lambda)$ — $N = 20$, $s = 5$.

Figure 5.8: The shapes of the curves of the quadrature error (lines on the left) for sufficiently small parameters δ correspond to each other such that the ratios (lines on the right) of two quadrature errors for two consecutive values of δ significantly coincide. Moreover, we observe a strict separation of the individual stages where the number of these stages is equal to the parameter s (2 stages in (a) and 5 stages in (b)).

In the left part of Figure 5.8, we observe that for sufficiently small parameters δ, the shape of the curves of the quadrature error is very similar; see in particular the left part of Figure 5.8(b). The correspondence among the curves is so tight that the ratios plotted in the right part of Figure 5.8 significantly overlaps each other, i.e., the results of our numerical experiments indicate that asymptotically (i.e., for sufficiently small parameters δ) the error decreases with the same rate for different values of δ. Moreover, we observe a significant separation of the rate of decrease into several stages. The formulated observations seems valid only in the asymptotic sense (i.e., for sufficiently small parameters δ) while these phenomenons were not so visible in the numerical experiments with larger parameters $\delta \approx 10^{-10}$, 10^{-20}.

We consider this topic worth of further study. In particular, we would like to study the phenomenon of several stages of the rate of decrease, we strongly believe that the correspondence between the number of stages and the number of clustered points of increase is not accidental. Furthermore, it may be interesting to perform these experiments also with other integrands $f(\lambda)$.

Conclusion

This thesis is focused on the convergence behaviour of the CG method both in exact and finite precision arithmetic. We have briefly reviewed the close link of the CG method with the Lanczos method, orthogonal polynomials, the problem of moments and the Gauss-Christoffel quadrature. These relationships can help to understand that the CG method represents a highly nonlinear finite process and that its analysis requires mathematical tools different from those used for the linear stationary or semi-iterative methods. We have described in detail the fundamental difference between the CG method and the Chebyshev semi-iterative method and we have explained that the widespread linear convergence bound based on the extremal properties of the Chebyshev polynomials is relevant for the CSI method but, in general, has a little in common with the practical rate of convergence of the CG method.

We would like to emphasize that the CG method is computationally based on short recurrences and thus the analysis relevant to practical computations must take into account the delay of convergence caused by the loss of orthogonality among the computed direction vectors. We have briefly reviewed the theoretical results which enable to understand the mechanism of this delay of convergence. We have demonstrated that the rate of convergence of the CG method in finite precision can be substantially different from the rate of convergence of CG in exact arithmetic and we have shown that the composite polynomial convergence bounds based on explicit annihilation of the large outlying eigenvalues (which hold assuming exact arithmetic) must inevitably fail in finite precision CG computations.

Whereas the CG convergence *rate* may substantially differ in finite precision and exact computations, we have observed that the *trajectory* of the approximations or the energy norm of the error is very similar. We have shifted back the results of finite precision computations by the numerical rank-deficiency of the computed Krylov subspaces and we have observed close correspondence to the results of exact computations. Moreover, we have observed that the computed rank-deficient Krylov subspace span numerically nearly the same subspace as the Krylov subspace of the corresponding rank generated by the CG method in exact arithmetic.

This correspondence is worth of further study. This should include the derivation of a technique which would quantitatively measure the distance between the generated subspaces. We also intend to study the trajectories of finite precision and exact computations in other Krylov subspace methods. Together with the phenomenon of the sensitivity of the Gauss-Christoffel quadrature outlined in the last chapter, these questions may motivate our further research.

Bibliography

[1] Axelsson, O. A class of iterative methods for finite element equations. *Comput. Methods Appl. Mech. Engrg. 9*, 2 (1976), 123–127.

[2] Benzi, M. Preconditioning techniques for large linear systems: a survey. *J. Comput. Phys. 182*, 2 (2002), 418–477.

[3] Björck, Å., Elfving, T., and Strakoš, Z. Stability of conjugate gradient and Lanczos methods for linear least squares problems. *SIAM J. Matrix Anal. Appl. 19*, 3 (1998), 720–736.

[4] Carpraux, J.-F., Godunov, S. K., and Kuznetsov, S. V. Condition number of the Krylov bases and subspaces. *Linear Algebra Appl. 248* (1996), 137–160.

[5] Cipra, B. A. The best of the 20th century: Editors name top 10 algorithms. *SIAM News 33* (2000).

[6] Dongarra, J., and Sullivan, F. The Top 10 Algorithms (Guest editors' introduction). *Comput. Sci. Eng. 2*, 1 (2000), 22–23.

[7] Gautschi, W. A survey of Gauss-Christoffel quadrature formulae. In *E. B. Christoffel (Aachen/Monschau, 1979)*. Birkhäuser, Basel, 1981, pp. 72–147.

[8] Gautschi, W. *Orthogonal Polynomials: Computation and Approximation.* Numerical Mathematics and Scientific Computation. Oxford University Press, New York, 2004. Oxford Science Publications.

[9] Golub, G. H., and Strakoš, Z. Estimates in quadratic formulas. *Numer. Algorithms 8*, 2-4 (1994), 241–268.

[10] Greenbaum, A. Behaviour of slightly perturbed Lanczos and conjugate-gradient recurrences. *Linear Algebra Appl. 113* (1989), 7–63.

[11] Greenbaum, A., and Strakoš, Z. Predicting the behavior of finite precision Lanczos and conjugate gradient computations. *SIAM J. Matrix Anal. Appl. 13*, 1 (1992), 121–137.

[12] Gutknecht, M. H., and Strakoš, Z. Accuracy of two three-term and three two-term recurrences for Krylov space solvers. *SIAM J. Matrix Anal. Appl. 22*, 1 (2000), 213–229.

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

[14] Higham, N. J. *Accuracy and Stability of Numerical Algorithms*, second ed. SIAM, Philadelphia, PA, 2002.

[15] Jennings, A. Influence of the eigenvalue spectrum on the convergence rate of the conjugate gradient method. *J. Inst. Math. Appl. 20*, 1 (1977), 61–72.

[16] Kuznetsov, S. V. Perturbation bounds of the Krylov bases and associated Hessenberg forms. *Linear Algebra Appl. 265* (1997), 1–28.

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

[18] Laurie, D. P. Computation of Gauss-type quadrature formulas. *J. Comput. Appl. Math. 127*, 1-2 (2001), 201–217.

[19] Liesen, J., and Strakoš, Z. *Krylov subspace methods: principles and analysis.* Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, 2012.

[20] Meurant, G. *The Lanczos and conjugate gradient algorithms: from theory to finite precision computations.* vol. 19 of Software, Environments, and Tools. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2006.

[21] Meurant, G., and Strakoš, Z. The Lanczos and conjugate gradient algorithms in finite precision arithmetic. *Acta Numer. 15* (2006), 471–542.

[22] O'Leary, D. P., Strakoš, Z., and Tichý, P. On sensitivity of Gauss-Christoffel quadrature. *Numer. Math. 107*, 1 (2007), 147–174.

[23] Paige, C. C. *The computation of eigenvalues and eigenvectors of very large and sparse matrices.* PhD thesis, London University, London, England, 1971.

[24] Paige, C. C. Computational variants of the Lanczos method for the eigenproblem. *J. Inst. Math. Appl. 10* (1972), 373–381.

[25] Paige, C. C. Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix. *J. Inst. Math. Appl. 18*, 3 (1976), 341–349.

[26] Paige, C. C. Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem. *Linear Algebra and Its Applications 34* (1980), 235–258.

[27] Paige, C. C., and Van Dooren, P. Sensitivity analysis of the Lanczos reduction. *Numer. Linear Algebra Appl. 6*, 1 (1999), 29–50. Czech-US Workshop in Iterative Methods and Parallel Computing, Part I (Milovy, 1997).

[28] Saad, Y. *Iterative Methods for Sparse Linear Systems*, second ed. SIAM, Philadelphia, PA, 2003.

[29] Strakoš, Z. On the real convergence rate of the conjugate gradient method. *Linear Algebra Appl. 154/156* (1991), 535–549.

[30] Strakoš, Z., and Tichý, P. On error estimation in the conjugate gradient method and why it works in finite precision computations. *Electron. Trans. Numer. Anal. 13* (2002), 56–80.

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

[32] van der Sluis, A., and van der Vorst, H. A. The convergence behavior of Ritz values in the presence of close eigenvalues. *Linear Algebra Appl. 88/89* (1987), 651–694.

A. List of experimental data

Here we describe the matrices (or points of increase of piecewise constant distribution function) used in this thesis.

Spectrum 1($N, \lambda_1, \lambda_N, \rho$)

A diagonal matrix with the following spectrum. Given $N, \lambda_1 > 0,\ \lambda_N > 0$ and $\rho \in (0, 1]$ the inner eigenvalues are given by the formula

$$\lambda_i = \lambda_1 + \frac{i-1}{N-1}(\lambda_N - \lambda_1)\rho^{N-i} \quad i = 2, \ldots, N-1.$$

The parameter ρ determines the non-uniformity of the spectrum. For $\rho \ll 1$ the eigenvalues tend to cumulate near λ_1 and for $\rho = 1$ the spectrum is distributed uniformly. This type of spectrum was introduced in [29].

Spectrum 1-Q($N, \lambda_1, \lambda_N, \rho$)

Given a diagonal matrix Λ of type **Spectrum 1** and random unitary matrix Q, we define $A = Q\Lambda Q^T$.

Spectrum 2($N, \lambda_1, \lambda_N, \rho_{out}, m, \rho_{in}$)

Diagonal matrix generated in two steps:

1. Run **Spectrum 1**($N, \lambda_1, \lambda_N, \rho_{out}$)
2. Run **Spectrum 1**($N-m, \lambda_1, \lambda_{N-m}, \rho_{in}$) and thus rewrite $\lambda_2, \ldots, \lambda_{N-m-1}$.

This type of spectrum is used in the enclosed paper; see Appendix B.

Spectrum 2-Q($N, \lambda_1, \lambda_N, \rho_{out}, m, \rho_{in}$)

Given a diagonal matrix Λ of type **Spectrum 2** and random unitary matrix Q, we define $A = Q\Lambda Q^T$.

Bcsstk04

Matrix from the `MatrixMarket` database; see `http://math.nist.gov/MatrixMarket/`

Basic properties: symmetric positive definite; $N = 132$; $\kappa \approx 5.6e + 06$.

B. Paper

Composite convergence bounds based on Chebyshev polynomials and finite precision conjugate gradient computations

Tomáš Gergelits [†] · **Zdeněk Strakoš** [§]

Abstract The conjugate gradient method (CG) for solving linear systems of algebraic equations represents a *highly nonlinear finite process.* Since the original paper of Hestenes and Stiefel published in 1952, it has been linked with the Gauss-Christoffel quadrature approximation of Riemann-Stieltjes distribution functions determined by the data, i.e., with a simplified form of the *Stieltjes moment problem.* This link, developed further by Vorobyev, Brezinski, Golub, Meurant and others, indicates that a general description of the CG rate of convergence using an asymptotic convergence factor has principal limitations. Moreover, CG is computationally based on *short recurrences.* In finite precision arithmetic its behaviour is therefore affected by a possible loss of orthogonality among the computed direction vectors. Consequently, *any* consideration concerning the CG rate of convergence relevant to practical computations must include analysis of effects of rounding errors. Through the example of composite convergence bounds based on Chebyshev polynomials, this paper argues that the facts mentioned above should become a part of common considerations on the CG rate of convergence. It also explains that the spectrum composed of small number of well separated tight clusters of eigenvalues does not necessarily imply a fast convergence of CG or other Krylov subspace methods.

