Charles University in Prague
Faculty of Mathematics and Physics

HABILITATION THESIS

Petr Tichý

Analysis of Krylov subspace methods

Departments of Numerical Mathematics

Prague 2015

WWW: `http://www.mff.cuni.cz/~ptichy`
e-mail: `ptichy@karlin.mff.cuni.cz`

Contents

Preface

The presented habilitation thesis summarizes results of 15 journal papers. The common denominator of all the presented works is the investigation and analysis of iterative methods for solving system of linear algebraic equations called Krylov subspace methods. The papers are written by Petr Tichý in collaboration with various co-authors. The most frequent co-authors are Vance Faber (5 papers), Jörg Liesen (10 papers), and Zdeněk Strakoš (4 papers). The collaboration with Zdeněk Strakoš started during Ph.D. studies of Petr Tichý in Prague, the collaboration with Jörg Liesen and Vance Faber during his long-term stay (4.5 years) at TU-Berlin. The other co-authors are Dianne O'Leary and Gérard Meurant. In the thesis we do not present a systematic overview of Krylov subspace methods. We instead concentrate on the introduction to and summary of our contribution to the investigation of Krylov subspace methods. A comprehensive summary about principles and analysis of Krylov subspace methods can be found in the recent book by Liesen and Strakoš [27].

The first part of the thesis consists of two chapters that describe the mathematical background of the considered problems and emphasize the author's contribution to the presented topic. In more detail, we first introduce the reader to the world of Krylov subspace methods (Chapter 1), and then study various issues in the analysis of Krylov subspace methods (Chapter 2). We start with investigating convergence bounds for linear systems with normal matrices (Section 2.1) and with nonnormal matrices (Section 2.2). In Section 2.3 we give the answer to the question when the iterates of an optimal Krylov subspace method can be computed by an algorithm with low memory requirements. Finally, Sections 2.4 and 2.5 are devoted to the error estimation and the behavior of considered algorithms and error estimators in finite precision arithmetic.

In the second part of the thesis, we attach reprints of the above mentioned 15 papers. Below we give the list of these papers sorted alphabetically by names of the authors.

- V. Faber, J. Liesen, and P. Tichý, *The Faber-Manteuffel theorem for linear operators*, SIAM J. Numer. Anal., 46 (2008), pp. 1323–1337.

- V. Faber, J. Liesen, and P. Tichý, *On orthogonal reduction to Hessenberg form with small bandwidth*, Numer. Algorithms, 51 (2009), pp. 133–142.

- V. Faber, J. Liesen, and P. Tichý, *On Chebyshev polynomials of matrices*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 2205–2221.

- V. Faber, J. Liesen, and P. Tichý, *Properties of worst-case GMRES*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 1500–1519.

- J. Liesen and P. Tichý, *The worst-case GMRES for normal matrices*, BIT, 44 (2004), pp. 79–98.

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

- J. Liesen and P. Tichý, *On the worst-case convergence of MR and CG for symmetric positive definite tridiagonal Toeplitz matrices*, Electron. Trans. Numer. Anal., 20 (2005), pp. 180–197.

- J. Liesen and P. Tichý, *On best approximations of polynomials in matrices in the matrix 2-norm*, SIAM J. Matrix Anal. Appl., 31 (2009), pp. 853–863.

- J. Liesen and P. Tichý, *Max-min and min-max approximation problems for normal matrices revisited*, Electron. Trans. Numer. Anal., 41 (2014), pp. 159–166.

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

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

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

- Z. Strakoš and P. Tichý, *Error estimation in preconditioned conjugate gradients*, BIT, 45 (2005), pp. 789–817.

- Z. Strakoš and P. Tichý, *On efficient numerical approximation of the bilinear form $c^*A^{-1}b$*, SIAM J. Sci. Comput., 33 (2011), pp. 565–587.

- P. Tichý, J. Liesen, and V. Faber, *On worst-case GMRES, ideal GMRES, and the polynomial numerical, hull of a Jordan block*, Electron. Trans. Numer. Anal., 26 (2007), pp. 453–473.

Chapter 1

Krylov subspace methods

1.1 Introduction

One of the most powerful tools for solving large and sparse systems of linear algebraic equations is a class of iterative methods called Krylov subspace methods. Their significant advantages like low memory requirements and good approximation properties make them very popular, and they are widely used in applications throughout science and engineering.

Mathematically, Krylov subspace methods are based on projection techniques. Instead of solving a possibly very large problem, the idea is to find approximations in Krylov subspaces of small dimensions. To generate the Krylov subspaces, the central operation is matrix-vector multiplication with the input matrix (or its transpose). Krylov subspaces can be build up using only a function that computes the multiplication of the matrix and a vector, so that the matrix itself never has to be formed or stored explicitly. Hence Krylov subspace methods are particularly well suited for application to large and sparse linear systems. The repeated multiplication with the input matrix may reveal dominant properties of the problem at an early stage and give satisfactory approximations at a low iteration number.

1.2 Krylov subspace methods

This work is concerned with Krylov subspace methods for solving linear algebraic systems

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

where A is a real or complex nonsingular N by N matrix, and b is a real or complex vector of length N. Let x_0 be an initial guess for the solution x, and define the initial residual $r_0 = b - Ax_0$. Krylov subspace methods can be derived from the following *projection process*: The nth iterate x_n, $n = 1, 2, \ldots$, is of the form

$$x_n \in x_0 + \mathcal{S}_n\,, \tag{1.2}$$

where $\mathcal{S}_n$ is some n-dimensional space, called the *search space*. Because of the n degrees of freedom, n constraints are required to make x_n unique. This is done by choosing an n-dimensional space $\mathcal{C}_n$, called the *constraints space*, and by requiring that the nth residual is orthogonal to that space, i.e.,

$$r_n \;=\; b - Ax_n \;\in\; r_0 + A\mathcal{S}_n\,, \qquad r_n \;\perp\; \mathcal{C}_n\,. \tag{1.3}$$

Orthogonality here is meant in the Euclidean inner product. A similar type of projection process appears in many areas of mathematics. The choice of spaces usually depends on properties of A. In particular, if A is Hermitian and positive definite, the typical choice is $\mathcal{C}_n = \mathcal{S}_n$ which corresponds to the Galerkin method. Another important choice for a general case is $\mathcal{C}_n = A\mathcal{S}_n$.

The method defined by the conditions (1.2)-(1.3) is called a *Krylov subspace method* when the spaces $\mathcal{C}_n$ and $\mathcal{S}_n$ are defined by using so-called Krylov subspaces $\mathcal{K}_n(A, r_0)$,

$$\mathcal{K}_n(A, r_0) \;\equiv\; \operatorname{span}\{r_0, Ar_0, \ldots, A^{n-1}r_0\}\,, \quad n = 1, 2, \ldots\,. \tag{1.4}$$

The Krylov subspaces form a nested sequence that ends with a subspace of maximal dimension $d = \dim \mathcal{K}_N(A, r_0)$, i.e.,

$$\mathcal{K}_1(A, r_0) \subset \ldots \subset \mathcal{K}_d(A, r_0) = \ldots = \mathcal{K}_N(A, r_0)\,.$$

Naturally, we are interested in projection methods that ensure existence and uniqueness of their iterates x_n for each step $n \le d$, and that terminate with the exact solution in the step d (such a method will be called *well-defined*). Some properties of A ensure that a method is well-defined.

There are many choices of the spaces $\mathcal{S}_n$ and $\mathcal{C}_n$ where Krylov subspaces are involved. In this work we always choose the search space as the Krylov subspace,

$$\mathcal{S}_n \;\equiv\; \mathcal{K}_n(A, r_0)\,.$$

We mostly limit our discussion to the two important cases of well-defined methods. First, if A is Hermitian and positive definite (HPD), we consider the constrained space

$$\mathcal{C}_n \;\equiv\; \mathcal{K}_n(A, r_0)\,.$$

This choice leads to the construction of orthogonal residual vectors $r_n = b - Ax_n$. Since A is HPD, the method is always well-defined and the generated approximations are optimal in the sense that the errors are minimized in the norm defined by the matrix A. A particular implementation in this case is the conjugate gradient (CG) method [22].

Second, if A is nonsingular, one can choose the constrained space as

$$\mathcal{C}_n \;\equiv\; A\mathcal{K}_n(A, r_0)$$

which leads to residual vectors r_n that have minimal Euclidean norm over the whole affine subspace $r_0 + A\mathcal{K}_n(A, r_0)$. Therefore, the corresponding method is called the minimal residual method. The most well-known implementations are the MINRES method [39] for Hermitian indefinite matrices and the GMRES method [41] for general nonsingular matrices.

The conditions $x_n \in x_0 + \mathcal{K}_n(A, r_0)$ and $r_n \in r_0 + A\mathcal{K}_n(A, r_0)$ imply that the error $e_n = x - x_n$ and the residual r_n can be written in the form

$$e_n = p_n(A)e_0, \qquad r_n = p_n(A)r_0, \tag{1.5}$$

where p_n is a polynomial of degree at most n and with value one at the origin. This is the reason why Krylov subspace methods are sometimes called *polynomial methods.* The form (1.5) is a starting point for the convergence analysis of these methods.

Chapter 2

Analysis of Krylov subspace methods

In exact arithmetic, well-defined Krylov subspace methods terminate in a finite number of steps. Therefore no limit can be formed, and terms like "convergence" or "rate of convergence" loose their classical meaning. This situation requires approaches that are substantially different from the analysis of classical fixed point iteration methods such as Gauß-Seidel or SOR. Krylov subspace methods (in combination with preconditioning) find or should find a good approximate solution to the problem, usually after several iterations. It is therefore important to describe their convergence in the transient phase. As we will see in the following, the question about the convergence behavior leads to bounds represented by complicated nonlinear problems.

Convergence of Krylov subspace methods applied to linear systems with a symmetric matrix (or, in general, with a normal matrix) can be described using classical approximation problems of the form "find the best polynomial approximation of a function on a set of points (the eigenvalues), measured in a suitable norm". For linear systems with a nonnormal matrix, characterization of convergence in terms of classical approximation theory is generally very difficult. This fact is underlined by the classical results [20, 18] which indicate that the distribution of the matrix eigenvalues alone need not determine convergence behavior. We divide our discussion about convergence and convergence bounds into two parts. In the first part (Section 2.1) we consider normal matrices A and show that in this case the spectral information is important for analyzing the convergence. The second part (Section 2.2) shows the difficulties with understanding and bounding the convergence for the nonnormal matrices.

For numerical and implementational reasons, it is often advisable to use orthogonal bases when implementing Krylov subspace methods. For efficiency reasons (low memory requirements) it is desirable to compute such bases with a short recurrence meaning that in each iteration step only a few of the latest basis vectors are required to generate the new basis vector. In Section 2.3 we study the question if and when these two goals can be achieved simultaneously, and also the closely related question of conditions when a given matrix can be orthogonally reduced to upper Hessenberg form with small bandwidth.

An important advantage of Krylov subspace methods (and iterative methods in general) is that one can stop the algorithm at any iteration step and consider the updated vector x_n to be the approximate solution (in contrast to Gauss elimination where we do not have any intermediate approximations). To make the right decision when to stop the algorithm we need to have a mechanism how to measure the quality of the current approximation x_n. Typically, we need *some information* about the size of the error $x - x_n$. In some applications (such as in image processing) the Euclidean norm of the error plays an important role. In many other applications where the matrix A is Hermitian and positive definite, the A-norm of the error is the right measure of convergence. In the nonsymmetric case, one can even use some kinds of bilinear forms that do not represent a norm, but that still measure in some sense the quality of the approximate solution. This topic is discussed in Section 2.4.