Keywords Conjugate gradient method · Stieltjes moment problem · Chebyshev semi-iterative method · Composite polynomial convergence bounds · Finite precision computations · Clusters of eigenvalues

Mathematics Subject Classification (2010) 65F10, 65B99, 65G50, 65N15

The work was supported by the ERC-CZ project LL1202, by the GACR grant 201/09/0917 and by the GAUK grant 695612.

[†] Charles University in Prague, Faculty of Mathematics and Physics, Sokolovská 83, 186 75 Prague, Czech Republic (email: `gergelits.tomas@seznam.cz`)

[§] Charles University in Prague, Faculty of Mathematics and Physics, Sokolovská 83, 186 75 Prague, Czech Republic (email: `strakos@karlin.mff.cuni.cz`)

B.1 Introduction

In this paper we consider the method of conjugate gradients (CG) {33} for solving linear algebraic systems $Ax = b$, where $A \in \mathbb{C}^{N\times N}$ is Hermitian and positive definite (HPD) matrix which is typically large and sparse. Given an initial guess x_0 and $r_0 = b - Ax_0$, the CG approximations x_k are uniquely determined by the relations

$$x_k \in x_0 + \mathcal{K}_k(A, r_0), \quad r_k \perp \mathcal{K}_k(A, r_0), \quad k = 1, 2, \ldots,$$

where $r_k = b - Ax_k$ is the k-th residual and

$$\mathcal{K}_k(A, r_0) \equiv \operatorname{span}\{r_0, Ar_0, \ldots, A^{k-1}r_0\}$$

is the k-th Krylov subspace associated with the matrix A and the vector r_0.

Apart from simple examples, CG can not be applied without *preconditioning*. Throughout this paper we assume that $Ax = b$ represents the preconditioned system. CG can be introduced in more general infinite dimensional Hilbert space settings; see, e.g. {Chapter III, Sections 2 and 4, 61}, {19, 62}, and also the recent descriptions using the Riesz map in, e.g., {38, 30}. Throughout this paper, the finite dimensional linear algebraic formulation will be sufficient. If A and b results from preconditioning of discretized operator equation (as in numerical solution of partial differential equations), then the preconditioning is often motivated by the operator context; see, e.g. {59, 22, 5, 34, 7, 52, 38}. In practical computations, preconditioning is incorporated into the algorithm and the preconditioned system $Ax = b$ is not formed. For an analytic investigation of the rate of convergence assuming exact arithmetic this difference is not important. In finite precision arithmetic, convergence is delayed due to the loss of orthogonality among the computed direction (residual) vectors. This can be conveniently demonstrated using the preconditioned system $Ax = b$ without going into further details on the particular preconditioning technique. An example of a detailed rounding error analysis can be found, e.g., in {56}.

B.1.1 CG, Gauss-Christoffel quadrature and the Stieltjes moment problem

Throughout the paper we assume that $A \in \mathbb{C}^{N\times N}$ is HPD with the spectral decomposition

$$A = U \operatorname{diag}(\lambda_1, \ldots, \lambda_N)\, U^*, \; U^*U = UU^* = I \tag{B.1}$$

where for simplicity of notation $0 < \lambda_1 < \ldots < \lambda_N$ and $U = [u_1, \ldots, u_N]$. Using this spectral decomposition, $v_1 \equiv r_0/\|r_0\|$ and $\omega_j \equiv |(v_1, u_j)|^2$, $j = 1, \ldots, N$, the moments of the distribution function $\omega(\lambda)$ determined by the nodes $\lambda_1, \ldots, \lambda_N$ and the weights $\omega_1, \ldots, \omega_N$ are given by

$$\sum_{j=1}^{N} \omega_j \lambda_j^k = v_1^* A^k v_1, \quad k = 0, 1, 2, \ldots. \tag{B.2}$$

The n-node Gauss-Christoffel quadrature of the monomials then determines the n nodes $\theta_l^{(n)}$ and weights $\omega_l^{(n)}$, $l = 1, \ldots, n$, of the distribution function $\omega^{(n)}(\lambda)$

such that the first $2n$ moments of the distribution function $\omega(\lambda)$ are matched, i.e.,

$$\sum_{l=1}^{n} \omega_l^{(n)} \{\theta_j^{(n)}\}^k = v_1^* A^k v_1, \quad k = 0, 1, 2, \dots, 2n. \tag{B.3}$$

Here the sums on the left hand sides of (B.2) and (B.3) can be expressed via the Riemann-Stieltjes integrals for the monomials with respect to the distribution functions $\omega(\lambda)$ and $\omega^{(n)}(\lambda)$ respectively.

As explained in {Section 3.5, 37} with references to many earlier publications, CG applied to $Ax = b$ with the initial residual r_0 can be understood as a process generating the sequence of the distribution functions $\omega^{(n)}(\lambda)$, $n = 1, \dots, N$ approximating the original distribution function $\omega(\lambda)$ in the sense of the Gauss-Christoffel quadrature. Equivalently, CG (implicitly) solves the (simplified) Stieltjes moment problem (B.2)–(B.3). The energy norm of the CG error is then given by

$$\|x - x_n\|_A^2 = \|r_0\|^2 \left(\sum_{j=1}^{N} \omega_j \lambda_j^{-1} - \sum_{l=1}^{n} \omega_l^{(n)} \{\theta_l^{(n)}\}^{-1} \right) \tag{B.4}$$

$$= \|r_0\|^2 \sum_{j=1}^{N} \prod_{l=1}^{n} \left(\frac{1}{\lambda_j^{1/2}} - \frac{\lambda_j^{1/2}}{\theta_l^{(n)}} \right)^2 \omega_j\,; \tag{B.5}$$

see {Section 5.6.1, Corollary 5.6.2 and Theorem 5.6.3, 37}. The nodes $\theta_l^{(n)}$ and the weights $\omega_l^{(n)}$ are the eigenvalues and the squared first components of the associated normalized eigenvectors of the Jacobi matrix T_n generated in the first n steps of the Lanczos process applied to the matrix A with the initial vector v_1. The matrix T_n represents the *operator* $A : \mathbb{C}^N \to \mathbb{C}^N$ restricted and orthogonally projected onto the n-th Krylov subspace $\mathcal{K}_n(A, r_0)$, which reveals the degree of nonlinearity with respect to A; see, e.g., {61, 12, 37}[1].

Recalling the previous facts prior to starting a discussion of a-priori bounds or estimates for the CG rate of convergence (based on some simplified information extracted from A and b) makes a good sense for the following reason. *Any* such bound or estimate has to deal with the tremendous *nonlinear complexity* of the expressions (B.4) and (B.5). Further details can be found, e.g., in {37, 24, 41}.

B.1.2 Comments on the a-priori analysis of the CG rate of convergence

A-priori analysis of the rate of convergence of CG (as well as of other Krylov subspace methods) focuses on certain relatively simple characteristics of the problem which can conveniently be linked (if applicable) with the underlying system of infinite dimensional operator equations, its preconditioning and discretisation. A condition number of the preconditioned discretized operator in combination with some information on large or small eigenvalues may serve as the most typical example of such characteristics. Following the functional analysis-based investigation in {45} as well as experimental observations, it is *assumed* that the rate

[1] The nonlinearity with respect to b has recently been studied in {26}.

of convergence follows the following three consecutive phases (see {Section 1.3, 45}):

> "*in the early sweeps the convergence is very rapid but then slows down, this is the sublinear behavior. The convergence then settles down to a roughly constant linear rate. ... Towards the end new speed may be picked up again, corresponding to the superlinear behavior.*"

Heuristic arguments on CG based on the spectrum of A are used to support this assumption (see also {Section 1, 6}). It should be taken into account, however, that this assumption and the supporting heuristics are based on experience with *some* spectral distributions. It can not be generalized to all practical problems. This is made clear in {45} by the sentence almost immediately following the quoted one given above:

> "*In practice all phases need not be identifiable, nor they appear only once and in this order.*"

The sublinear, linear and superlinear phases are analysed in literature using various tools; see, e.g., {45, 62, 6} or the survey in {Sections 2–4, 7}. Section 3.2 of {6} gives a nice example on how the reasoning about an initial sublinear phase can be applied in practice; see also {4}.

Applications of the results associated with particular phases to practical computations or to analysis of a particular problem requires verification whether the assumptions used in derivations are met in the given problems. Here the asymptotic reasoning requires a special attention. As stated in {p. 113, 22}:

> "*Methods with similar asymptotic work estimates may behave quite differently in practice*".

Krylov subspace methods are mathematically *finite*. Therefore, strictly speaking, in Krylov subspace methods there is no asymptotic present at all.

In relation to the last point it is sometimes argued in literature that due to rounding errors Krylov subspace methods do not terminate in a finite number of steps and *therefore* they are considered iterative methods which also justifies use of asymptotic bounds. In our opinion this point is not valid. First, effects of rounding errors depend on whether methods are implemented via short or long recurrences; see {Sections 5.9 and 5.10, 37}. Second, the standard CG implementation is based (for a good reason; see, e.g., the surveys in {41} and {31}) on coupled two-term recurrences. In finite precision arithmetic the orthogonality of the computed residuals (or direction vectors) can not be, in general, preserved, which results in a *delay of convergence.* The mechanism of this delay is well understood, and its consequences should not be interpreted as making the iteration process infinite.

This is immediately clear from the other effect of rounding errors, called *maximal attainable accuracy.* The accuracy of the computed approximate solution can not be improved below some level of the error determined by the implementation, computer arithmetic and the input data; see, e.g., {Section 7.3, 28}, {Section 5.4, 41}, {Section 5.9.3, 37} and the references given there. CG as well as other Krylov subspace methods are considered iterative because the iteration can be stopped whenever the user-specified accuracy is reached; see, e.g. {Section 2.4.2, 32} and {p. 450, 3}. The stopping criteria must be based on *a-posteriori* error analysis; see, e.g. {in particular Section 4.1, 1} for a recent survey of the context in adaptive numerical solution of elliptic partial differential equations, as well as {20} and {Appendix A, 3} for some early examples.

Throughout this paper we *assume* that the iteration is stopped before the maximal attainable accuracy is reached. Such assumption can not be taken in practical computations for granted. It must be justified by a proper numerical stability analysis (a simple *a-posteriori* check can be based on comparison of the iteratively and directly computed residuals). A detailed exposition of the related issues is out of the scope of this paper and we refer the interested reader to the literature given above.

In summary, *a-priori* analysis of the CG rate of convergence must take into account a possible delay of convergence due to rounding errors. Since in CG computations keeping short recurrences is essential, which inevitably results in a loss of orthogonality, *developing bounds or estimates which are to be applied to practical computations can not assume exact arithmetic.*

B.1.3 Analysis based on Chebyshev polynomials

In this paper we focus on the most common *a-priori* analysis of the CG convergence rate based on Chebyshev polynomials. The rate of convergence of CG is associated with linear convergence bounds derived using scaled and shifted Chebyshev polynomials in hundreds of papers and essentially in every textbook covering the CG method. As argued in Section B.1.1 above, the CG method and therefore also its convergence rate are, however, nonlinear and its convergence often tends to accelerate, with more or less pronounced variations, during the iteration process. Axelsson {2} and Jennings {35} suggested in this context composite polynomial bounds based on explicit annihilation of the outlying eigenvalues. Such bounds seemed to offer an illustrative explanation especially in case when large outlying eigenvalues were present in the spectrum.[2] These composite polynomial bounds assumed exact arithmetic. As rounding errors may substantially delay convergence of the CG method, it is not clear whether the composite polynomial bounds and the conclusions based on them apply to finite precision CG computations. A motivating example is presented in Figure B.1. It indeed shows that a composite polynomial bound can fail to describe CG convergence quantitatively and even *qualitatively.* The difficulty has been to some extent noticed already by Jennings in the paper {35}, and also by van der Sluis and van der Vorst {58} who therefore restrict themselves to the case of small outlying eigenvalues, where the difficulty caused by finite precision arithmetic is not strongly pronounced. In the rest of the paper we deal with the composite polynomial bounds with large outlying eigenvalues. They are used for quantitative evaluation of CG convergence and conclusions based on them are published in recent literature.

The paper is organized as follows. In Section B.2 we briefly clarify the relationship between the CG method, the CSI method and the well known linear

[2] It should be understood, however, that the spectral upper bounds do not necessarily describe the actual CG convergence behaviour for particular right hand sides (initial residuals); see, e.g., {Sections 5.6.1–5.6.3, 37} and {9, 10, 11, 43, 44}.

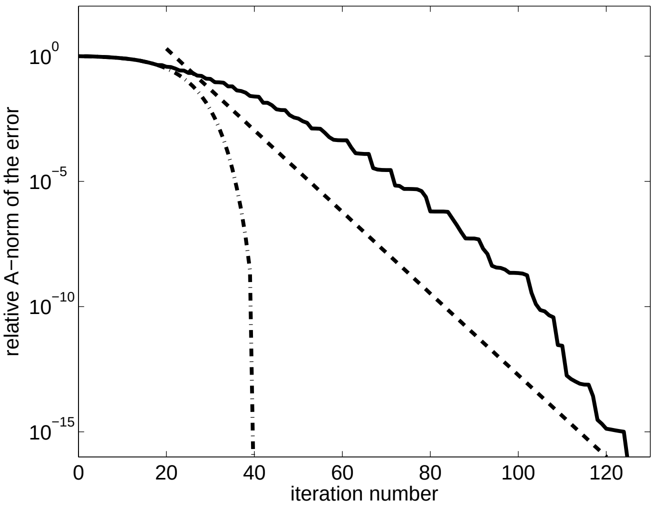


Figure B.1: Rounding errors can cause a substantial delay of convergence in finite precision CG computations (solid line) in comparison to their exact precision counterpart (dash-dotted line). A composite polynomial bound (dashed line) fails to describe the finite precision CG behaviour *quantitatively* (the slope given by the bound is not descriptive) and *qualitatively* (the staircase-like shape of the convergence curve).

convergence bound derived using Chebyshev polynomials. Section B.3 describes the construction of the composite polynomial bounds and comments on their properties. In Section B.4 we use results of the backward-like analysis by Greenbaum and compare *exact* CG computations where matrices have well separated individual eigenvalues, with exact CG computations where matrices have corresponding well separated *clusters* of eigenvalues. We conclude that a “bird’s eye view” of the spectrum can be misleading in Krylov subspace methods. Based on that we examine validity of the composite polynomial bounds for finite precision CG computations. We conclude and numerically demonstrate that in the presence of large outlying eigenvalues such bounds have, apart from simple exceptions, little in common with the finite precision behaviour of the CG method. Section B.5 presents numerical experiments which illustrate in detail shortcomings of the composite polynomial bounds. Concluding remarks summarize the presented clarifications and formulate recommendations for evaluation of the CG rate of convergence.

Writing this paper is motivated by persisting misunderstandings reappearing in literature. This is not meant as a criticism or a negative statement. Our point is that the whole matter is very complex and this should be taken into account whenever any simplification is made. The presented formulas are not new, but, except for the Chapter 5 of the monograph {37}, they have not been, to our knowledge, presented in a comprehensive way in a single publication. Most of the points are presented in {37}, but their placement is subordinate to the organization of the whole monograph, which addresses many related as well as many distant topics. Therefore we consider useful to publish this focused presentation, which in some parts (in particular Section B.4 and Section B.5) complements the presentation in {37} by some new observations. In comparison to a monograph covering much larger area, presentation in the paper allows to focus on interpretation of the formulas. We believe that here the interpretation is more important than the formulas themselves. A need for the correct interpretation can be underlined by the following quote presented (in a somewhat related context) in the instructive paper by Faber, Manteuffel and Parter {p. 113, 22}:

> "There is no flaw in the analysis, only a flaw in the conclusions drawn from the analysis."

B.2 Chebyshev semi-iterative method, conjugate gradient method and their convergence bounds

The idea of the Chebyshev semi-iterative (CSI) method can be linked, with the works of Flanders and Shortley {23}, Lanczos {36} and Young {63}. The CSI method requires a knowledge or estimation of the extreme eigenvalues $\lambda_1 < \lambda_N$ of A and it can be implemented using the three-term recurrence relation for the Chebyshev polynomials; see, e.g., {Chapter 5, 60}.

The CSI method can be viewed as a polynomial acceleration of the stationary Richardson iterations {50} where the k-th error can be written as

$$x - x_k = \phi_k^R(A)(x - x_0), \tag{B.6}$$

and the iteration polynomial

$$\phi_k^R(\lambda) = \left(1 - \frac{2\lambda}{\lambda_1 + \lambda_N}\right)^k$$

belongs to the set of polynomials of degree k with the constant term equal to one (i.e. having the value one at zero). As has been already observed by Richardson in {50}, replacing the k-multiple root of the iteration polynomial $\phi_k^R(\lambda)$ by k distinct roots may lead to faster convergence. The CSI method is motivated by the following reasoning. Let

$$x - x_k = \phi_k(A)(x - x_0),$$

where $\phi_k(\lambda),\ \phi_k(0) = 1$, represents the polynomial of degree at most k. Then the A-norm of the error

$$\|x - x_k\|_A = \{(x - x_k)^* A (x - x_k)\}^{\frac{1}{2}}$$

is given by

$$\|x - x_k\|_A = \|\phi_k(A)(x - x_0)\|_A \tag{B.7}$$

and using the spectral decomposition (B.1) of A the relative A-norm of the error satisfies

$$\frac{\|x - x_k\|_A}{\|x - x_0\|_A} \leq \|\phi_k(A)\| = \max_{j=1,\ldots,N} |\phi_k(\lambda_j)|\,. \tag{B.8}$$

The right hand side in (B.8) is independent of the right hand side b and thus it represents the worst case upper bound. Maximizing over the whole interval $[\lambda_1, \lambda_N]$ instead of the discrete set of eigenvalues $\lambda_1, \ldots, \lambda_N$ gives the bound

$$\frac{\|x - x_k\|_A}{\|x - x_0\|_A} \le \max_{\lambda \in [\lambda_1, \lambda_N]} |\phi_k(\lambda)| \,. \tag{B.9}$$

Setting the roots of the iteration polynomial $\phi_k(\lambda)$ as the roots of the shifted Chebyshev polynomial

$$\chi_k(\lambda) = \begin{cases} \cos\left(k \arccos\left(\dfrac{2\lambda - \lambda_N - \lambda_1}{\lambda_N - \lambda_1}\right)\right) & \text{for } \lambda \in [\lambda_1, \lambda_N], \\ \cosh\left(k \operatorname{arccosh}\left(\dfrac{2\lambda - \lambda_N - \lambda_1}{\lambda_N - \lambda_1}\right)\right) & \text{for } \lambda \notin [\lambda_1, \lambda_N], \end{cases} \tag{B.10}$$

is motivated by the fact that

$$\phi_k(\lambda) \equiv \chi_k(\lambda)/\chi_k(0) \tag{B.11}$$

represents the unique solution of the minimization problem

$$\min_{\substack{\phi(0)=1 \\ \deg(\phi) \le k}} \max_{\lambda \in [\lambda_1, \lambda_N]} |\phi(\lambda)| \tag{B.12}$$

originally solved by Markov {39}. In words, the k-th shifted and scaled Chebyshev polynomial has the minimal maximum norm on the interval $[\lambda_1, \lambda_N]$ among the set of all polynomials of degree at most k having the value one at zero.

Substituting (B.11) into (B.9) and using $|\chi_k(\lambda)| \le 1$ for $\lambda \in [\lambda_1, \lambda_N]$ results in the bound for the relative A-norm of the error

$$\frac{\|x - x_k\|_A}{\|x - x_0\|_A} \le |\chi_k(0)|^{-1}, \quad k = 0, 1, 2, \ldots; \tag{B.13}$$

see {Section 2, 63}. The alternative definition of the Chebyshev polynomials

$$\chi_k(\gamma) = \frac{1}{2}\left(\left(\gamma + (\gamma^2 - 1)^{\frac{1}{2}}\right)^k + \left(\gamma + (\gamma^2 - 1)^{\frac{1}{2}}\right)^{-k}\right) \tag{B.14}$$

(see, e.g., {Section 1.1, 51}) gives with the shift $\gamma = (2\lambda - \lambda_N - \lambda_1)/(\lambda_N - \lambda_1)$ used in (B.10) after a simple manipulation

$$|\chi_k(0)| = \frac{1}{2}\left[\left(\frac{\sqrt{\kappa(A)} + 1}{\sqrt{\kappa(A)} - 1}\right)^k + \left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1}\right)^k\right] \ge \frac{1}{2}\left(\frac{\sqrt{\kappa(A)} + 1}{\sqrt{\kappa(A)} - 1}\right)^k \tag{B.15}$$

where $\kappa(A) = \lambda_N/\lambda_1$ is the condition number of A. This gives the convergence bound *for the CSI method*, which was published in this form by Rutishauser {II.23, 21} in 1959,

$$\frac{\|x - x_k\|_A}{\|x - x_0\|_A} \le 2\left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1}\right)^k, \quad k = 0, 1, 2, \ldots. \tag{B.16}$$

The CG approximations x_k minimize the A-norm of the error over the manifolds $x_0 + \mathcal{K}_k(A, r_0)$; cf. {Theorem 4.1, 33}. Equivalently,

$$\|x - x_k\|_A = \min_{\substack{\varphi(0)=1\\ \deg(\varphi)\leq k}} \|\varphi(A)(x - x_0)\|_A \tag{B.17}$$

$$= \min_{\substack{\varphi(0)=1\\ \deg(\varphi)\leq k}} \left\{ \sum_{j=1}^{N} |\xi_j|^2 \lambda_j \varphi^2(\lambda_j) \right\}^{1/2}, \tag{B.18}$$

where $|\xi_j|$ represents the size of the component of the initial error $x - x_0$ in the direction of the eigenvector u_j corresponding to λ_j, i.e.,

$$x - x_0 = \sum_{j=1}^{N} \xi_j u_j \tag{B.19}$$

and, similarly to (B.18),

$$\|x - x_0\|_A = \left\{ \sum_{j=1}^{N} |\xi_j|^2 \lambda_j \right\}^{1/2}. \tag{B.20}$$

The formula (B.17) leads, using the spectral decomposition (B.1) of A, to the bound for the relative A-norm of the error

$$\frac{\|x - x_k\|_A}{\|x - x_0\|_A} \leq \min_{\substack{\varphi(0)=1\\ \deg(\varphi)\leq k}} \max_{j=1,\ldots,N} |\varphi(\lambda_j)|; \tag{B.21}$$

cf. (B.8). This bound is independent on the right-hand side b and thus it represents the worst case upper bound for the CG method. Since

$$\min_{\substack{\varphi(0)=1\\ \deg(\varphi)\leq k}} \max_{j=1,\ldots,N} |\varphi(\lambda_j)| \leq |\chi_k(0)|^{-1} \max_{j=1,\ldots,N} |\chi_k(\lambda_j)| \leq |\chi_k(0)|^{-1}, \tag{B.22}$$

we can apply (B.15) and conclude that the bound (B.16) must also hold for the CG method.