Finally, in Section 2.5 we study and analyze some problems related to computations in finite precision arithmetic. To understand what is going on in finite precision arithmetic, it is desirable to create (if it is possible) a *mathematical model* that describes results of computation of the given method (algorithm) in finite precision arithmetic. Using such a model one can justify that the computed results are reliable and meaningful. Another problem consists in *application of mathematical formulas* used, e.g., for estimating some convergence characteristics. Such formulas are often derived based on mathematical assumptions that need not be satisfied during finite precision computations. Consequently, even though a formula can compute very accurately, it can happen that it produces results that do not correspond to the estimated convergence characteristic. Hence, to justify that a mathematical formula for estimating a convergence characteristic works well also during finite precision computations, rounding errors in the whole computation, not only in the computation of the current formula, must be taken into account. For example, in the conjugate gradient method, the computed residual vectors become typically non-orthogonal (and even linearly dependent) after a few iterations. Then, without a proper rounding error analysis one can never be sure that mathematical formulas for estimating the A-norm derived assuming the exact orthogonality of the residual vectors really produce results that correspond to the A-norm of the actual error.

2.1 Convergence bounds – normal matrices

Consider a nonsingular and *normal* matrix A, and let

$$A = V\Lambda V^*, \qquad \text{where} \qquad V^*V = I, \qquad \Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_N)\,,$$

be its eigendecomposition (the superscript $*$ denotes Hermitian transposed). The orthogonality of the eigenvector basis leads to a significant simplification in the convergence analysis of Krylov subspace methods: Considering A^n in the form $V\Lambda^n V^*$ and using (1.5), the errors and residuals of a Krylov subspace method satisfy

$$e_n = V p_n(\Lambda) V^* e_0, \qquad r_n = V p_n(\Lambda) V^* r_0\,. \tag{2.1}$$

The projection property usually refers to some sort of optimality, and we can expect that Krylov subspace methods for normal matrices solve some weighted polynomial minimization problem on the matrix spectrum. In the following we explain that in the worst case, the convergence of well-known Krylov subspace methods (CG, MINRES, GMRES) is determined by the value

$$\min_{p \in \pi_n} \max_k |p(\lambda_k)| \,, \tag{2.2}$$

where π_n denotes the set of polynomials of degree at most n and with value one at the origin. Note that the value (2.2) represents a min-max approximation problem on the discrete set of the matrix eigenvalues. The value (2.2) depends in a complicated (nonlinear) way on the eigenvalue distribution. Assume, for simplicity, that all eigenvalues are real and distinct. The results in [13, 29] show that there exists a subset of $n+1$ (distinct) eigenvalues $\{\mu_1, \ldots, \mu_{n+1}\} \subseteq \{\lambda_1, \ldots, \lambda_N\}$, such that

$$\min_{p \in \pi_n} \max_k |p(\lambda_k)| \;=\; \left(\sum_{j=1}^{n+1} \prod_{\substack{k=1 \\ k \neq j}}^{n+1} \frac{|\mu_k|}{|\mu_k - \mu_j|} \right)^{-1} . \tag{2.3}$$

If at least one eigenvalue of A is complex, the equality (2.3) does not hold in general, cf. [29]. Nevertheless, in [29] we formulate a conjecture, supported by numerical experiments and by some theoretical results, that there exist a set of $n+1$ eigenvalues such that the value on the right hand side of (2.3) is equal to (2.2) up to a factor between 1 and $4/\pi$.

Consider a *Hermitian positive definite* matrix A. Each such matrix defines a norm (the so-called A-norm),

$$\|u\|_A \;=\; (u^* A u)^{\frac{1}{2}} \,. \tag{2.4}$$

It is well known that, for the choice $\mathcal{S}_n = \mathcal{C}_n = \mathcal{K}_n(A, r_0)$, the Krylov subspace iterates x_n are uniquely defined in each iterative step n and can be computed using the CG method. The CG errors $e_n = x - x_n$ satisfy

$$\|e_n\|_A \;=\; \min_{p \in \pi_n} \|p(A) e_0\|_A \,. \tag{2.5}$$

A simple algebraic manipulation shows that the value (2.2) represents an upper bound on the relative A-norm of the error,

$$\frac{\|e_n\|_A}{\|e_0\|_A} \;\leq\; \min_{p \in \pi_n} \max_k |p(\lambda_k)| \,. \tag{2.6}$$

This convergence bound is sharp, i.e., for each iteration step n there exist a right hand side b or an initial guess x_0 (depending on n and A) such that equality holds in (2.6), see [13] and [32]. In this sense, the bound (2.6) completely describes the *worst-case behavior* of the CG method (for a given matrix A). When the whole spectrum of A is known, one can try to determine the value of the right hand side of (2.6) using the formula (2.3). However, it is in general an open problem which subset of $n+1$ eigenvalues leads to equality in (2.3).

Obviously, the bound (2.6) depends only on the matrix eigenvalues and not on any other properties of A, b, or x_0. If a particular right hand side b is known, it is sometimes possible to incorporate the information about b into the analysis, and thus to obtain a better estimate of the actual convergence behavior.

The convergence behavior of the CG method is relatively well understood, but some open problems still remain. The right approach for investigating the convergence behavior is to use all information about the eigenvalue distribution we have at our disposal. If a particular right hand side b and initial guess x_0 are given, they should be incorporated in the analysis. An example for such an approach for the model problem of the one-dimensional Poisson equation is given in our paper [28].

Consider now a nonsingular and *normal* matrices A. It is well known that the iterates x_n of the minimal residual Krylov subspace method are for any such matrix uniquely defined in each iterative step n, and that the nth residual $r_n = b - Ax_n$ satisfies

$$\|r_n\| = \min_{p \in \pi_n} \|p(A)r_0\|. \tag{2.7}$$

In general, no strict monotonicity of the residual norms is guaranteed. In particular, for any (finite) nonincreasing sequence of numbers one can find a normal A and a right hand side b such that the minimal residual method exhibits the prescribed convergence behavior [1, 18]. That normal matrix can even be chosen to be unitary. In the normal case, the relative residual norm of the minimal residual method can be bounded similarly as in (2.6),

$$\frac{\|r_n\|}{\|r_0\|} \leq \min_{p \in \pi_n} \max_k |p(\lambda_k)| \tag{2.8}$$

and again, the bound (2.8) is sharp [17, 25, 32]. If full spectral information is available, then the approach in [29] (cf. the discussion of formula (2.3)) can be used for estimating the worst-case convergence behavior.

In our recent paper [32] we answer the question of the sharpness of the bounds (2.6) and (2.8) in a very general setting, using classical results of approximation theory. In particular, let A be a real or complex square matrix, i.e., $A \in \mathbb{F}^{N \times N}$ with $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$. Suppose that f and $\varphi_1, \ldots, \varphi_n$ are given (scalar) functions so that $f(A) \in \mathbb{F}^{N \times N}$ and $\varphi_1(A), \ldots, \varphi_n(A) \in \mathbb{F}^{N \times N}$ are well defined matrix functions in the sense of [23, Definition 1.2]. Let $\mathcal{P}_n(\mathbb{F})$ denote the linear span of the functions $\varphi_1, \ldots, \varphi_n$ with coefficients in $\mathbb{F}$, so that in particular $p(A) \in \mathbb{F}^{N \times N}$ for each linear combination $p = \alpha_1\varphi_1 + \ldots + \alpha_k\varphi_n \in \mathcal{P}_n(\mathbb{F})$.

With this notation, the optimality property of many useful methods of numerical linear algebra can be formulated as an approximation problem of the form

$$\min_{p \in \mathcal{P}_n(\mathbb{F})} \|f(A)v - p(A)v\|, \tag{2.9}$$

where $v \in \mathbb{F}^N$ is a given vector and $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{F}^N$. An example of such a method is the GMRES method for solving $Ax = b$ with $A \in \mathbb{F}^{N \times N}$, $b \in \mathbb{F}^N$, and the initial guess $x_0 \in \mathbb{F}^N$. Its optimality property is of the form (2.9) with $f(z) = 1$, $\varphi_i(z) = z^i$ for $i = 1, \ldots, n$, and $v = b - Ax_0$.

If the given vector v has unit norm, which usually can be assumed without loss of generality, then an upper bound on (2.9) is given by

$$\min_{p \in \mathcal{P}_n(\mathbb{F})} \|f(A) - p(A)\|, \tag{2.10}$$

where $\|\cdot\|$ denotes the matrix norm associated with the Euclidean vector norm, i.e., the matrix 2-norm or spectral norm on $\mathbb{F}^{N \times N}$. In (2.10) we seek a best approximation (with respect to the given norm) of the matrix $f(A) \in \mathbb{F}^{N \times N}$ from the subspace of $\mathbb{F}^{N \times N}$ spanned by the matrices $\varphi_1(A), \ldots, \varphi_n(A)$. An example of this type is the Chebyshev matrix approximation problem with $A \in \mathbb{F}^{N\times N}$, $f(z) = z^n$, and $\varphi_i(z) = z^{i-1}$, $i = 1, \ldots, n$. This problem was introduced in [21] and later studied, for example, in [49] and [7].

In order to analyze how close the upper bound (2.10) can possibly be to the quantity (2.9), one can maximize (2.9) over all unit norm vectors $v \in \mathbb{F}^N$ and investigate the sharpness of the inequality

$$\max_{\substack{v\in\mathbb{F}^N\\ \|v\|=1}} \min_{p\in\mathcal{P}_n(\mathbb{F})} \|f(A)v - p(A)v\| \quad\leq\quad \min_{p\in\mathcal{P}_n(\mathbb{F})} \|f(A) - p(A)\|\,. \tag{2.11}$$

From analyses of the GMRES method it is known that the inequality (2.11) can be strict. For example, certain nonnormal matrices $A \in \mathbb{R}^{4\times 4}$ were constructed in [4, 48], for which (2.11) is strict with $n = 3$, $f(z) = 1$, and $\varphi_i(z) = z^i$, $i = 1, 2, 3$. More recently, nonnormal matrices $A \in \mathbb{R}^{2N\times 2N}$, $N \geq 2$, were derived in [8], for which the inequality (2.11) is strict for all $n = 3, \ldots, 2N-1$, $f(z) = 1$, and $\varphi_i(z) = z^i$, $i = 1, \ldots, n$. On the other hand, the following result is well known. *Under the assumptions made above, if $A \in \mathbb{F}^{N\times N}$ is normal, then equality holds in (2.11).*

At least three different proofs of this theorem or variants of it can be found in the literature. Greenbaum and Gurvits proved it for $\mathbb{F} = \mathbb{R}$ using mostly methods from matrix theory; see [17, Section 2]. Using (analytic) methods of optimization theory, Joubert proved the equality for the case of the GMRES method with $f(z) = 1$, $\varphi_i(z) = z^i$, $i = 1, \ldots, n$, and he distinguished the cases $\mathbb{F} = \mathbb{R}$ and $\mathbb{F} = \mathbb{C}$; see [25, Theorem 4]. Finally, Bellalij, Saad, and Sadok also considered the GMRES case with $\mathbb{F} = \mathbb{C}$, and they applied methods from constrained convex optimization; see [2, Theorem 2.1].

In [32] we present yet another proof of this statement which is rather simple because it fully exploits the link between matrix approximation problems for normal matrices and scalar approximation problems in the complex plane. We observe that when formulating the matrix approximation problems in (2.11) in terms of scalar approximation problems, the proof reduces to a straightforward application of a well-known characterization theorem of best approximation in the complex plane; see, e.g., [33, Theorem 3, p. 22] or [40, pp. 672-674]). While the proof of the theorem for $\mathbb{F} = \mathbb{C}$ can be accomplished in just a few lines, the case $\mathbb{F} = \mathbb{R}$ contains some technical details that require additional attention.