Now we come to the point which is fundamental but still very rarely mentioned in literature. It should be acknowledged that (B.16) *represents the bound for the CSI method*; see the very clear description given by Rutishauser in {21}. This bound holds for the CG method because the optimal polynomial giving the minimum in (B.21) can be bounded using (B.22). The behaviour of $\|x - x_k\|_A$ for some given initial error (residual) is, however, given by (B.18), which can be substantially different than suggested by (B.21) and therefore certainly substantially different than suggested by the CSI error bound (B.16). The different nature of the CG and CSI methods is clear also from the comparison of the minimization problems (B.12) and (B.18). Whereas the CSI norm of the error can be *tightly* bounded by the minimization problem over the whole interval $[\lambda_1, \lambda_N]$, the CG norm of the error is determined by the discrete minimization problem.

We have presented the (known) derivation in detail in order to avoid further misinterpretations of the relationship between the CSI and CG methods and of the relationship of the bound (B.16) to the CG rate of convergence. In short, as described in Section B.1.1, CG solves the simplified Stieltjes moment problem. Therefore the CG iteration polynomials $\varphi_k(\lambda),\ k = 0, 1, \ldots, N$ defined by (B.17) are orthogonal with respect to the (discrete) inner product determined by the Riemann-Stieltjes integral with the distribution function $\omega(\lambda)$. The Chebyshev polynomials $\chi_k(\lambda),\ k = 0, 1, \ldots$ are orthogonal with respect to the certain continuous and discrete inner products which contain apart from the extremal eigenvalues λ_1 and λ_N no further information about the data A, b and r_0 (or $x - x_0$); see, e.g. {Section 1.5, 51} and {Theorem 4.5.20, 16}. Polynomials orthogonal with different inner products can indeed be very different. Therefore it is beyond any doubt that, except for very special situations, *the bound* (B.16) *relevant for the CSI method has a very little in common with the rate of convergence of the CG method.* Further details and extensive historical comments can be found in {Section 5.6.2, 37}.

The upper bound (B.16) implies that, in *exact arithmetic*,

$$k_\epsilon = \left\lceil \frac{1}{2} \ln\left(\frac{2}{\epsilon}\right) \sqrt{\kappa(A)} \right\rceil \tag{B.23}$$

iterations ensure the decrease of the relative energy norm of the CSI (and therefore also CG) error below the given level of accuracy $\epsilon > 0$ (here $\lceil\cdot\rceil$ denotes rounding up to the nearest integer). As justified in {27, 29}, using results of a thorough analysis, the presented results hold, with a small correction, also for *finite precision arithmetic* CG computations. When $\kappa(A) = \lambda_N/\lambda_1 \approx 1$, the linear system is easily solvable. Using the bound (B.16) and the iteration count (B.23) for CG computations then does not cause any harm. But in such cases one should also ask whether the CG method is really needed for solving such problems. Simpler methods might be fast enough. If $\kappa(A) \gg 1$, then, depending on the *distribution of the spectrum inside the interval* $[\lambda_1, \lambda_N]$, the CG method and the CSI method can naturally perform very differently. In such cases an application of the bound (B.16) to the CG method should always be accompanied with an appropriate justification.

B.3 Composite polynomial bounds and superlinear convergence assuming exact arithmetic

As mentioned above, the superlinear convergence behaviour of the CG method in *exact arithmetic* was explained by Axelsson {2} and Jennings {35} using composite polynomial bounds. For any given polynomial $q_m(\lambda)$ of degree $m \leq k$ satisfying $q_m(0) = 1$ we obtain

$$\min_{\substack{\varphi(0)=1\\ \deg(\varphi)\leq k}} \max_{j=1,\ldots,N} |\varphi(\lambda_j)| \leq \min_{\substack{\varphi(0)=1\\ \deg(\varphi)\leq k-m}} \max_{j=1,\ldots,N} |q_m(\lambda_j)\varphi(\lambda_j)|\,, \tag{B.24}$$

where the minimax problem on the right hand side considers the composite polynomial $q_m(\lambda)\varphi(\lambda)$. In order to describe the superlinear convergence in case of large outlying eigenvalues, Axelsson and Jennings propose in {2, 35} the following natural choice

$$q_m(\lambda) = \prod_{j=N-m+1}^{N} \left(1 - \frac{\lambda}{\lambda_j}\right). \tag{B.25}$$

Since the polynomial $q_m(\lambda)$ given by (B.25) has by construction its roots at the m largest eigenvalues, the relative A-norm of the error is bounded, using (B.21) and (B.24), as

$$\begin{aligned} \frac{\|x - x_k\|_A}{\|x - x_0\|_A} &\le \min_{\substack{\varphi(0)=1 \\ \deg(\varphi)\le k-m}} \max_{j=1,\ldots,N} |q_m(\lambda_j)\varphi(\lambda_j)| && \text{(B.26)} \\ &\le \max_{j=1,\ldots,N-m} |q_m(\lambda_j)| \min_{\substack{\varphi(0)=1 \\ \deg(\varphi)\le k-m}} \max_{j=1,\ldots,N-m} |\varphi(\lambda_j)| \,. && \text{(B.27)} \end{aligned}$$

The polynomial $\varphi(\lambda)$ is evaluated only at the eigenvalues $\lambda_1, \ldots, \lambda_{N-m}$. Therefore the use of the composite polynomial

$$q_m(\lambda)\chi_{k-m}(\lambda)/\chi_{k-m}(0), \tag{B.28}$$

where $\chi_{k-m}(\lambda)$ denotes the Chebyshev polynomial of degree $k - m$ shifted to the interval $[\lambda_1, \lambda_{N-m}]$, results using $|q_m(\lambda_j)| \le 1$ for $j = 1, \ldots, N - m$, analogously to Section B.2, in the bound

$$\frac{\|x - x_k\|_A}{\|x - x_0\|_A} \le 2\left(\frac{\sqrt{\kappa_m(A)} - 1}{\sqrt{\kappa_m(A)} + 1}\right)^{k-m}, \quad k = m, m+1, \ldots, \tag{B.29}$$

where $\kappa_m(A) \equiv \lambda_{N-m}/\lambda_1$ is the so-called *effective condition number.* This quantity is typically substantially smaller than the condition number $\kappa(A)$ which indicates possibly faster convergence after m initial iterations. Illustration of the composite polynomial (B.28) is for $k = 8$, $m = 2$, and the eigenvalues $\lambda_1 = 0.1, \lambda_{N-2} = 6, \lambda_{N-1} = 9$ and $\lambda_N = 15$ given in Figure B.2. As we can immediately observe, the composite polynomial has even for small N, k and small $\kappa(A)$ and $\kappa_m(A)$ very large gradients close to the outlying eigenvalues λ_{N-1} and λ_N. This observation will be important below.

Using an idea analogous to {58}, CG computations with the initial error $x-x_0$ are compared in {Theorem 5.6.9, 37} to CG computations with the initial error $x - \tilde{x}_0$ obtained from $x - x_0$ by neglecting the components ξ_j in the direction of the m eigenvectors corresponding to the m largest eigenvalues,

$$\|x - \tilde{x}_0\|_A = \left\{ \sum_{j=1}^{N-m} |\xi_j|^2 \lambda_j \right\}^{1/2}. \tag{B.30}$$

This comparison gives the following formula

$$\|x - \tilde{x}_k\|_A \le \|x - x_k\|_A \le \|x - \tilde{x}_{k-m}\|_A \,, \quad k = m, m+1, \ldots \tag{B.31}$$

The right inequality in (B.31) shows that CG computation for $Ax = b$ with the initial error $x - x_0$ (the initial residual $r_0 = b - Ax_0$) is from its m-th iteration at

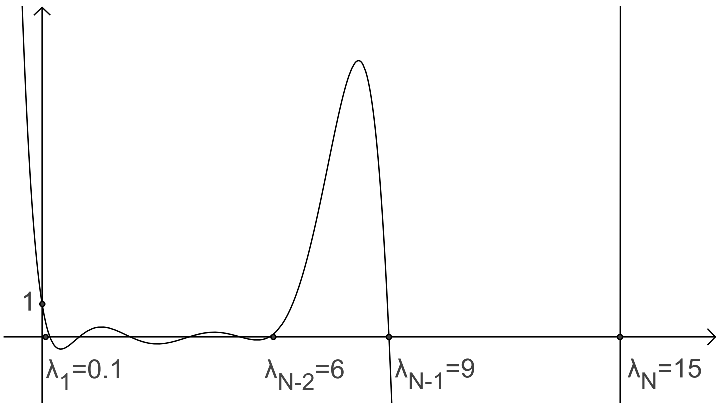


Figure B.2: Illustration of the composite polynomial (B.28) with $k = 8$ and $m = 2$. The polynomial has roots at two large outlying eigenvalues and on the rest of the spectrum is small due to the minimax property of the Chebyshev polynomials. Here the underlying matrix of dimension N would have two largest eigenvalues $\lambda_N = 15, \lambda_{N-1} = 9$ and the remaining eigenvalues would be arbitrarily distributed in the interval $[0.1, 6]$.

least as fast as CG computations for $Ax = b$ with the initial error $x - \tilde{x}_0$ from the start. Dividing this inequality by $\|x - x_0\|_A$ and using $\|x - \tilde{x}_0\|_A \leq \|x - x_0\|_A$ we get the upper bound (B.29) based on the idea of composite polynomial, indeed

$$\frac{\|x - x_k\|_A}{\|x - x_0\|_A} \leq \frac{\|x - \tilde{x}_{k-m}\|_A}{\|x - \tilde{x}_0\|_A} \leq 2\left(\frac{\sqrt{\kappa_m(A)} - 1}{\sqrt{\kappa_m(A)} + 1}\right)^{k-m}, \quad k = m, m+1, \ldots. \tag{B.32}$$

This upper bound can be interpreted as if the first m CG iterations "annihilate" the m large outlying eigenvalues with the subsequent convergence rate bounded linearly by (B.32). It should be noted, however, that this is nothing but an *interpretation*. CG computations do not work that way; see also {Section 5.6.4, 37}.

Analogously to (B.23) in Section B.2 we get from the upper bound (B.32) that after

$$k_\epsilon = m + \left\lceil \frac{1}{2} \ln\left(\frac{2}{\epsilon}\right) \sqrt{\kappa_m(A)} \right\rceil \tag{B.33}$$

iterations, the relative A-norm of the error drops below the given tolerance ϵ; see {p. 132, 2}, {relation (5.9), 35} as well as the recent application of this formula in {Theorem 2.5, 53}.

It should be emphasized, however, that all this is true only in exact arithmetic. The rest of the paper explains that, in general, this approach *must fail in finite precision arithmetic.* The failure of the composite polynomial bounds in finite precision CG computations can be explained by the fact that the closely related Lanczos method computes in finite precision arithmetic repeated approximations of large outlying eigenvalues. This was observed by many authors and it led to results explaining finite precision behaviour of the Lanczos and CG methods; see, in particular, {49, 27, 40} and the survey {41} referring to extensive further literature. Despite the theoretical and experimental counterarguments, the composite polynomial bounds and the related asymptotic convergence factor ideas with neglecting eigenvalues away from the rest of the spectrum as insignificant are tempting to be used for justification of cost in CG computations; see e.g. {Remark 2.1, 38}, {Section 20.4, 57} and {Theorem 2.5, 53}. In the rest of the paper we restrict ourselves to investigation of the bound (B.32) and the formula (B.33). Other approaches not based on Chebyshev polynomials should be in the presence of large outlying eigenvalues examined analogously.

B.4 Analysis of the composite polynomial bounds in finite precision arithmetic

The CG method determines in exact arithmetic an orthogonal basis of the Krylov subspace $\mathcal{K}_k(A, r_0)$ given by the residuals r_j, $j = 0, 1, \ldots, k-1$. However, in finite precision CG computations the orthogonality of the computed residual vectors is (usually quickly) lost and they often become even (numerically) linearly dependent. Consequently, the computed residual vectors may at the step k span a subspace of dimension smaller than k. This *rank-deficiency* of computed Krylov subspace bases thus determines *delay* of convergence of finite precision computations, which can be defined as the difference between the number of iterations required to attain a prescribed accuracy in finite precision computations and the number of iterations required to attain the same accuracy assuming exact arithmetic.

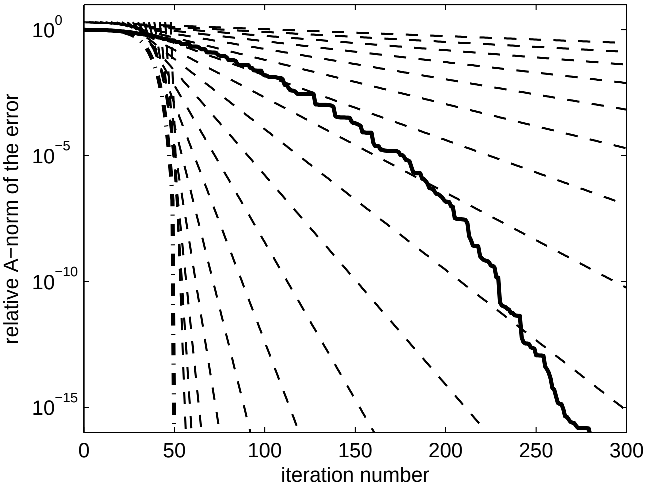


Figure B.3: The sequence of the composite polynomial bounds (B.29) (dashed lines) for increasing number of deflated large eigenvalues ($m = 0, 3, 6, \ldots$) is compared with the results of finite precision CG computations (bold solid line) and exact CG computations (dash-dotted line).

The bound (B.29) and the number of iterations (B.33) were derived assuming exact arithmetic and therefore they do not reflect possible delay of convergence. In finite precision CG computations they suffer from a fundamental difficulty outlined in Figure B.1 and illustrated in more detail in Figure B.3. Here the dashed lines plot the sequence of the composite polynomial bounds (B.29) with increasing number of the large eigenvalues of A considered as outliers ($m = 0, 3, 6, \ldots$). The bold solid line represents the convergence curve of the finite precision CG and the dash-dotted line the CG behaviour assuming exact arithmetic[3]. Computations were performed using a symmetric positive definite diagonal matrix A of the size $N = 50$ with the eigenvalues $0 < \lambda_1 < \lambda_2 < \ldots < \lambda_{N-1} < \lambda_N$, where $\lambda_1 = 0.1$, $\lambda_N = 10^4$, the inner eigenvalues were given by the formula

$$\lambda_i = \lambda_1 + \frac{i-1}{N-1}(\lambda_N - \lambda_1)\rho^{N-i} \quad i = 2, \ldots, N-1 \tag{B.34}$$

and $\rho = 0.8$; see {54, 29, 40}. The parameter $\rho \in (0, 1]$ determines the non-uniformity of the spectrum. For $\rho \ll 1$ the eigenvalues tend to cumulate near λ_1 and for $\rho = 1$ the spectrum is distributed uniformly. In our experiments we use the vector b of all ones, i.e., $b = [1, \ldots, 1]^T$. We observe that the linear convergence bounds determine (a close) envelope for the exact arithmetic CG convergence curve. This is in correspondence with the intuitive explanation of the superlinear convergence behaviour of CG in exact arithmetic presented in literature. The data in this example do not represent a purely academic case. Spectra with large outlying eigenvalues do appear in practice; see e.g., {8} for an early study on this related to preconditioning techniques.

The point is that *none* of the straight lines describes the finite precision convergence behaviour, as can be seen by comparing the dashed lines with the bold solid line. Evidently, the composite polynomial bounds (B.29) can not be used, in general, as upper bounds.

The finite precision behaviour of the Lanczos and CG methods was analyzed, in particular, by Paige and Greenbaum; see {27, 48, 49}. Shortly speaking, Greenbaum has proved that

> the finite precision Lanczos computation for a matrix A and a given starting vector v produces in steps 1 through k the same eigenvalue approximations (the same Jacobi matrix T_k) as the exact Lanczos computation for some particular larger matrix $\widehat{A}(k)$ and some particular starting vector $\widehat{v}(k)$ while the eigenvalues of $\widehat{A}(k)$ all lie *within tiny intervals around the eigenvalues of* A. The size as well as (all) individual entries of $\widehat{A}(k)$ and $\widehat{v}(k)$ depend on the rounding errors in the steps 1 through k.

It should be emphasized that $\widehat{A}(k)$ *is not given by a slight perturbation of* A, as sometimes stated in literature; $\widehat{A}(k)$ is typically much larger than A. This is illustrated on Figure B.4. An analogous statement is valid, with a small inaccuracy specified in {27}, also for the behaviour of finite precision CG computations. This explains why (B.29) and (B.33) must fail, in general, in finite precision arithmetic, where m CG steps are not enough to annihilate the influence of the m large outlying eigenvalues. One may suggest to resolve the matter by adding several penalty steps which account for the effects of rounding errors. The number of such additional steps, however, depends on current iteration k and it can not be determined a-priori. The difficulty is illustrated in Figure B.1 above where the "penalty" is given by the horizontal differences between the dashed line (the bound) and the solid line (computed results).

[3] CG behaviour assuming exact arithmetic is simulated throughout the paper by double reorthogonalization of the residual vectors; see {29, 48}.

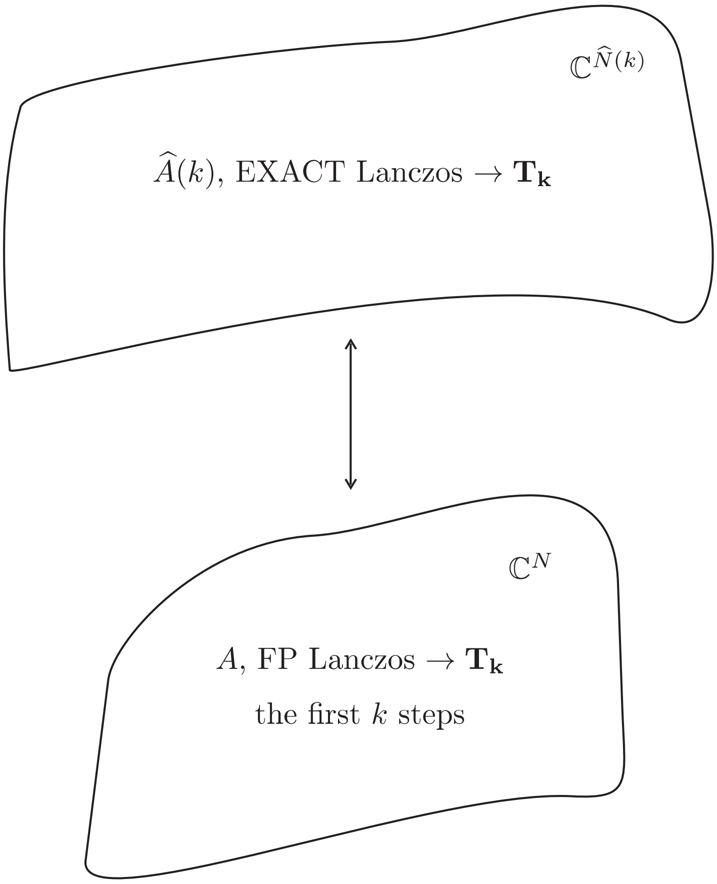

$\mathbb{C}^{\widehat{N}(k)}$
$\widehat{A}(k)$, EXACT Lanczos → $\mathbf{T_k}$
$\mathbb{C}^{N}$
A, FP Lanczos → $\mathbf{T_k}$
the first k steps

Figure B.4: For any $k = 1, 2, \ldots$ the first k steps of the finite precision Lanczos computation for $A \in \mathbb{C}^{N \times N}$ can be analyzed as the first k steps of the exact Lanczos for the (possibly much larger) matrix $\widehat{A}(k) \in \mathbb{C}^{\widehat{N}(k) \times \widehat{N}(k)}$ depending on k which generates the same $k \times k$ Jacobi matrix T_k.

As stated above, the matrix $\widehat{A}(k)$ and the vector $\widehat{v}(k)$ depend on the iteration step k. The reasoning about the delay in finite precision CG computations suggests (it was experimentally confirmed in {29}) that the particular matrix $\widehat{A}(k)$ constructed for the k steps of the given finite precision CG computation can be replaced (with an acceptable inaccuracy) by a matrix $\widehat{A}$ having sufficiently many eigenvalues in *tight clusters around each eigenvalue of* A; see also the detailed argumentation in {41} and, in particular, in {Section 5.9, 37}. The appropriate starting vector associated with $\widehat{A}$ can be constructed from A and b independently of k. As an example, the matrix $\widehat{A}$ used in our experiments below has l eigenvalues $\widehat{\lambda}_{j,1} < \widehat{\lambda}_{j,2} < \ldots < \widehat{\lambda}_{j,l}$ uniformly distributed in tiny intervals $[\lambda_j - \Delta, \lambda_j + \Delta]$ around each original eigenvalue λ_j of A, $j = 1, 2, \ldots, N$, where l is sufficiently large in correspondence to the maximal number of the performed iterations steps. The associated right hand side $\widehat{b}$ is obtained from b by splitting each individual entry β_j of b into l equal parts $\widehat{\beta}_{j,1}, \ldots, \widehat{\beta}_{j,l}$ such that $\sum_{s=1}^{l} \widehat{\beta}_{j,s}^2 = \beta_j^2$, $j = 1, 2, \ldots, N$; see {29}.