2.2 Convergence bounds – nonnormal matrices

In this section we consider the case of a general nonsingular and *nonnormal* matrix A. In this general case, a minimal residual Krylov subspace method such as GMRES yields uniquely defined iterates x_n so that the nth residual $r_n = b - Ax_n$ satisfies (2.7). Similarly to the convergence analysis for normal matrices presented above, we are interested in finding a (sharp) bound on the right hand side of (2.7).

If A is diagonalizable,

$$A = V\Lambda V^{-1}\,, \qquad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)\,,$$

then the following convergence bound easily follows from (2.7),

$$\frac{\|r_n\|}{\|r_0\|} \;=\; \min_{p\in\pi_n} \frac{\|Vp(\Lambda)V^{-1}r_0\|}{\|r_0\|} \;\leq\; \kappa(V)\, \min_{p\in\pi_n}\max_{k} |p(\lambda_k)|\,. \tag{2.12}$$

Here $\kappa(V) = \|V\|\,\|V^{-1}\|$ denotes the condition number of the eigenvector matrix V. A bound similar to (2.12) can be derived for nondiagonalizable matrices.

The bound (2.12) frequently is the basis for discussions of the GMRES convergence behavior. As mentioned in the previous section, this bound is sharp when A is normal. When $\kappa(V)$ is small, the right hand side of (2.12) typically represents a good convergence bound, and its value can be estimated. However, when V is far from unitary, the bound (2.12) may fail to provide any reasonable information about the GMRES convergence. Apart from the fact, that the factor $\kappa(V)$ can be very large in case of ill-conditioned eigenvectors, the principal weakness of the bound (2.12) is that the min-max problem on the matrix eigenvalues need not have any connection with the GMRES convergence, since the eigenvalues may have nothing to do with the convergence behavior at all [1, 18]. As a consequence, the curve produced by the min-max approximations on matrix eigenvalues can be substantially different from the (worst-case) GMRES convergence curve and the bound can fail to give any reasonable convergence information. On the other hand, it needs to be stressed that from an analytic point of view the principal difficulty of nonnormality is *not* the often met belief that the convergence is slower for nonnormal than for normal matrices.

It should be clear by now that in the nonnormal case the GMRES convergence behavior is *significantly more difficult to analyze* than in the normal case. A general approach to understand the worst-case GMRES convergence in the nonnormal case is to replace the complicated minimization problem (2.7) by another one that is easier to analyze and that, in some sense, approximates the original problem (2.7). Natural bounds on the GMRES residual norm arise by excluding the influence of the initial residual r_0,

$$\begin{align} \frac{\|r_n\|}{\|r_0\|} &= \min_{p\in\pi_n} \frac{\|p(A)r_0\|}{\|r_0\|} && \text{(GMRES)} \notag\\ &\le \max_{\|v\|=1}\min_{p\in\pi_n} \|p(A)v\| && \text{(worst-case GMRES)} \tag{2.13}\\ &\le \min_{p\in\pi_n} \|p(A)\| && \text{(ideal GMRES).} \tag{2.14} \end{align}$$

The bound (2.13) corresponds to the *worst-case* GMRES behavior and represents a sharp upper bound, i.e. a bound that is attainable by the GMRES residual norm. In this sense, (2.13) is the best bound on $\|r_n\|/\|r_0\|$ that is independent of r_0. Despite the independence of r_0, it is not clear in general, which properties of A influence the bound (2.13); see, e.g., [4]. The expression (2.13) can be bounded by the *ideal* GMRES approximation problem (2.14), which was introduced by Greenbaum and Trefethen [21]. To justify the relevance of the bound (2.14), several researchers tried to identify cases in which (2.13) is equal to (2.14). The best known result of this type is that (2.13) is equal to (2.14) whenever A is normal [17, 25, 32]. Despite the existence of some counterexamples [4, 48], it is still an open question whether (2.13) is equal or close to (2.14) for larger classes of nonnormal matrices. In [47] we consider this problem for a Jordan block, a representative of a nonnormal matrix, and prove equality of the expressions (2.13) and (2.14) in some steps.

The main goal of our paper [8] is to contribute to the understanding of the worst-case GMRES approximation problem (2.13). We show that the worst case behavior of GMRES for the matrices A and A^* is the same, and we analyze properties of initial vectors for which the worst-case residual norm is attained. In particular, we prove that such vectors satisfy a certain "cross equality". We show that the worst-case GMRES polynomial may not be uniquely determined, and we consider the relation between the worst-case and the ideal GMRES approximations, giving new examples in which the inequality between the two quantities is strict at all iteration steps $n \geq 3$. Finally, we give a complete characterization of how the values of the approximation problems change in the context of worst-case and ideal GMRES for a real matrix, when one considers complex (rather than real) polynomials and initial vectors.

A possible way to approximate the value of the matrix approximation problem (2.14) is to determine sets $\Omega \subset \mathbb{C}$ and $\hat{\Omega} \subset \mathbb{C}$, that are somehow associated with A, and that provide lower and upper bounds on (2.14),

$$c_1 \min_{p\in\pi_n} \max_{z\in\Omega} |p(z)| \;\leq\; \min_{p\in\pi_n} \|p(A)\| \;\leq\; c_2 \min_{p\in\pi_n} \max_{z\in\hat{\Omega}} |p(z)|.$$

Here c_1 and c_2 should be some (moderate size) constants depending on A and possibly on n. This approach represents a generalization of the idea for normal matrices, where the appropriate set associated with A is the spectrum of A and $c_1 = c_2 = 1$.

One approach is to take $\hat{\Omega}$ to be the *field of values* of A,

$$\mathcal{F}(A) = \left\{ v^* A v \;:\; \|v\| = 1\,,\; v \in \mathbb{C}^N \right\}\,.$$

A generalization of the field of values of A is the *polynomial numerical hull*, introduced by Nevanlinna [36], and defined as

$$\mathcal{H}_n(A) = \{ z \in \mathbb{C} : \|p(A)\| \geq |p(z)| \text{ for all } p \in \mathcal{P}_n \}\,,$$

where $\mathcal{P}_n$ denotes the set of polynomials of degree n or less. It can be shown that $\mathcal{F}(A) = \mathcal{H}_1(A)$. The set $\mathcal{H}_n(A)$ provides a lower bound on (2.14),

$$\min_{p\in\pi_n} \max_{z\in\mathcal{H}_n(A)} |p(z)| \;\leq\; \min_{p\in\pi_n} \|p(A)\|. \tag{2.15}$$

In some way, $\mathcal{H}_n(A)$ reflects the complicated relation between the polynomial of degree n and the matrix A, and provides often a very good estimate of the value of the ideal GMRES approximation (2.14). Greenbaum and her co-workers [3, 15, 16] have obtained theoretical results about $\mathcal{H}_n(A)$ for Jordan blocks, banded triangular Toeplitz matrices and block diagonal matrices with triangular Toeplitz blocks. Clearly, for a larger applicability of the bound (2.15), the class of matrices for which $\mathcal{H}_n(A)$ is known needs to be extended. But in general, the determination of these sets represents a nontrivial open problem.

In [47] we investigate the bound (2.15) for a single Jordan block J_λ. We study the relation between ideal and worst-case GMRES approximations (2.13) and (2.14) as well as the problem of estimating the ideal GMRES approximation using the set $\mathcal{H}_n(J_\lambda)$. We prove new results about the radii of the polynomial numerical hulls of Jordan blocks. Using these, we discuss the closeness of the lower bound on the ideal GMRES approximation that is derived from the radius of the polynomial numerical hull.

The ideal GMRES approximation problem is a special case of more general matrix approximation problems that are not well understood. In our paper [31] we consider the following problem: Let f be a function that is analytic in a neighborhood of the spectrum of a given matrix $A \in \mathbb{C}^{N\times N}$, so that $f(A)$ is well defined, let $\|\cdot\|$ be the spectral norm (2-norm) and let m be a nonnegative integer. Consider the *matrix approximation problem*

$$\min_{p \in \mathcal{P}_n} \|f(A) - A^m p(A)\|, \tag{2.16}$$

where $\mathcal{P}_n$ is the set of polynomials of degree at most n. Special cases of the problem (2.16) are the *ideal Arnoldi* and *ideal GMRES* approximation problems. Greenbaum and Trefethen [21] seem to be the first who studied existence and uniqueness of polynomials that solve ideal Arnoldi and ideal GMRES problems. In our paper [31] we generalize their results to problems of the form (2.16). Our main result is the following*:* Provided that the minimum in (2.16) is nonzero and A is nonsingular, the problem (2.16) has *a unique minimizer.* In the subsequent paper [7] we study general properties of so called Chebyshev polynomials of matrices, the polynomials that solve the ideal Arnoldi approximation problem. In some cases, these properties turn out to be generalizations of well known properties of Chebyshev polynomials of compact sets in the complex plane.

2.3 Short recurrences

At the Householder Symposium VIII held in Oxford in July 1981, Golub posed as an open question to characterize necessary and sufficient conditions on a matrix A for the existence of a three-term conjugate gradient type method for solving linear systems with A (cf. SIGNUM Newsletter, vol. 16, no. 4, 1981). This important question was answered by Faber and Manteuffel in 1984 [9]. They formulated a fundamental theorem in the area of iterative methods known as the Faber-Manteuffel theorem. It shows that a short recurrence for orthogonalizing Krylov subspace bases for a matrix A exists if and only if the adjoint of A is a low degree polynomial in A. This result is important, since it characterizes all matrices, for which an optimal Krylov subspace method with short recurrences can be constructed. Here optimal means that the error is minimized in the norm induced by the given inner product. Of course, such methods are highly desirable, due to convenient work and storage requirements for generating the orthogonal basis vectors. Examples are the CG method [22] for solving systems of linear algebraic equations with a symmetric positive definite matrix A, or the MINRES method [39] for solving symmetric but indefinite systems.

Now we briefly describe the result of Faber and Manteuffel. Let A be a nonsingular matrix and v be a vector of grade d (d is the degree of the uniquely determined monic polynomial of smallest degree that annihilates v). For theoretical as well as practical purposes it is often convenient to orthogonalize the basis $v, \dots, A^{d-1}v$ of the subspace $\mathcal{K}_d(A, v)$. The classical approach to orthogonalization is to use the Arnoldi method, that produces mutually orthogonal vectors $v_1, \dots, v_d$ satisfying $\text{span}\{v_1, \dots, v_n\} = \text{span}\{v, \dots, A^{n-1}v\}$, $n = 1, \dots, d$. The algorithm can be written in a matrix form

$$v_1 = v\,, \tag{2.17}$$

$$A \underbrace{[v_1,\ldots,v_{d-1}]}_{\equiv V_{d-1}} = \underbrace{[v_1,\ldots,v_d]}_{\equiv V_d} \underbrace{\begin{bmatrix} h_{1,1} & \cdots & h_{1,d-1} \\ 1 & \ddots & \vdots \\ & \ddots & h_{d-1,d-1} \\ & & 1 \end{bmatrix}}_{\equiv H_{d,d-1}}, \tag{2.18}$$

$$(v_i, v_j) = 0 \quad \text{for} \quad i \neq j\,, \quad i,j = 1,\ldots,d\,. \tag{2.19}$$

As described above, for efficiency reasons, it is desirable to generate such an orthogonal basis with a short recurrence, meaning that in each iteration step only a few of the latest basis vectors are required to generate the new basis vector. This corresponds to the situation when the matrix $H_{d,d-1}$ in (2.18) is, for each starting vector v_1, low-band Hessenberg matrix. Note that an unreduced upper Hessenberg matrix is called $(s+2)$-band Hessenberg, when its s-th superdiagonal contains at least one nonzero entry, and all its entries above its s-th superdiagonal are zero. We say that A *admits an optimal* $(s+2)$*-term recurrence* if $H_{d,d-1}$ is for each starting vector at most $(s+2)$-band Hessenberg and, moreover, there exists an initial vector such that $H_{d,d-1}$ is exactly $(s+2)$-band Hessenberg (the s-th superdiagonal contains at least one nonzero entry). The fundamental question is, what properties are necessary and sufficient for A to admit an optimal $(s+2)$-term recurrence. This question was answered by Faber and Manteuffel [9].