As an immediate consequence of the results from {27, 29} we get that convergence behaviour of exact CG applied to a matrix with the spectrum having well separated clusters of eigenvalues is both *qualitatively* and *quantitatively* different from the convergence behaviour of *exact* CG applied to a matrix with a spectrum where each cluster is replaced by a single eigenvalue. We can conclude that even for the CG method, the HPD matrix and *assuming exact arithmetic*,

> a spectrum composed of a small number of tight clusters can not be associated, in general, with fast convergence.

Indeed, the associated Stieltjes moment problems from Section B.1.1 can be for different distribution of eigenvalues very different. This is true, in particular, when clusters of eigenvalues are replaced by single (representing) eigenvalues of the same weights; see {47}. This point contradicts the common belief which seems widespread.

We will now explain how this fact is reflected in the composite polynomial convergence bounds (B.29). Using the relationship with the exact CG computations applied to $\widehat{A}$, the corresponding minimization problem which bounds the CG convergence behaviour *in finite precision arithmetic* is not

$$\min_{\substack{\varphi(0)=1\\ \deg(\varphi)\le k}} \max_{j=1,\dots,N} |\varphi(\lambda_j)| \,, \tag{B.35}$$

where $\lambda_1, \dots, \lambda_N$ are the eigenvalues of A; see (B.21). Instead, one must use

$$\min_{\substack{\varphi(0)=1\\ \deg(\varphi)\le k}} \max_{\lambda\in\sigma(\widehat{A})} |\varphi(\lambda)| \,, \tag{B.36}$$

where the spectrum of the matrix $\widehat{A}$ consists of the union of the individual clusters around the original eigenvalues λ_j, $j = 1, \dots, N$, i.e., in our case

$$\sigma(\widehat{A}) \equiv \bigcup_{j=1,\dots,N} \left\{\widehat{\lambda}_{j,1}, \dots, \widehat{\lambda}_{j,l}\right\}. \tag{B.37}$$

Consequently, in order to be valid for finite precision CG computations, the upper bound based on the composite polynomial (B.28) from Section B.3 must use instead of

$$\max_{j=1,\dots,N} |q_m(\lambda_j)\chi_{k-m}(\lambda_j)| \,/\, |\chi_{k-m}(0)| \,, \tag{B.38}$$

which considers the values of the composite polynomial at the eigenvalues $\lambda_1, \dots, \lambda_N$ of A, the modification

$$\max_{\lambda\in\sigma(\widehat{A})} |q_m(\lambda)\chi_{k-m}(\lambda)| \,/\, |\chi_{k-m}(0)| \,, \tag{B.39}$$

which considers the values of the composite polynomial at the eigenvalues of the matrix $\widehat{A}$. As a consequence of the minimality property of the Chebyshev polynomial $\chi_{k-m}(\lambda)$ over the interval $[\lambda_1, \lambda_{N-m}]$, its values outside this interval become even for small k very large. More specifically, the Chebyshev polynomial is outside the minimality interval the fastest growing polynomial of the given degree; see, e.g., {Section 2.7, rel. (2.37), 51} and {Section 3.2.3, 16}. The composite polynomial has, by construction, large values of its gradient at the large outlying eigenvalues of A; see the illustration in Figure B.2 above. The values of the composite polynomial at the points located in the tight clusters around such large outlying eigenvalues can therefore be huge even for small k, and the upper

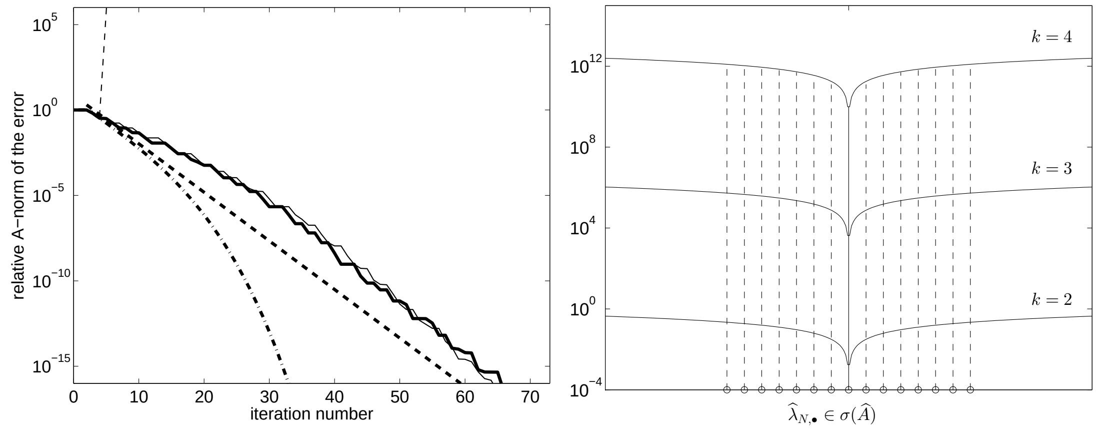


Figure B.5: Left: Whereas the exact CG convergence behaviour corresponding to $\widehat{A}, \widehat{b}$ (solid line) is both quantitatively and qualitatively different from the exact CG convergence behaviour corresponding to A, b (dash-dotted line), it nicely matches the finite precision CG computation (bold solid line) using A, b. The composite polynomial bound (B.29) (bold dashed line) is irrelevant and the bound (B.39) (dashed line) becomes after several iterations meaningless due to huge values of the composite polynomial in the neighborhood of the outlying eigenvalues of A. Right: Using the logarithmic vertical scale we plot a detail of the absolute values of the composite polynomial (with restriction to the values in the interval $[10^{-4},\ 10^{13}]$) corresponding to the k-th iteration with $k = 2, 3$ and 4. The values of the composite polynomial at the eigenvalues $\widehat{\lambda}_{N,s}$, $s = 1, \ldots, l$ clustered around the largest eigenvalue λ_N blow up even for the smallest degrees of the corresponding shifted Chebyshev polynomial $\chi_{k-m}(\lambda)$ ($k - m = 1$ and 2). Here the width of the cluster around λ_N is $4\varepsilon \, \|A\| \approx 10^{-9}$.

bound based on the expression (B.39) becomes after several iterations in practical computations meaningless; see the illustration in Figure B.5. The left part shows finite precision CG convergence behaviour (bold solid line) corresponding to the right hand side b of ones and the matrix A of dimension $N = 40$ with $m = 2$ large outlying eigenvalues $\lambda_{N-1} = 10^4$, $\lambda_N = 10^6$ and with the eigenvalues $\lambda_1, \ldots, \lambda_{N-2}$ determined using

$$\lambda_i = \lambda_1 + \frac{i-1}{N-m-1}\,(\lambda_{N-m} - \lambda_1)\,\rho_{in}^{N-m-i} \quad i = 2, \ldots, N-m-1 \qquad \text{(B.40)}$$

with $\rho_{in} = 0.9$, $\lambda_1 = 0.1$ and $\lambda_{N-2} = 1$. We compare it with exact CG convergence behaviour (solid line) corresponding to the associated vector $\widehat{b}$ and matrix $\widehat{A}$ with $\Delta = 2\varepsilon \, \|A\|$ and $l = 15$, where $\varepsilon = 2^{-52}$ is machine roundoff unit; cf. {p. 126, 29}. In agreement with {29} we observe quantitative and qualitative similarity of both convergence curves. The composite polynomial bound (B.29) (bold dashed line) with $m = 2$ (i.e. considering 2 largest eigenvalues of the matrix A as outliers) is for the finite precision computations irrelevant and the associated bound (B.39) (dashed line) practically immediately blows up. The latter is a consequence of the evaluation of the composite polynomial at the eigenvalues of $\widehat{A}$ clustered around the outlying eigenvalues of A as visualized in the right part of the figure.

The spectral upper bound applicable to finite precision CG computations based on the minimization problem (B.36) was investigated, following {27, 29}, by Notay in {46}. He considered the composite polynomial where the part dealing with the outlying eigenvalues has possibly many roots in the neighborhood of the large outlying eigenvalues. The paper presents an estimate of the number of iterations needed to deal with the outlying eigenvalues as the number of iterations increases. This requires estimating the frequency of forming multiple copies of the large outlying eigenvalues, which unavoidably uses partially empirical arguments and requires knowledge of all large outlying eigenvalues. The paper {46} instructively demonstrates that *a-priori* investigation of the CG rate of convergence, which aims at realistic results including effects of rounding errors, is inevitably rather technical. Consequently, a practical application of a realistic *a-priori* analysis which is not specialized to some particular cases is limited.

B.5 Other shortcomings of composite polynomial bounds

In this section we will comment and numerically demonstrate several other drawbacks of the composite polynomial bound (B.29). Our observations can be summarized in the following points.

a) The composite polynomial bound (B.29) by construction does not depend on distribution of the eigenvalues within the interval $[\lambda_1, \lambda_{N-m}]$. In contrast to that, a finite precision CG behaviour can significantly depend on this distribution.

b) Unlike the bound (B.29), finite precision CG computations depend on the position of the large outlying eigenvalues.

c) The failure of the composite polynomial bound (B.29) in finite precision CG computations can occur even for a small size and/or conditioning of the problem.

In the numerical illustrations below we used diagonal matrices A and the right hand side b of all ones.

Point a) In Figure B.6 we compare CG computations applied to two problems with the same outlying eigenvalues, the same effective condition number $\kappa_m(A) = \lambda_{N-m}/\lambda_1$ but with different distribution of the eigenvalues within the interval $[\lambda_1, \lambda_{N-m}]$. Computations were performed using diagonal matrices of dimension $N = 80$ with $m = 7$ large outlying eigenvalues $\lambda_{N-6}, \ldots, \lambda_N$ and the eigenvalue λ_{N-7} determined using (B.34) with $\lambda_1 = 0.1$, $\lambda_N = 10^5$ and $\rho \equiv \rho_{out} = 0.3$. The eigenvalues $\lambda_2, \ldots, \lambda_{N-8}$ are distributed in the interval $[\lambda_1, \lambda_{N-7}]$ either uniformly or using (B.40) with $\rho_{in} = 0.95$.

The composite polynomial bound (B.29) with $m = 7$ (dashed line) is the same for both computations, as it does not reflect the distribution of the eigenvalues within the interval $[\lambda_1, \lambda_{N-m}]$. On the contrary, the convergence of the CG method depends in exact arithmetic slightly (dash-dotted lines) and in finite precision arithmetic very significantly (bold solid lines) on the distribution of *all* eigenvalues, including those in the interval $[\lambda_1, \lambda_{N-m}]$.

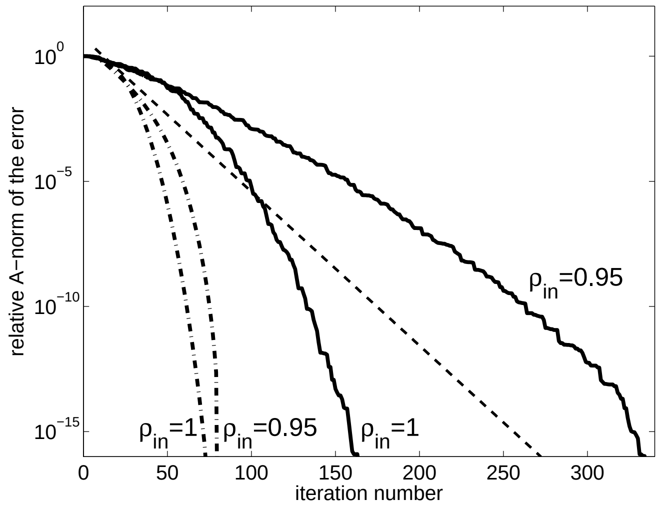


Figure B.6: Unlike the composite polynomial bound (dashed line), both exact (dash-dotted lines) and finite precision (bold solid lines) CG convergence behaviour are sensitive to the change of distribution of the eigenvalues in the interval $[\lambda_1, \lambda_{N-m}]$. In finite precision computations the difference between the uniform distribution with $\rho_{in} = 1$ and the distribution with $\rho_{in} = 0.95$ is significant.

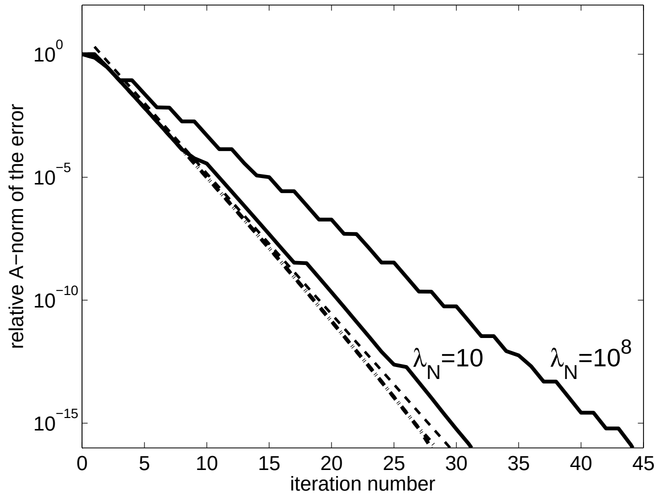


Figure B.7: Finite precision CG computations (bold solid lines) are, in contrast to the exact CG convergence behaviour (dash-dotted lines), sensitive to the position of the single large outlying eigenvalue λ_N. The frequency of forming multiple approximations of the largest eigenvalue is seriously affected by its position. The bounds based on the composite polynomial (B.28) (dashed line) can fail also in the presence of only a single large outlier.

Point b) As mentioned in the previous paragraph, the convergence behaviour of the CG method depends on distribution of all eigenvalues. Thus the position of the outlying eigenvalues is of importance. In Figure B.7 we plot the finite precision CG convergence curves (bold solid lines) and CG behaviour assuming exact arithmetic (dash-dotted lines) using the diagonal matrices of dimension $N = 50$ whose largest eigenvalue $\lambda_N = 10$ respectively $\lambda_N = 10^8$ is considered as *the only outlier* and the eigenvalues $\lambda_1, \ldots, \lambda_{N-1}$ are distributed uniformly within the interval $[\lambda_1, \lambda_{N-1}]$, $\lambda_1 = 0.1,\ \lambda_{N-1} = 0.3$.

The exact CG convergence behaviour is in both cases nearly identical. The delay of convergence in the finite precision CG computation with the outlying eigenvalue $\lambda_N = 10^8$ is naturally more significant than with the outlying eigenvalue $\lambda_N = 10$. This happens due to more frequent occurrence of the multiple approximations of the largest eigenvalue. Thus the information about the *number* of eigenvalues lying above some given number $\bar{\lambda}$ (as used, e.g., in {Corollary 2.2, 53} or {p. 4, 38}) is without further analysis of the problem not sufficient for estimating the actual convergence rate of finite precision CG computations. A single large outlying eigenvalue *can* affect the "asymptotic" rate of convergence. The composite polynomial bound (B.29) can fail even in this case.

Point c) Depending on the distribution of eigenvalues, the composite convergence bound can fail even for small and well-conditioned problems. We will use diagonal matrices with spectrum determined in the following way. We consider 4 different problems with $N = 30$ or 100 and $\lambda_N = 10$ or 10^6. The $m = 8$ large outlying eigenvalues $\lambda_{N-7}, \ldots, \lambda_N$ and the eigenvalue λ_{N-8} are given by (B.34) with $\lambda_1 = 0.1$, $\rho_{out} = 0.6$ for $N = 30$, $\rho_{out} = 0.2$ for $N = 100$. The rest of the eigenvalues is distributed in the interval $[\lambda_1, \lambda_{N-8}]$ using (B.40) with $\rho_{in} = 0.8$. Each of the subplots in Figure B.8 shows that the composite polynomial bound (bold dashed line) and the finite precision CG convergence behaviour (bold solid line) have a little in common. We also plot the exact CG convergence behaviour (solid line) corresponding to the matrix $\widehat{A}$ which is determined using $\Delta = \varepsilon \, \|A\|$ and $l = 15$. Similarly as in Section B.4 we observe that it qualitatively matches the finite precision CG computations. The associated upper bound (B.39) (dashed line) becomes after several iterations meaningless.

B.6 Concluding remarks

This paper demonstrates that the composite polynomial bound (B.29) based on a Chebyshev polynomial and a fixed part having roots at large outlying eigenvalues of A has, in general, a little in common with actual finite precision CG computations. Related to that, CG method applied to a problem $Ax = b$ with a spectrum of the matrix A consisting of t tiny clusters does not necessarily produce a good approximation to the solution x within t steps. Many more steps may be needed, depending on the position of the individual clusters (this holds in exact arithmetic as well as in finite precision arithmetic). Our experimental illustration use small examples with diagonal matrices. In our opinion this makes the message appealing also for computations with real data.

Although this paper concentrates on bounds based on Chebyshev polynomials, the main point that the large outlying eigenvalues can challenge the relevance of *a-priori* CG convergence rate analysis when applied to practical computations is valid in general. Any *a-priori* CG convergence rate analysis is based on a substantial simplification of the very complex phenomena. We must admit this fact and verify any conclusion drawn from such analysis by justification of the assumptions incorporated in the whole development. *A-priori* convergence bounds are often used in connection with evaluation of preconditioning strategies and

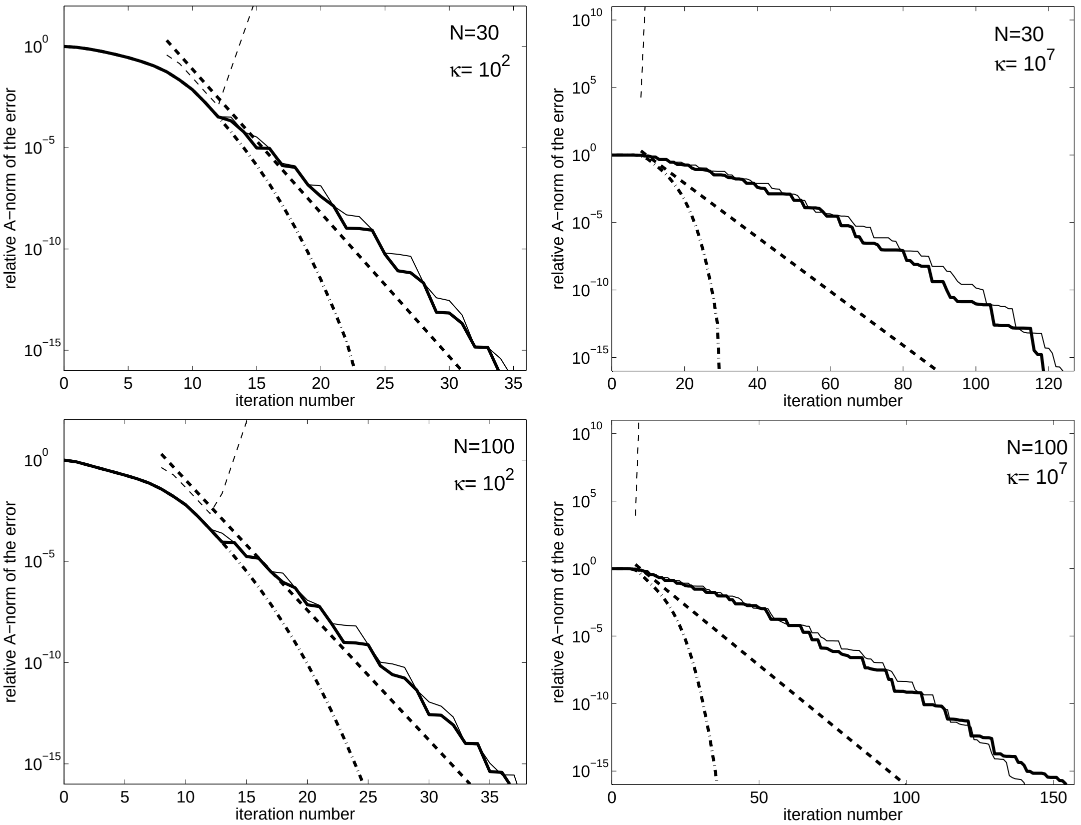


Figure B.8: The failure of the composite polynomial bound (bold dashed line) in finite precision CG computations (bold solid line) for well-conditioned (left) resp. ill-conditioned (right) smaller (top) and larger (bottom) problems. The exact CG convergence behaviour corresponding to $\widehat{A}$ (solid line) matches the finite precision CG computations performed using A and it differs both qualitatively and quantitatively from the exact CG convergence behaviour corresponding to A (dash-dotted line). The upper bound (B.39) (dashed line) which evaluates the composite polynomial in the neighborhood of outliers gives no relevant information.

their optimality. Here the validity of the bounds in the presence of rounding errors and the *tightness* of the bounds should be taken as a strict requirement, otherwise the conclusions are not mathematically justified. There is an obvious exception, when preconditioning ensures very fast convergence, so that the tightness of the bounds does not matter. In such cases rounding errors have no chance to spoil significantly the computation.

In order to limit the effects of rounding errors, it would be useful to avoid pro-actively presence of large outlying eigenvalues in the spectrum of the preconditioned matrix; cf. {8}. Reorthogonalization procedures known from the Lanczos method for computing several dominating eigenvalues are in the CG context not generally applicable for efficiency reasons. They might be worth investigating, however, together with combined arithmetic techniques, in parallel implementations.

Finally, actual error in CG computations should be estimated and analyzed *a-posteriori*. This field has been thoroughly investigated by Golub and his collaborators, with early works {17, 18}; see also {13, 14, 15}. As pointed out in

{55}, important steps in this direction can be found already in the original paper by Hestenes and Stiefel {33}. As in the *a-priori* analysis, the *a-posteriori* estimates and bounds can not be reliably applied to practical computations unless they are accompanied by a thorough rounding error analysis; see the arguments and examples given in {25, 55, 42}. For a survey we refer, e.g., to {Sections 3.3 and 5.3, 41}, {Chapter 12, 24}. In the context of numerical solution of partial differential equations, the *a-posteriori* analysis of the algebraic iterations should be incorporated into the *a-posteriori* analysis of the whole solution process; see, e.g. the recent survey {1} and some possible challenges related to applications of CG formulated in {Chapter 5, 37}.