THEOREM (Faber-Manteuffel) *Let A be a nonsingular matrix with minimal polynomial degree $d_{\min}(A)$. Let s be a nonnegative integer, $s+2 < d_{\min}(A)$. Then A admits an optimal $(s+2)$-term recurrence if and only if $A^* = p(A)$, where p is a polynomial of smallest degree s having this property (i.e. A is normal(s)).*

While the sufficiency of the normal(s) condition is rather easy to prove, the proof of necessity given by Faber and Manteuffel is based on a clever, highly nontrivial construction by using results from mathematical analysis ("continuous function"), topology ("closed set of smaller dimension") or multilinear algebra ("wedge product").

In [26], Liesen and Strakoš discuss and clarify the existing important results in the context of the Faber-Manteuffel Theorem. They suggest that, in light of the fundamental nature of the result, it is desirable to find an alternative, and possibly simpler proof of the necessity part.

In our paper [5] we address this issue. We formulate here this theorem in terms of linear operators on finite dimensional Hilbert spaces. We have chosen this setting because we believe that the proof of necessity is easier to follow when we use linear operators rather than matrices. We give two different proofs of the necessity part, both based on restriction of the linear operator A to certain cyclic invariant subspaces. The resulting technicalities in the matrix formulation would obstruct rather than help the understanding. Moreover, our formulation may serve as a starting point for extending the results to infinite dimensional spaces. We are not aware that any such extensions have been obtained yet.

In the subsequent paper [6] we study a problem that is closely related to Faber-Manteuffel theorem. In particular, at the dth iteration step the relation (2.18) becomes

$$AV_d = V_d H_d\,. \tag{2.20}$$

Here H_d can be interpreted as the matrix representation of the linear operator A restricted to the A-invariant subspace $\mathcal{K}_d(A, v)$. Or, H_d can be interpreted as a *reduction of A to upper Hessenberg form.* In [6] we study necessary and sufficient conditions on A so that the orthogonal Hessenberg reduction yields a Hessenberg matrix with small bandwidth. This includes the orthogonal reduction to tridiagonal form as a special case. Orthogonality here is meant with respect to some given but unspecified inner product. We prove the following theorem.

THEOREM *Let $A \in \mathbb{C}^{N\times N}$, let $B \in \mathbb{C}^{N\times N}$ be a Hermitian positive definite matrix, and let s be a nonnegative integer, $s + 2 \leq d_{\min}(A)$. The matrix A is reducible for the given B to $(s + 2)$-band Hessenberg form if and only if $A^* = Bp(A)B^{-1}$, where p is a polynomial of smallest degree s having this property (i.e. A is B-normal(s)).*

While this result is already implied by the Faber-Manteuffel theorem on short recurrences for orthogonalizing Krylov sequences (see [26]), we consider it useful to present a new, less technical proof. Our proof utilizes the idea of a “minimal counterexample”, which is standard in combinatorial optimization, but rarely used in the context of linear algebra.

2.4 Error estimation and related problems

In this section we concentrate on two problems: Error estimation in the conjugate gradient (CG) method that solves systems of linear algebraic equations $Ax = b$ with Hermitian and positive definite matrices, and approximation of the bilinear form $c^*A^{-1}b$.

Today the (preconditioned) Conjugate Gradient (CG) algorithm by Hestenes and Stiefel [22] is the iterative method of choice for solving linear systems $Ax = b$ with a Hermitian positive definite symmetric matrix A. An important question is when to stop the iterations. Ideally, one would like to stop the iterations when some norm of the error $e_n = x - x_n$, where x_n are the CG iterates, is small enough. However, the error is unknown and most CG implementations rely on stopping criteria that use the residual norm $\|r_n\| = \|b - Ax_n\|$ as a measure of convergence. These types of stopping criteria can provide misleading information about the actual error. It can stop the iterations too early when the norm of the error is still too large, or too late in which case too many floating point operations have been done for obtaining the required accuracy. This motivated researchers to look for ways to compute estimates of some norms of the error during CG iterations. The norm of the error which is particularly interesting for CG is the A-norm which is minimized at each iteration,

$$\|e_n\|_A = (e_n^* A e_n)^{1/2}.$$

Inspired by the connection of CG with the Gauss quadrature rule for a Riemann-Stieltjes integral, a way of research on this topic was started by Gene Golub in the 1970s and continued throughout the years with several collaborators (e.g., Dahlquist, Eisenstat, Fischer, Meurant, Strakoš). The main idea of Golub and his collaborators was to obtain bounds for the integral using different quadrature rules. It turns out that these bounds can be computed without the knowledge of the stepwise constant measure and at almost no cost during the CG iterations.

These techniques were used by Golub and Meurant [11] for providing lower and upper bounds on quadratic forms $u^*f(A)u$ where f is a smooth function, A is a Hermitian matrix and u is a given vector. Their algorithm GQL (Gauss Quadrature and Lanczos) was based on the Lanczos algorithm and on computing functions of Jacobi matrices. Later, these techniques were adapted to CG to compute lower and upper bounds on the A-norm of the error for which the function is $f(\lambda) = \lambda^{-1}$. The idea was to use CG instead of the Lanczos algorithm, to compute explicitly the entries of the corresponding Jacobi matrices and their modifications from the CG coefficients, and then to use the same formulas as in GQL. The formulas were summarized in the CGQL algorithm (QL standing again for Quadrature and Lanczos), whose most recent version is described in the book [12]. Below we describe our contribution to the algorithmic development of the error estimators in CG.

The CG method of Hestenes and Stiefel is given by Algorithm 1).

Algorithm 1 Conjugate gradient algorithm

```
**input** $A$, $b$, $x_0$
$r_0 = b - Ax_0$
$p_0 = r_0$
**for** $n = 1, \ldots, N$ until convergence **do**
    $\gamma_{n-1} = \frac{r_{n-1}^* r_{n-1}}{p_{n-1}^* A p_{n-1}}$
    $x_n = x_{n-1} + \gamma_{n-1} p_{n-1}$
    $r_n = r_{n-1} - \gamma_{n-1} A p_{n-1}$
    $\delta_n = \frac{r_n^* r_n}{r_{n-1}^* r_{n-1}}$
    $p_n = r_n + \delta_n p_{n-1}$
**end for**
```

In the Ph.D. thesis [46] and in [43] we summarize the connection between CG and Gauss quadrature, and compare and analyze various ways how to compute the Gauss quadrature lower bound on the A-norm of the error. In the end, we recommend the estimate

$$\|e_n\|_A \approx \left(\sum_{i=n}^{n+k-1} \gamma_i \|r_i\|^2 \right)^{1/2} \tag{2.21}$$

to be incorporated into any software realization of the CG method. It is simple and numerically stable. Note that to compute the estimate we need to perform k extra steps of CG. It remains a subject of further work to design an adaptive error estimator, which would use some heuristics for adjusting k according to the desired accuracy of the estimate. Though we concentrate in [43] mostly on the lower bound for the A-norm of the error, we describe also an estimate for the Euclidean norm of the error. Further extension and popularization of our results [43] to practical users of the preconditioned conjugate gradient method (PCG) we published in [44]. Note that the estimate (2.21) appeared later to be very useful not only in the context of CG, but also, e.g., in stopping criteria for rational matrix functions of Hermitian and symmetric matrices [10] or, in a posteriori error estimates for the finite volume discretization of a second-order elliptic model problem, which take into account an inexact solution of the associated linear algebraic system [24].

The CGQL algorithm may seem complicated, particularly for computing bounds with the Gauss-Radau or Gauss-Lobatto quadrature rules. In our paper with Meurant

[35] we show that the CGQL formulas can be considerably simplified. We use the fact that CG computes the Cholesky decomposition of the Jacobi matrix which is given implicitly, and derive new algebraic formulas by working with the LDL^* factorizations of the Jacobi matrices and their modifications instead of computing the Lanczos coefficients explicitly. In other words, we obtain the bounds from the CG coefficients without computing the Lanczos coefficients. The new algorithm is called CGQ (Conjugate Gradients and Quadrature). The algebraic derivation of the new formulas is more difficult than it was when using Jacobi matrices but, in the end, the formulas are simpler. Obtaining simple formulas is a prerequisite for analyzing the behavior of the bounds in finite precision arithmetic and also for a better understanding of their dependence on the auxiliary parameters that are lower and upper bounds (or estimates) of the smallest and the largest eigenvalue of A.

Given a nonsingular square matrix $A \in \mathbb{C}^{N\times N}$ and vectors b and c of compatible dimensions, many applications require approximation of the quantity

$$c^*A^{-1}b \;=\; c^*x \tag{2.22}$$

where x is the solution of the linear algebraic system $Ax = b$. They arise in signal processing under the name scattering amplitude, as well as in nuclear physics, quantum mechanics, computational fluid dynamics. In numerical linear algebra they arise naturally in computing error bounds for iterative methods, in solving inverse problems, least and total least squares problems etc.; see [12].

Usually, $c^*A^{-1}b$ need not be approximated to a high accuracy; an approximation correct to very few digits of accuracy is sufficient. Therefore direct solution of $Ax = b$ is even for problems of moderate size inefficient. If A is sufficiently large or the elements of A are too costly to compute, then the direct solution is not possible. A strategy used by several authors is to generate a sequence $\{x_n\}$ of approximate solutions to $Ax = b$ using a Krylov subspace method, and to approximate $c^*A^{-1}b$ by c^*x_n for sufficiently large n. However, even when A is HPD, this approximation may require a large number of iterations as a result of rounding errors affecting x_n; see [43, 44]. In our paper [45] we presents an approach for approximating $c^*A^{-1}b$ that is designed to be computationally efficient. Algorithmically, this paper extends the results presented in [43, 44] from the HPD case and the conjugate gradient method (CG) to the general complex case and the biconjugate gradient method (BiCG). In more detail, the efficient approximation is based on the identity

$$c^*A^{-1}b = \sum_{j=0}^{n-1} \alpha_j s_j^* r_j + s_n^* A^{-1} r_n\,. \tag{2.23}$$

where r_n and s_n are residual and dual residual vectors generated by the BiCG method. The sum in (2.23) is easily computable during BiCG computations and it provides an approximation to the quantity $c^*A^{-1}b$. Note that (2.23) generalizes the result from the Hermitian positive definite case, in which $b^*A^{-1}b$ and $r_n^*A^{-1}r_n$ equal, respectively, to the squared A-norms of the errors at steps 0 and n; see [43]. In our paper [45] we also show the mathematical equivalence of the preferred approximation based on (2.23) to the existing estimates which use a complex generalization of Gauss quadrature, and discuss its numerical properties. The proposed estimate is compared with existing approaches using analytic arguments and numerical experiments on a practically important problem that arises from the computation of diffraction of light on media with periodic structure.

In our understanding, various approaches for numerical approximation of the quantity $c^*A^{-1}b$ can be viewed as applications of the general mathematical concept of *matching moments model reduction*, formulated and used in applied mathematics by Vorobyev in his book [50]. Using the Vorobyev moment problem, matching moments properties of Krylov subspace methods can be described in a very natural and straightforward way, see [42].

2.5 Influence of finite precision arithmetic

Almost all practical computations are done on computers that use finite precision arithmetic. Hence, understanding the behavior of an algorithm in the presence of rounding errors is very important. In iterative methods, any stopping criterion and also any theoretical consideration about convergence in solving a practical problem has to take into account that rounding errors may delay the convergence, limit the attainable accuracy, or influence significantly the information provided by error estimators. The goal of rounding error analysis is then to find algorithms that are numerically stable and to identify algorithms (or their parts) which are not. In the following we concentrate on behavior of the CG and Lanczos algorithms in finite precision arithmetic, and on reliability of error estimators in CG. We also briefly comment on the numerical behavior of the approximation to the quantity $c^*A^{-1}b$ based on the identity (2.23) and the BiCG method.

For more than 20 years the effects of rounding errors to the Lanczos and CG methods seemed devastating. Orthogonality among the computed vectors was usually lost very quickly, with a subsequent loss of linear independence. Consequently, the finite termination property was lost. Still, despite a total loss of orthogonality among the vectors, the Lanczos and the CG methods produced reasonable results. A fundamental work was done by Paige who proved that loss of orthogonality among the computed Lanczos vectors was possible only in the directions of the converged Ritz vectors; see, e.g., [38] or the review paper [34]. Another step was made by Greenbaum in [14]. On the foundations laid by Paige she developed a backward-like analysis of the Lanczos algorithm (and also of the closely related conjugate gradient algorithm). Roughly speaking, she showed that Lanczos (and CG) computations can be simulated (modeled) for a given number of iterations by the *exact* Lanczos (and CG) applied to a larger system with the matrix having eigenvalues clustered in small intervals about the eigenvalues of the original matrix. Using the relation between CG and Gauss quadrature, the results by Greenbaum [14] and Greenbaum and Strakoš [19] can also be formulated in the following way: Finite precision CG computations can be viewed as computations of exact CG applied to a modified problem, for which the convergence is determined by a Riemann-Stieltjes integral with a slightly perturbed distribution function of the original problem.

To further understand the mathematical model that describes results of CG computation we investigated in [37] the problem of sensitivity of Gauss-Christoffel quadrature with respect to small perturbations of the distribution function. Note that the question how much does a function change under perturbations of its arguments is of central importance in numerical computations. In more detail, consider a sufficiently smooth integrated function uncorrelated with the perturbation of the distribution function. Then it seems natural that given the same number of function evaluations, the difference between the quadrature approximations is of the same order as the difference between the (original and perturbed) approximated integrals. That is perhaps one of the reasons why, to our knowledge, the sensitivity question has not been formulated and addressed in the literature, though several other sensitivity problems, motivated, in particular, by computation of the quadrature nodes and weights from moments, have been thoroughly studied by many authors. In [37] we survey existing particular results and show that even a small perturbation of a distribution function can cause large differences in Gauss-Christoffel quadrature estimates. This can happen for analytic integrands and discontinuous, continuous, and even analytic distribution functions. We also discuss conditions under which the Gauss-Christoffel quadrature is insensitive under perturbation of the distribution function, present illustrative examples, and relate our observations to known conjectures on some sensitivity problems.

Consider now the CG finite precision computations. As described in Section 2.4, it is desirable to control the quality of the actual approximate solution which can be done using error estimators like (2.21). Rounding errors in the computation of the sum in (2.21) do not represent a problem. However, does this sum really approximate the A-norm of the actual error even though the computed approximate solution is far away (the orthogonality is lost) from its exact precision counterpart? The answer was given in our papers [43] for CG and in [44] for preconditioned CG.

The error estimator (2.21) is based on the identity

$$\|x - x_n\|_A^2 = \sum_{i=n}^{n+k-1} \gamma_i \|r_i\|^2 + \|x - x_{n+k}\|_A^2 \tag{2.24}$$

that holds in exact arithmetic. If $\|x - x_n\|_A^2 \gg \|x - x_{n+k}\|_A^2$, then the square root of the sum in (2.24) is a tight estimate of $\|x - x_n\|_A$. In [43] we have shown that this identity holds (up to some small inaccuracy) also for numerically computed quantities, even though they do not usually correspond to their exact precision counterparts. This result holds thanks to the fact that the local orthogonality among the consecutive residuals and direction vectors is preserved also during finite precision CG computations. Since the A-norm of the error is nonincreasing in finite precision arithmetic (recall results by Greenbaum [14]), the sum in (2.24) can be used for the estimation of the A-norm of the actual error. A similar consideration can be done also for preconditioned CG [44] where the condition number of the preconditioner plays an important role.

The results published in [43, 44] are encouraging for investigating the numerical stability of further estimates of the A-norm of the error in CG, based on Gauss-Radau (or Gauss-Lobatto) quadrature. Our experiments performed jointly with Gérard Meurant predict that estimates based on Gauss-Radau quadrature can be significantly influenced by the behavior of CG in finite precision arithmetic, and need not always provide a good estimate. We would like to analyze and understand this phenomenon.

Finally, let us discuss the finite precision behavior of the approximation to the quantity $c^* A^{-1} b$ based on the identity (2.23); see [45]. For BiCG one can hardly expect results of the same strength as for CG. Formally, we can derive analogous formulas as in the rounding error analysis of (2.24). In [45] we have shown that the identity (2.23) holds also for numerically computed quantities (up to some small inaccuracy), if the local biorthogonality among consecutive residual and direction vectors is well preserved. But in BiCG, a close preservation of the local biorthogonality cannot be proved due to the possible occurrence of the so-called breakdowns. Note that breakdowns are not caused by rounding errors; they can occur in exact arithmetic. Nevertheless, there is no need of preserving the global orthogonality among the computed vectors in order the approximation based on (2.23) to work during finite precision computations. This represents a strong numerical argument in favor of the proposed estimate.

2.6 Short summary of our contribution

Convergence results. In the survey [28] we summarize known convergence results for three well-known Krylov subspace methods (CG, MINRES and GMRES) and formulate open questions in this area.

In [29, 30, 32] we investigate the convergence of Krylov subspace methods for systems of linear algebraic equations with normal matrices. In particular, in [29] we explore the standard bound based on the min-max approximation problem (2.2). We ask the questions how to evaluate it and whether it is possible to characterize the initial residual for which this bound is attained. In [30] we apply the previous results [29] to a particular problem, to one-dimensional reaction-diffusion equations with Dirichlet boundary conditions, and discuss the question which source term and boundary condition in the underlying differential equation lead to the slowest possible convergence of a Krylov subspace method. In [32] we revise the question of the sharpness of the bounds (2.6) and (2.8). We formulate this problem in a more general setting (2.11) and prove the sharpness result using classical results of approximation theory.

Our work [47, 31, 7, 8] is concerned with convergence results for nonnormal matrices. Since in the nonnormal case the situation is much less clear, we concentrate on a case study in [47]. We study the question about the quality of the bound (2.15) based on polynomial numerical hull, or the question about the sharpness of the inequality (2.14) for a single Jordan block J_λ. In [8] we contribute to the understanding of the worst-case GMRES approximation problem (2.13). For example, we show that the worst-case GMRES polynomial may not be uniquely determined. In [31] we generalize the ideal GMRES problem (2.14) to (2.16) and prove the uniqueness of the minimizer. Finally, in [7] we investigate so called Chebyshev polynomials of a square matrix A (solutions of the problem (2.16) for $f(x) = 1$ and $m = 1$) that are related to the Arnoldi method for approximating eigenvalues of matrices. We study general properties of these polynomials, which in some cases turn out to be generalizations of well-known properties of Chebyshev polynomials of compact sets in the complex plane.

Short recurrences. In the paper [5] we formulate the Faber-Manteuffel theorem in terms of linear operators on finite dimensional Hilbert spaces, and give two new different proofs of the necessity part. In the subsequent paper [6] we give a new proof of necessary and sufficient conditions on A so that the orthogonal reduction yields a Hessenberg matrix with small bandwidth.

Error estimation. In [43] we summarize the connection between CG and Gauss quadrature and analyze various ways how to compute the Gauss quadrature lower bound. We recommend the estimate (2.21) that is simple and numerically stable. In [44] we extend our results [43] to practical users of the PCG method. In [35] we formulate a new algorithm called CGQ for computing quadrature bounds for the A-norm of the error in CG. This new algorithm represents a considerable simplification of the CGQL algorithm by Golub and Meurant. Finally, in [45] we presents an approach for approximating $c^*A^{-1}b$ that is designed to be computationally efficient. Algorithmically, we extend the results presented in [43, 44] from the Hermitian and positive definite case and CG to the general complex case and BiCG.

Finite precision arithmetic. To further understand the mathematical model that describes results of CG computation we investigated in [37] the problem of sensitivity of Gauss-Christoffel quadrature with respect to small perturbations of the distribution function. In [43, 44] we present the rounding error analysis of the error estimators in CG and PCG. We explain, why the proposed error estimator works also in finite precision arithmetic, despite the fact that the computed approximate solutions are usually far away from its exact precision counterparts. In [45] we have shown that the proposed approximation of the bilinear form $c^*A^{-1}b$ works well also in finite precision arithmetic, if the local biorthogonality among consecutive BiCG residual and direction vectors is well preserved.

Bibliography

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

[2] M. Bellalij, Y. Saad, and H. Sadok, *Analysis of some Krylov subspace methods for normal matrices via approximation theory and convex optimization*, Electron. Trans. Numer. Anal., 33 (2008/09), pp. 17–30.

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

[4] V. Faber, W. Joubert, E. Knill, and T. Manteuffel, *Minimal residual method stronger than polynomial preconditioning*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 707–729.

[5] V. Faber, J. Liesen, and P. Tichý, *The Faber-Manteuffel theorem for linear operators*, SIAM J. Numer. Anal., 46 (2008), pp. 1323–1337.

[6] ——, *On orthogonal reduction to Hessenberg form with small bandwidth*, Numer. Algorithms, 51 (2009), pp. 133–142.

[7] V. Faber, J. Liesen, and P. Tichy, *On chebyshev polynomials of matrices*, SIAM Journal on Matrix Analysis and Applications, 31 (2010), pp. 2205–2221.

[8] V. Faber, J. Liesen, and P. Tichý, *Properties of worst-case GMRES*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 1500–1519.

[9] V. Faber and T. Manteuffel, *Necessary and sufficient conditions for the existence of a conjugate gradient method*, SIAM J. Numer. Anal., 21 (1984), pp. 352–362.

[10] A. Frommer and V. Simoncini, *Stopping criteria for rational matrix functions of Hermitian and, symmetric matrices*, SIAM J. Sci. Comput., 30 (2008), pp. 1387–1412.

[11] G. H. Golub and G. Meurant, *Matrices, moments and quadrature*, in Numerical analysis 1993 (Dundee, 1993), vol. 303 of Pitman Res. Notes Math. Ser., Longman Sci. Tech., Harlow, 1994, pp. 105–156.

[12] G. H. Golub and G. Meurant, *Matrices, moments and quadrature with applications*, Princeton University Press, USA, 2010.

[13] A. Greenbaum, *Comparison of splittings used with the conjugate gradient algorithm*, Numer. Math., 33 (1979), pp. 181–193.