As in numerical solution of partial differential equations, *a-priori* and *a-posteriori* analysis has its place also in the iterative algebraic computations. In both fields *reliability* is the key requirement.

Acknowledgments. The authors are grateful to Jörg Liesen, Jan Papež and anonymous referee for useful comments.

Bibliography

{1} Arioli, M., Liesen, J., Miedlar, A., and Strakos, Z. Interplay between discretization and algebraic computation in adaptive numerical solution of elliptic pde problems. *GAMM Mitt. Ges. Angew. Math. Mech.* (to appear).

{2} Axelsson, O. A class of iterative methods for finite element equations. *Comput. Methods Appl. Mech. Engrg. 9*, 2 (1976), 123–127.

{3} Axelsson, O. *Iterative solution methods.* Cambridge University Press, Cambridge, 1994.

{4} Axelsson, O. A generalized conjugate gradient minimum residual method with variable preconditioners. In *Advanced mathematics: computations and applications (Novosibirsk, 1995).* NCC Publ., Novosibirsk, 1995, pp. 14–25.

{5} Axelsson, O. Optimal preconditioners based on rate of convergence estimates for the conjugate gradient method. *Numer. Funct. Anal. Optim. 22*, 3-4 (2001), 277–302.

{6} Axelsson, O., and Kaporin, I. On the sublinear and superlinear rate of convergence of conjugate gradient methods. *Numer. Algorithms 25*, 1-4 (2000), 1–22. Mathematical journey through analysis, matrix theory and scientific computation (Kent, OH, 1999).

{7} Axelsson, O., and Karátson, J. Equivalent operator preconditioning for elliptic problems. *Numer. Algorithms 50*, 3 (2009), 297–380.

{8} Axelsson, O., and Lindskog, G. On the eigenvalue distribution of a class of preconditioning methods. *Numer. Math. 48*, 5 (1986), 479–498.

{9} Beckermann, B., and Kuijlaars, A. B. J. On the sharpness of an asymptotic error estimate for conjugate gradients. *BIT 41*, 5, suppl. (2001), 856–867.

{10} Beckermann, B., and Kuijlaars, A. B. J. Superlinear convergence of conjugate gradients. *SIAM J. Numer. Anal. 39*, 1 (2001), 300–329.

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

{12} Brezinski, C. *Projection Methods for Systems of Equations*, vol. 7 of *Studies in Computational Mathematics.* North-Holland Publishing Co., Amsterdam, 1997.

{13} Brezinski, C. Error estimates for the solution of linear systems. *SIAM J. Sci. Comput. 21*, 2 (1999), 764–781.

{14} Calvetti, D., Morigi, S., Reichel, L., and Sgallari, F. Computable error bounds and estimates for the conjugate gradient method. *Numer. Algorithms 25*, 1-4 (2000), 75–88. Mathematical journey through analysis, matrix theory and scientific computation (Kent, OH, 1999).

{15} Calvetti, D., Morigi, S., Reichel, L., and Sgallari, F. An iterative method with error estimators. *J. Comput. Appl. Math. 127*, 1-2 (2001), 93–119. Numerical analysis 2000, Vol. V, Quadrature and orthogonal polynomials.

{16} Dahlquist, G., and Björck, Å. *Numerical methods in scientific computing. Vol. I.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008.

{17} Dahlquist, G., Eisenstat, S. C., and Golub, G. H. Bounds for the error of linear systems of equations using the theory of moments. *J. Math. Anal. Appl. 37* (1972), 151–166.

{18} Dahlquist, G., Golub, G. H., and Nash, S. G. Bounds for the error in linear systems. In *Semi-infinite programming (Proc. Workshop, Bad Honnef, 1978)*, vol. 15 of *Lecture Notes in Control and Information Sci.* Springer, Berlin, 1979, pp. 154–172.

{19} Daniel, J. W. The conjugate gradient method for linear and nonlinear operator equations. *SIAM J. Numer. Anal. 4* (1967), 10–26.

{20} Deuflhard, P. Cascadic conjugate gradient methods for elliptic partial differential equations: algorithm and numerical results. In *Domain decomposition methods in scientific and engineering computing (University Park, PA, 1993)*, vol. 180 of *Contemp. Math.* American Mathematical Society, Providence, RI, 1994, pp. 29–42.

{21} Engeli, M., Ginsburg, T., Rutishauser, H., and Stiefel, E. Refined iterative methods for computation of the solution and the eigenvalues of self-adjoint boundary value problems. *Mitt. Inst. Angew. Math. Zürich. No. 8* (1959), 107.

{22} Faber, V., Manteuffel, T. A., and Parter, S. V. On the theory of equivalent operators and application to the numerical solution of uniformly elliptic partial differential equations. *Adv. in Appl. Math. 11*, 2 (1990), 109–163.

{23} Flanders, D. A., and Shortley, G. Numerical determination of fundamental modes. *J. Appl. Phys. 21* (1950), 1326–1332.

{24} Golub, G. H., and Meurant, G. *Matrices, Moments and Quadrature with Applications.* Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, 2010.

{25} Golub, G. H., and Strakoš, Z. Estimates in quadratic formulas. *Numer. Algorithms 8*, 2-4 (1994), 241–268.

{26} Gratton, S., Titley-Peloquin, D., Toint, P., and Tshimanga, J. Linearizing the method of conjugate gradients. Technical Report naXys-15-2012, Namur Centre for Complex Systems, FUNDP–University of Namur, Belgium (2012).

{27} Greenbaum, A. Behaviour of slightly perturbed Lanczos and conjugate-gradient recurrences. *Linear Algebra Appl. 113* (1989), 7–63.

{28} Greenbaum, A. *Iterative Methods for Solving Linear Systems*, vol. 17 of *Frontiers in Applied Mathematics*. SIAM, Philadelphia, PA, 1997.

{29} Greenbaum, A., and Strakoš, Z. Predicting the behavior of finite precision Lanczos and conjugate gradient computations. *SIAM J. Matrix Anal. Appl. 13*, 1 (1992), 121–137.

{30} Günnel, A., Herzog, R., and Sachs, E. A Note on Preconditioners and Scalar Products for Krylov Methods in Hilbert Space. (preprint).

{31} Gutknecht, M. H., and Strakoš, Z. Accuracy of two three-term and three two-term recurrences for Krylov space solvers. *SIAM J. Matrix Anal. Appl. 22*, 1 (2000), 213–229.

{32} Hackbusch, W. *Iterative Solution of Large Sparse Systems of Equations*, vol. 95 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1994. Translated and revised from the 1991 German original.

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

{34} Hiptmair, R. Operator preconditioning. *Comput. Math. Appl. 52*, 5 (2006), 699–706.

{35} Jennings, A. Influence of the eigenvalue spectrum on the convergence rate of the conjugate gradient method. *J. Inst. Math. Appl. 20*, 1 (1977), 61–72.

{36} Lanczos, C. Chebyshev polynomials in the solution of large-scale linear systems. In *Proceedings of the Association for Computing Machinery, Toronto, 1952* (1953), Sauls Lithograph Co. (for the Association for Computing Machinery), Washington, D. C., pp. 124–133.

{37} Liesen, J., and Strakoš, Z. *Krylov subspace methods: principles and analysis*. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, 2012.

{38} Mardal, K.-A., and Winther, R. Preconditioning discretizations of systems of partial differential equations. *Numer. Linear Algebra Appl. 18*, 1 (2011), 1–40.

{39} Markoff, A. Démonstration de certaines inégalités de M. Tchébychef. *Math. Ann. 24*, 2 (1884), 172–180.

{40} Meurant, G. *The Lanczos and conjugate gradient algorithms: from theory to finite precision computations.* vol. 19 of Software, Environments, and Tools. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2006.

{41} Meurant, G., and Strakoš, Z. The Lanczos and conjugate gradient algorithms in finite precision arithmetic. *Acta Numer. 15* (2006), 471–542.

{42} Meurant, G., and Tichý, P. On computing quadrature-based bounds for the A-norm of the error in conjugate gradients. *Numer. Algorithms 62*, 2 (2013), 163–191.

{43} Naiman, A. E., Babuška, I. M., and Elman, H. C. A note on conjugate gradient convergence. *Numer. Math. 76*, 2 (1997), 209–230.

{44} Naiman, A. E., and Engelberg, S. A note on conjugate gradient convergence. II, III. *Numer. Math. 85*, 4 (2000), 665–683, 685–696.

{45} Nevanlinna, O. *Convergence of iterations for linear equations.* Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 1993.

{46} Notay, Y. On the convergence rate of the conjugate gradients in presence of rounding errors. *Numer. Math. 65*, 3 (1993), 301–317.

{47} O'Leary, D. P., Strakoš, Z., and Tichý, P. On sensitivity of Gauss-Christoffel quadrature. *Numer. Math. 107*, 1 (2007), 147–174.

{48} Paige, C. C. Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix. *J. Inst. Math. Appl. 18*, 3 (1976), 341–349.

{49} Paige, C. C. Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem. *Linear Algebra and Its Applications 34* (1980), 235–258.

{50} Richardson, L. F. The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam. *Phil. Trans. Roy. Soc. London A, 210* (1911), 307–357.

{51} Rivlin, T. J. *Chebyshev Polynomials*, second ed. Pure and Applied Mathematics. John Wiley & Sons Inc., New York, 1990.

{52} Silvester, D. J., and Simoncini, V. An optimal iterative solver for symmetric indefinite systems stemming from mixed approximation. *ACM Trans. Math. Software 37*, 4 (2011), Art. 42, 22.

{53} Spielman, D. A., and Woo, J. A note on preconditioning by low-stretch spanning trees. *Computing Research Repository* (2009).

{54} Strakoš, Z. On the real convergence rate of the conjugate gradient method. *Linear Algebra Appl. 154/156* (1991), 535–549.

{55} Strakoš, Z., and Tichý, P. On error estimation in the conjugate gradient method and why it works in finite precision computations. *Electron. Trans. Numer. Anal. 13* (2002), 56–80.

{56} Strakoš, Z., and Tichý, P. Error estimation in preconditioned conjugate gradients. *BIT 45*, 4 (2005), 789–817.

{57} Tyrtyshnikov, E. E. *A brief introduction to numerical analysis.* Birkhäuser Boston Inc., Boston, MA, 1997.

{58} van der Sluis, A., and van der Vorst, H. A. The rate of convergence of conjugate gradients. *Numer. Math. 48*, 5 (1986), 543–560.

{59} van der Vorst, H. A. Iterative solution methods for certain sparse linear systems with a nonsymmetric matrix arising from PDE-problems. *J. Comput. Phys. 44*, 1 (1981), 1–19.

{60} Varga, R. S. *Matrix iterative analysis*, expanded ed., vol. 27 of *Springer Series in Computational Mathematics.* Springer-Verlag, Berlin, 2000.

{61} Vorobyev, Y. V. *Methods of Moments in Applied Mathematics.* Translated from the Russian by Bernard Seckler. Gordon and Breach Science Publishers, New York, 1965.

{62} Winther, R. Some superlinear convergence results for the conjugate gradient method. *SIAM J. Numer. Anal. 17*, 1 (1980), 14–17.

{63} Young, D. On Richardson's method for solving linear systems with positive definite matrices. *J. Math. Physics 32* (1954), 243–255.