[14] ——, *Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences*, Linear Algebra Appl., 113 (1989), pp. 7–63.

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

[16] ——, *Some theoretical results derived from polynomial numerical hulls of Jordan blocks*, Electron. Trans. Numer. Anal., 18 (2004), pp. 81–90.

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

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

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

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

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

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

[23] N. J. Higham, *Functions of Matrices. Theory and Computation*, SIAM, Philadelphia, PA, 2008.

[24] P. Jiránek, Z. Strakoš, and M. Vohralík, *A posteriori error estimates including algebraic error and stopping criteria for iterative solvers*, SIAM J. Sci. Comput., 32 (2010), pp. 1567–1590.

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

[26] J. Liesen and Z. Strakoš, *On optimal short recurrences for generating orthogonal Krylov subspace bases*, SIAM Rev., 50 (2008), pp. 485–503.

[27] ——, *Krylov subspace methods: Principles and Analysis*, Oxford University Press, Oxford, 2013.

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

[29] ——, *The worst-case GMRES for normal matrices*, BIT, 44 (2004), pp. 79–98.

[30] ——, *On the worst-case convergence of MR and CG for symmetric positive definite tridiagonal Toeplitz matrices*, Electron. Trans. Numer. Anal., 20 (2005), pp. 180–197.

[31] ——, *On best approximations of polynomials in matrices in the matrix 2-norm*, SIAM J. Matrix Anal. Appl., 31 (2009), pp. 853–863.

[32] ——, *Max-min and min-max approximation problems for normal matrices revisited*, Electron. Trans. Numer. Anal., 41 (2014), pp. 159–166.

[33] G. G. Lorentz, *Approximation of Functions*, Chelsea Publishing Co., New York, second ed., 1986.

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

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

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

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

[38] C. C. Paige, *The Computation of Eigenvalues and Eigenvectors of Very Large Sparse Matrices*, PhD thesis, Intitute of Computer Science, University of London, London, U.K., 1971.

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

[40] T. J. Rivlin and H. S. Shapiro, *A unified approach to certain problems of approximation and minimization*, J. Soc. Indust. Appl. Math., 9 (1961), pp. 670–699.

[41] Y. Saad and M. H. Schultz, *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[42] Z. Strakoš, *Model reduction using the Vorobyev moment problem*, Numer. Algorithms, 51 (2009), pp. 363–379.

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

[44] ——, *Error estimation in preconditioned conjugate gradients*, BIT, 45 (2005), pp. 789–817.

[45] ——, *On efficient numerical approximation of the bilinear form $c^*A^{-1}b$*, SIAM J. Sci. Comput., 33 (2011), pp. 565–587.

[46] P. Tichý, *O některých otevřených problémech v Krylovovských metodách*, PhD thesis, Faculty of Mathematics and Physics, Charles University, 2002.

[47] P. Tichý, J. Liesen, and V. Faber, *On worst-case GMRES, ideal GMRES, and the polynomial numerical, hull of a Jordan block*, Electron. Trans. Numer. Anal., 26 (2007), pp. 453–473.

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

[49] K.-C. Toh and L. N. Trefethen, *The Chebyshev polynomials of a matrix*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 400–419 (electronic).

[50] Y. V. Vorobyev, *Methods of moments in applied mathematics*, Gordon and Breach Science Publishers, New York, 1965.

SIAM J. NUMER. ANAL.
Vol. 46, No. 3, pp. 1323–1337

THE FABER–MANTEUFFEL THEOREM FOR LINEAR OPERATORS[*]

V. FABER[†], J. LIESEN[‡], AND P. TICHÝ[§]

Abstract. A short recurrence for orthogonalizing Krylov subspace bases for a matrix A exists if and only if the adjoint of A is a low-degree polynomial in A (i.e., A is normal of low degree). In the area of iterative methods, this result is known as the Faber–Manteuffel theorem [V. Faber and T. Manteuffel, *SIAM J. Numer. Anal.*, 21 (1984), pp. 352–362]. Motivated by the description by J. Liesen and Z. Strakoš, we formulate here this theorem in terms of linear operators on finite dimensional Hilbert spaces and give two new proofs of the necessity part. We have chosen the linear operator rather than the matrix formulation because we found that a matrix-free proof is less technical. Of course, the linear operator result contains the Faber–Manteuffel theorem for matrices.

Key words. cyclic subspaces, Krylov subspaces, orthogonal bases, orthogonalization, short recurrences, normal matrices

AMS subject classifications. 65F10, 65F25

DOI. 10.1137/060678087

1. Introduction.

At the Householder Symposium VIII held in Oxford in July 1981, Golub posed as an open question to characterize necessary and sufficient conditions on a matrix A for the existence of a three-term conjugate gradient–type method for solving linear systems with A (cf. *SIGNUM Newsletter*, vol. 16, no. 4, 1981). This important question was answered by Faber and Manteuffel in 1984 [4]. They showed that an $(s+2)$-term conjugate gradient type method for A, based on some given inner product, exists if and only if the adjoint of A with respect to the inner product is a polynomial of degree s in A (i.e., A is normal of degree s). In the area of iterative methods this result is known as the Faber–Manteuffel theorem; see, e.g., [7, Chapter 6] or [13, Chapter 6.10].

The theory of [4] and some further developments have recently been surveyed in [12]. There the Faber–Manteuffel theorem is formulated independently of the conjugate gradient context and solely as a result on the existence of a short recurrence for generating orthogonal bases for Krylov subspaces of the matrix A. A new proof of the sufficiency part is given, and the normality condition on A is thoroughly characterized. For the proof of the (significantly more difficult) necessity part, however, the authors refer to [4]. In particular, they suggest that, in light of the fundamental nature of the result, it is desirable to find an alternative, and possibly simpler, proof. Note that a proof similar to the one of Faber and Manteuffel but for other classes of matrices has been given in [14].

Motivated by the description in [12], we here take a new approach to formulate and prove the necessity part of the Faber–Manteuffel theorem. Instead of a matrix

[*] Received by the editors December 20, 2006; accepted for publication (in revised form) September 10, 2007; published electronically March 26, 2008.
http://www.siam.org/journals/sinum/46-3/67808.html

[†] BD Biosciences-Bioimaging Systems (vance_faber@bd.com).

[‡] Institute of Mathematics, Technical University of Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany (liesen@math.tu-berlin.de). The work of this author was supported by the Emmy Noether-Programm of the Deutsche Forschungsgemeinschaft.

[§] Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 2, 18207 Prague, Czech Republic (tichy@cs.cas.cz). The work of this author was supported by the Emmy Noether-Programm of the Deutsche Forschungsgemeinschaft and by the Czech National Program of Research Information Society under project 1ET400300415.

we consider a given linear operator A on a finite dimensional Hilbert space V. By the cyclic decomposition theorem, the space V decomposes into cyclic invariant subspaces, i.e., Krylov subspaces, of A (see section 2 for details). The Faber–Manteuffel theorem then gives a necessary (and sufficient) condition on A, so that the standard Gram–Schmidt algorithm for generating orthogonal bases of the cyclic subspaces reduces from a full to a short recurrence.

We have chosen this setting because we believe that the proof of necessity is easier to follow when we use linear operators rather than matrices. In this paper we give two different proofs of the necessity part, both based on restriction of the linear operator A to certain cyclic invariant subspaces. The resulting technicalities in the matrix formulation would obstruct rather than help the understanding. Moreover, our formulation may serve as a starting point for extending the results to infinite dimensional spaces. We are not aware that any such extensions have been obtained yet.

The paper is organized as follows. In section 2 we introduce the notation and the required background from the theory of linear operators. In section 3 we translate the matrix concepts introduced in [12] into the language of linear operators. In section 4 we state and prove several technical lemmas that are required in the proof of the main result, which is given in section 5. In section 6 we give an alternative proof, which we consider elementary and constructive. This proof involves structure-preserving orthogonal transformations of Hessenberg matrices, which may be of interest beyond our context here. In section 7 we discuss our rather theoretical analysis in the preceeding sections. This discussion includes a matrix formulation of the Faber–Manteuffel theorem, a high-level description of the strategies of our two proofs of the necessity part, and our reasoning why necessity is more difficult to prove than sufficiency. For obtaining a more detailed overview of the results in this paper, section 7 may be read before the other sections.

2. Notation and background.

In this section we introduce the notation and recall some basic results from the theory of linear operators; see Gantmacher's book [6, Chapters VII and IX] for more details.

Let V be a finite dimensional Hilbert space, i.e., a complex vector space equipped with a (fixed) inner product $(\cdot,\cdot)$. Let $A : V \to V$ be a given invertible linear operator. For any vector $v \in V$, we can form the sequence

$$v, Av, A^2v, \dots . \tag{2.1}$$

Since V is finite dimensional, there exists an integer $d = d(A, v)$ such that the vectors $v, Av, \dots, A^{d-1}v$ are linearly independent, while $A^d v$ is a linear combination of them. This means that there exist scalars, $\alpha_1, \dots, \alpha_{d-1}$, not all equal to zero, such that

$$A^d v \;=\; -\sum_{j=0}^{d-1} \alpha_j A^j v\,. \tag{2.2}$$

Defining the monic polynomial $\phi(z) = z^d + \alpha_{d-1}z^{d-1} + \cdots + \alpha_0$, we can rewrite (2.2) as

$$\phi(A)v \;=\; 0\,. \tag{2.3}$$

We say that ϕ *annihilates* v. It would be more accurate to say "ϕ annihilates v with respect to A," but when it is clear which operator A is meant, the reference to A is omitted for the sake of brevity. The monic polynomial ϕ is the uniquely determined monic polynomial of smallest degree that annihilates v, and it is called the *minimal polynomial of* v. Its degree, equal to $d(A,v)$, is called the *grade of* v, and v is said to be of grade $d(A,v)$.

Consider any basis of V, and define the polynomial Φ as the least common multiple of the minimal polynomials of the basis vectors. Then Φ is the uniquely defined (independent of the choice of the basis!) monic polynomial of smallest degree that annihilates all vectors $v \in V$, and it is called the *minimal polynomial of* A. We denote its degree by $d_{\min}(A)$. Apparently, $d_{\min}(A) \geq d(A,v)$ for all $v \in V$, and Φ is divisible by the minimal polynomial of every vector $v \in V$.

If $v \in V$ is any vector of grade d, then

$$\text{span}\{v, \dots A^{d-1}v\} \;\equiv\; \mathcal{K}_d(A,v) \tag{2.4}$$

is a d-dimensional invariant subspace of A. Because of (2.2) and the special character of the basis vectors, the subspace $\mathcal{K}_d(A,v)$ is called *cyclic.* The letter $\mathcal{K}$ has been chosen because this space is often called the *Krylov subspace* of A and v. The vector v is called the *generator* of this subspace.

A central result in the theory of linear operators on finite dimensional vector spaces is that the space V can be decomposed into cyclic subspaces. This result has several equivalent formulations, and in this paper we will use the one from [6, Chapter VII, section 4, Theorem 3]: there exist vectors $w_1, \dots, w_j \in V$ of respective grades $d_1, \dots, d_j$ such that

$$V \;=\; \mathcal{K}_{d_1}(A,w_1) \oplus \cdots \oplus \mathcal{K}_{d_j}(A,w_j)\,, \tag{2.5}$$

where the minimal polynomial of w_1 is equal to the minimal polynomial of A, and for $k = 1, \dots, j-1$, the minimal polynomial of w_k is divisible by the minimal polynomial of w_{k+1}.

Since the decomposition (2.5) is an important tool in this paper, we illustrate it by a simple example (adapted from [9, section 7.2]; also see [10] for a short and self-contained proof of the decomposition (2.5)). Suppose that A is the linear operator on $V = \mathbb{R}^3$ whose matrix representation in the canonical basis of $\mathbb{R}^3$ is

$$\begin{bmatrix} 2 & -3 & -3 \\ -3 & 2 & 3 \\ 3 & -3 & -4 \end{bmatrix}.$$

The characteristic polynomial of A is $(z-2)(z+1)^2$, while the minimal polynomial is $\Phi = (z-2)(z+1)$, so that $d_{\min}(A) = 2$. Any nonzero vector in $\mathbb{R}^3$ is either of grade one (and hence is an eigenvector) or of grade two. It is easy to see that the first canonical basis vector is not an eigenvector. Thus, $w_1 \equiv [1,\,0,\,0]^T$ is of grade $d_1 = 2$, i.e., $\mathcal{K}_{d_1}(A,w_1)$ has dimension two, and the minimal polynomial of w_1 is Φ. Note that

$$\mathcal{K}_{d_1}(A,w_1) \;=\; \text{span}\left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ -3 \\ 3 \end{bmatrix} \right\} \;=\; \left\{ \begin{bmatrix} \alpha \\ \beta \\ -\beta \end{bmatrix} \;:\; \alpha, \beta \in \mathbb{R} \right\}.$$

Since $V = \mathbb{R}^3$ has dimension three, it remains to find a vector $w_2 \notin \mathcal{K}_{d_1}(A,w_1)$ that is of grade one and has minimal polynomial $z+1$, i.e., w_2 is an eigenvector with respect to the eigenvalue -1, that is not contained in $\mathcal{K}_{d_1}(A,w_1)$. These requirements are satisfied by $w_2 \equiv [1,\,0,\,1]^T$, giving

$$\mathbb{R}^3 \;=\; \mathcal{K}_{d_1}(A,w_1) \,\oplus\, \mathcal{K}_{d_2}(A,w_2) \;=\; \text{span}\,\{w_1,\, Aw_1\} \,\oplus\, \text{span}\,\{w_2\}\,.$$

In the basis of $\mathbb{R}^3$ given by w_1, Aw_1, w_2, the linear operator A has the matrix representation

$$\left[\begin{array}{cc|c} 0 & 2 & \\ 1 & 1 & \\ \hline & & -1 \end{array}\right].$$

Here the two diagonal blocks correspond to the decomposition (2.5), where each block is the companion matrix of the minimal polynomial of the respective cyclic subspace generators. This matrix representation is sometimes called the *rational canonical form*. When this canonical form consists of a single diagonal block in companion form, A is called *nonderogatory*. Hence in our example A is derogatory, but the restriction of A to the cyclic subspace generated by w_1 is nonderogatory. Loosely speaking, this restriction is the "largest nonderogatory part" of A.

3. Orthogonalization of a cyclic subspace basis. Let $v \in V$ be a vector of grade d. For theoretical as well as practical purposes it is often convenient to *orthogonalize* the basis $v, \dots, A^{d-1}v$ of the cyclic subspace $\mathcal{K}_d(A, v)$. The classical approach to orthogonalization, which appears in different mathematical areas (see, e.g., [2, p. 15], [5, p. 74]) is the Gram–Schmidt algorithm:

$$v_1 = v\,, \tag{3.1}$$

$$v_{n+1} = Av_n - \sum_{m=1}^{n} h_{m,n} v_m\,, \tag{3.2}$$

$$h_{m,n} = \frac{(Av_n, v_m)}{(v_m, v_m)}\,, \quad m = 1, \dots, n\,, \quad n = 1, \dots, d-1\,. \tag{3.3}$$

The resulting vectors $v_1, \dots, v_d$ are mutually orthogonal, and for $n = 1, \dots, d$ they satisfy $\text{span}\{v_1, \dots, v_n\} = \text{span}\{v, \dots, A^{n-1}v\}$. We call v (or v_1) the *initial vector* of the algorithm (3.1)–(3.3). When A is a (square) matrix, this algorithm is usually referred to as Arnoldi's method [1]. It can be equivalently written as

$$v_1 = v\,, \tag{3.4}$$

$$A \underbrace{[v_1, \dots, v_{d-1}]}_{\equiv V_{d-1}} = \underbrace{[v_1, \dots, v_d]}_{\equiv V_d} \underbrace{\begin{bmatrix} h_{1,1} & \cdots & h_{1,d-1} \\ 1 & \ddots & \vdots \\ & \ddots & h_{d-1,d-1} \\ & & 1 \end{bmatrix}}_{\equiv H_{d,d-1}}\,, \tag{3.5}$$

$$(v_i, v_j) = 0 \quad \text{for} \quad i \neq j\,, \quad i, j = 1, \dots, d\,. \tag{3.6}$$

The matrix $H_{d,d-1}$ is an unreduced upper Hessenberg matrix of size $d \times (d-1)$. Its band structure determines the length of the recurrence (3.2) that generates the orthogonal basis. To state this formally, we need the following definition [12, Definition 2.1].

DEFINITION 3.1. *An unreduced upper Hessenberg matrix is called $(s+2)$-band Hessenberg when its sth superdiagonal contains at least one nonzero entry and all its entries above its sth superdiagonal are zero.*

If $H_{d,d-1}$ is $(s+2)$-band Hessenberg, then for $n = 1, \ldots, d-1$, the recurrence (3.2) reduces to

$$v_{n+1} = Av_n - \sum_{m=n-s}^{n} h_{m,n} v_m \,, \tag{3.7}$$

and thus the orthogonal basis is generated by an $(s+2)$-*term recurrence.* Since precisely the latest $s+1$ basis vectors $v_n, \ldots, v_{n-s}$ are required to determine v_{n+1}, and only one operation with A is performed, an $(s+2)$-term recurrence of the form (3.7) is called *optimal.*

DEFINITION 3.2 (linear operator version of [12, Definition 2.4]). *Let A be an invertible linear operator with minimal polynomial degree $d_{\min}(A)$ on a finite dimensional Hilbert space, and let s be a nonnegative integer, $s+2 \leq d_{\min}(A)$.*

(1) *If for an initial vector the matrix $H_{d,d-1}$ in* (3.4)–(3.6) *is $(s+2)$-band Hessenberg, then we say that A admits for the given initial vector an optimal $(s+2)$-term recurrence.*

(2) *If A admits for any given initial vector an optimal recurrence of length at most $s+2$, while it admits for at least one given initial vector an optimal $(s+2)$-term recurrence, then we say that A admits an optimal $(s+2)$-term recurrence.*

We denote the *adjoint of* A by A^*. This is the uniquely determined operator that satisfies $(Av, w) = (v, A^*w)$ for all vectors v and w in the given Hilbert space. The operator A is called *normal* if it commutes with its adjoint, $AA^* = A^*A$. This holds if and only if A has a complete orthonormal system of eigenvectors. Equivalently, A^* can be written as a polynomial in A, $A^* = p(A)$, where p is completely determined by the condition that $p(\lambda_j) = \overline{\lambda}_j$ for all eigenvalues λ_j of A (cf. [6, Chapter IX, section 10]). We will be particularly interested in the degree of this polynomial.

DEFINITION 3.3. *Let A be an invertible linear operator on a finite dimensional Hilbert space. If the adjoint of A satisfies $A^* = p(A)$, where p is a polynomial of smallest degree s having this property, then A is called normal of degree s, or, shortly, normal(s).*

The condition that A is normal(s) is *sufficient* for A to admit an optimal $(s+2)$-term recurrence. The precise formulation of this statement is the following.

THEOREM 3.4. *Let A be an invertible linear operator with minimal polynomial degree $d_{\min}(A)$ on a finite dimensional Hilbert space. Let s be a nonnegative integer, $s+2 < d_{\min}(A)$. If A is normal(s), then A admits an optimal $(s+2)$-term recurrence.*

Proof. A matrix version of this result is given in [12, Theorem 2.9], and the proof given there can be easily adapted from matrices to linear operators. □

The main result we will prove in this paper is that the condition that A is normal(s) also is *necessary.*

THEOREM 3.5. *Let A be an invertible linear operator with minimal polynomial degree $d_{\min}(A)$ on a finite dimensional Hilbert space. Let s be a nonnegative integer, $s+2 < d_{\min}(A)$. If A admits an optimal $(s+2)$-term recurrence, then A is normal(s).*

4. Technical lemmas.

We prove Theorem 3.5 in section 5. To do so, we need several technical lemmas that are stated and proved in this section.

LEMMA 4.1. *Let A be an invertible linear operator with minimal polynomial degree $d_{\min}(A)$ on a finite dimensional Hilbert space. If $1 < i < n \leq d_{\min}(A)$ and $(u_1, Au_i) = 0$ for every initial vector u_1 of grade n, then $(v_1, Av_i) = 0$ for every initial vector v_1 of grade m, where $i \leq m \leq n$.*

(Here u_i, v_i are the ith basis vectors generated by (3.1)–(3.3) *with initial vectors u_1, v_1, respectively.)*

Proof. If $m = n$, there is nothing to prove. Hence, suppose that $m < n$, and let v_1 be a vector of grade m, and u_1 be a vector of grade n, such that the minimal polynomial of v_1 divides the minimal polynomial of u_1. Define

$$x_1 \;\equiv\; x_1(\gamma) \;\equiv\; v_1 + \gamma u_1\,, \tag{4.1}$$

where γ is a scalar parameter. It is clear that, except for finitely many choices of γ, the vector x_1 is of grade n.

Suppose that γ has been chosen so that x_1 is of grade n, and consider the corresponding ith basis vector x_i, where $1 < i \leq m$. By construction, $x_i = p(A)x_1$, where p is a polynomial of (exact) degree $i-1$. The vector x_i is defined uniquely (up to scaling) by the conditions

$$(A^j x_1, x_i) \;=\; (A^j x_1, p(A)x_1) \;=\; 0\,, \quad j = 0, \dots, i-2\,.$$

The hypothesis

$$(x_1, Ax_i) = (x_1, Ap(A)x_1) \;=\; 0$$

gives one additional condition. We thus have i conditions that translate into i homogeneous linear equations for the i coefficients of the polynomial p. The existence of x_i implies that the determinant of the matrix $M(x_1)$ of these equations must be zero, where

$$M(x_1) \;=\; \begin{bmatrix} (x_1, x_1) & (x_1, Ax_1) & \cdots & (x_1, A^{i-1}x_1) \\ (Ax_1, x_1) & (Ax_1, Ax_1) & \cdots & (Ax_1, A^{i-1}x_1) \\ \vdots & \vdots & \vdots & \vdots \\ (A^{i-2}x_1, x_1) & (A^{i-2}x_1, Ax_1) & \cdots & (A^{i-2}x_1, A^{i-1}x_1) \\ (x_1, Ax_1) & (x_1, A^2x_1) & \cdots & (x_1, A^i x_1) \end{bmatrix}.$$

Now note that $\det M(x_1)$ is a continuous function of γ. By construction, this function is zero for all but a finite number of choices of γ. Therefore $\det M(x_1) = 0$ for all γ, and in particular, $\det M(v_1) = 0$. Consequently, there exists a nontrivial solution of the linear system with $M(v_1)$, defining a vector $w = p(A)v_1$, where p is a polynomial of degree at most $i-1$. The first $i-1$ rows mean that w is orthogonal to $v_1, \dots, A^{i-2}v_1$, so w is a multiple of v_i. The last row means that Aw and hence Av_i is orthogonal to v_1. □

The decomposition (2.5) shows that for any linear operator A on a finite dimensional Hilbert space V, there exists a vector in V whose minimal polynomial coincides with the minimal polynomial of A. The following result shows that there in fact exists a basis of V consisting of vectors with this property.

LEMMA 4.2. *Let A be an invertible linear operator with minimal polynomial degree $d_{\min}(A)$ on a finite dimensional Hilbert space V. Then there exists a basis of V consisting of vectors that all are of grade $d_{\min}(A)$.*

Proof. From the cyclic decomposition theorem, cf. (2.5), we know that there exist vectors $w_1, \dots, w_j$ with minimal polynomials $\phi_1, \dots, \phi_j$ of respective degrees $d_1, \dots, d_j$, such that the space V can be decomposed as

$$V \;=\; \mathcal{K}_{d_1}(A, w_1) \oplus \cdots \oplus \mathcal{K}_{d_j}(A, w_j)\,,$$

where ϕ_1 equals the minimal polynomial of A, and ϕ_k is divisible by ϕ_{k+1} for $k = 1, \dots, j-1$. In particular, $d_1 = d_{\min}(A)$. Consequently, a basis of V is given by

$$w_1, \dots, A^{d_1-1}w_1, \quad w_2, \dots, A^{d_2-1}w_2, \quad \dots, \quad w_j, \dots, A^{d_j-1}w_j\,.$$

But then it is easy to see that

$$w_1, \dots, A^{d_1-1}w_1, \quad w_2+w_1, \dots, A^{d_2-1}w_2+w_1, \quad \dots, \quad w_j+w_1, \dots, A^{d_j-1}w_j+w_1$$

is a basis of V consisting of vectors that all are of grade d_1. □

The following result is a generalization of [11, Lemma 4.1], which in turn can be considered a (considerably) strengthened version of [4, Lemma 2].

LEMMA 4.3. *Let A be an invertible linear operator with minimal polynomial degree $d_{\min}(A)$ on a finite dimensional Hilbert space. Let B be a linear operator on the same space, and let s be a nonnegative integer, $s+2 \le d_{\min}(A)$. If*

$$Bv \in \operatorname{span}\{v, \dots, A^s v\} \quad \textit{for all vectors } v \textit{ of grade } d_{\min}(A)\,, \tag{4.2}$$

then $AB = BA$. In particular, if $B = A^*$, then A is normal(t) for some $t \le s$.

Proof. For notational convenience, we denote $\delta = d_{\min}(A)$. Let v be any vector of grade δ. Since A is invertible, $\mathcal{K}_\delta(A, v) = \mathcal{K}_\delta(A, Av)$. In addition, except possibly when γ is an eigenvalue of A, the vector $w = (A - \gamma I)v$ satisfies $\mathcal{K}_\delta(A, w) = \mathcal{K}_\delta(A, v)$. In the following, we exclude those values of γ. By assumption, there exist polynomials p_γ, q, and r of degree at most s, which satisfy

$$Bw = p_\gamma(A)w, \quad B(Av) = q(A)(Av), \quad Bv = r(A)v,$$

where p_γ depends on γ, but q and r do not. We can then write Bw as

$$Bw \;=\; p_\gamma(A)w \;=\; p_\gamma(A)Av - \gamma p_\gamma(A)v$$

and

$$Bw \;=\; BAv - \gamma Bv \;=\; q(A)Av - \gamma r(A)v\,.$$

Combining these two identities yields

$$t_\gamma(A)v \;=\; 0\,, \quad \text{where} \quad t_\gamma(z) = z(p_\gamma(z) - q(z)) - \gamma(p_\gamma(z) - r(z))\,.$$

The polynomial t_γ is of degree at most $s+1 < s+2 \le \delta$. Thus, except for finitely many γ, $t_\gamma = 0$. Some straightforward algebraic manipulation gives, for all but these γ,

$$\gamma(q(z) - r(z)) \;=\; (z - \gamma)\widehat{p}_\gamma(z)\,,$$

where $\widehat{p}_\gamma \equiv p_\gamma - q$ is of degree at most $s-1$. Therefore, every γ that is not an eigenvalue of A is a root of the polynomial $r - q$, which consequently must be identically zero.

But then

$$B(Av) \;=\; q(A)(Av) \;=\; Aq(A)v \;=\; Ar(A)v \;=\; A(Bv)\,.$$

By Lemma 4.2, there exists a basis consisting of vectors of grade δ. Hence $BAv = ABv$ for a basis of vectors v, so that $BA = AB$.

Finally, if $B = A^*$, then $AA^* = A^*A$, so that A is normal and hence $A^* = p(A)$ for some polynomial. From (4.2) we see that the degree of p is at most s, so that A is normal(t) for some $t \le s$. □

5. Proof of Theorem 3.5. Let A be an invertible linear operator on a finite dimensional Hilbert space, and let s be a nonnegative integer, $s+2 < d_{\min}(A)$. Suppose that A admits an optimal $(s+2)$-term recurrence.

Step 1. *Restriction to a cyclic subspace of dimension* $s+2$.

If u_1 is any vector of grade $s+3$, then (with the obvious meaning of u_{s+2})

$$0 \;=\; h_{1,s+2} \;=\; (u_1, Au_{s+2})\,. \tag{5.1}$$

Consider any v_1 of grade $s+2$, and the corresponding cyclic subspace $\mathcal{K}_{s+2}(A, v_1)$. Let $\widehat{A}$ be the restriction of A to $\mathcal{K}_{s+2}(A, v_1)$, i.e., the invertible linear operator

$$\widehat{A} : \mathcal{K}_{s+2}(A, v_1) \to \mathcal{K}_{s+2}(A, v_1)\,, \qquad v \mapsto Av \;\text{ for } v \in \mathcal{K}_{s+2}(A, v_1)\,.$$

Clearly, $d_{\min}(\widehat{A}) = s+2$. Let $\mathcal{K}_{s+2}(A, v_1)$ be equipped with the same inner product as the whole space.

Let $y_1 \in \mathcal{K}_{s+2}(A, v_1)$ be any vector of grade $s+2$. Obviously, the grade of y_1 with respect to A is the same as the grade of y_1 with respect to $\widehat{A}$. Since (5.1) holds for any u_1 of grade $s+3$ (with respect to A), Lemma 4.1 (with $i = m = s+2$ and $n = s+3$) implies that (with the obvious meaning of y_{s+2})

$$0 \;=\; (y_1, Ay_{s+2}) \;=\; (y_1, \widehat{A}y_{s+2}) \;=\; (\widehat{A}^* y_1, y_{s+2})\,, \tag{5.2}$$

where $\widehat{A}^* : \mathcal{K}_{s+2}(A, v_1) \to \mathcal{K}_{s+2}(A, v_1)$ is the adjoint operator of $\widehat{A}$. But this means that

$$\widehat{A}^* y_1 \;\in\; \operatorname{span}\{y_1, \ldots, \widehat{A}^s y_1\}\,. \tag{5.3}$$

Since this holds for any vector $y_1 \in \mathcal{K}_{s+2}(A, v_1) = \mathcal{K}_{s+2}(\widehat{A}, v_1)$ of grade $s+2 = d_{\min}(\widehat{A})$, Lemma 4.3 implies that $\widehat{A}$ is normal(t) for some $t \le s$. In particular, $\widehat{A}$ is normal, and has $s+2$ distinct eigenvalues, λ_k, $k = 1, \ldots, s+2$, with corresponding eigenvectors that are mutually orthogonal. Moreover, there exists a polynomial of degree at most s such that $p(\lambda_k) = \overline{\lambda}_k$, $k = 1, \ldots, s+2$. By definition, any eigenpair of $\widehat{A}$ is an eigenpair of A. Therefore, A acting on *any* vector of grade $s+2$ has $s+2$ distinct eigenvalues, and the corresponding eigenvectors are mutually orthogonal in the given inner product.

Step 2. *Extension to the whole space.*

Consider the cyclic decomposition of the whole space as in (2.5). Then the cyclic subspace $\mathcal{K}_{d_1}(A, w_1)$, where w_1 has the same minimal polynomial as A, can be further decomposed into

$$\mathcal{K}_{d_1}(A, w_1) \;=\; \mathcal{K}_{c_1}(A, z_1) \oplus \cdots \oplus \mathcal{K}_{c_\ell}(A, z_\ell)\,,$$

where the minimal polynomial of z_k is $(z-\lambda_k)^{c_k}$, $k = 1, \ldots, \ell$, and $\lambda_1, \ldots, \lambda_\ell$ are the distinct eigenvalues of A (see, e.g., [6, Chapter VII, section 2, Theorem 1]). In other words, $\mathcal{K}_{d_1}(A, w_1)$ is decomposed into ℓ cyclic invariant subspaces of A, where each of these corresponds to one of the ℓ distinct eigenvalues of A. (Recall that the restriction of A to $\mathcal{K}_{d_1}(A, w_1)$ is nonderogatory; see the example at the end of section 2.) In particular, if A is diagonalizable, then $\ell = d_{\min}(A)$, and $c_1 = \cdots = c_\ell = 1$, and $z_1, \ldots, z_\ell$ are eigenvectors of A corresponding to $\lambda_1, \ldots, \lambda_\ell$, respectively. In general, we can assume that $c_1 \ge c_2 \ge \cdots \ge c_\ell$. If $c_1 \ge s+2$, we can determine a vector v_1 of grade $s+2$ in $\mathcal{K}_{c_1}(A, z_1)$. But then the above implies that A acting on v_1 has $s+2$

distinct eigenvalues, which is a contradiction. Hence $c_1 < s+2$. We therefore can find an index m so that $c_1 + \cdots + c_{m-1} + \tilde{c}_m = s+2$, $0 \le \tilde{c}_m \le c_m$. Let $\tilde{z}_m$ be any vector of grade $\tilde{c}_m$ in $\mathcal{K}_{c_m}(A, z_m)$; then $w = z_1 + \cdots + z_{m-1} + \tilde{z}_m$ is of grade $s+2$. Hence A acting on w has $s+2$ distinct eigenvalues, which shows that $c_1 = c_2 = \cdots = c_\ell = 1$. To these eigenvalues correspond $s+2$ eigenvectors that are mutually orthogonal in the given inner product.

In the cyclic decomposition (2.5), the minimal polynomial of w_k is divisible by the minimal polynomial of w_{k+1}. Therefore the whole space completely decomposes into one-dimensional cyclic subspaces of A, i.e., A has a complete system of eigenvectors. We know that any $s+2$ of these corresponding to distinct eigenvalues of A must be mutually orthogonal. In the subspaces corresponding to a multiple eigenvalue we can find an orthogonal basis. Therefore A has a complete orthonormal system of eigenvectors, and hence A is normal. For every subset of $s+2$ distinct eigenvalues there exists a polynomial p of degree at most s that satisfies $p(\lambda_k) = \overline{\lambda_k}$ for all eigenvalues λ_k in the subset. If we take any two subsets having $s+1$ eigenvalues in common, the two corresponding polynomials must be identical. Thus all the polynomials are identical, so that A is normal(t) for some $t \le s$.

If $t < s$, then by the sufficiency result in Theorem 3.4, A admits an optimal $(t+2)$-term recurrence, which contradicts our initial assumption. Hence $t = s$, so that A is normal(s), which concludes the proof. □

6. Another proof based on the Rotation Lemma. In this section we discuss an elementary and more constructive approach to proving Theorem 3.5, which is based on orthogonal transformations ("rotations") of upper Hessenberg matrices. With this approach, we can prove Theorem 3.5 with the assumption $s+2 < d_{\min}(A)$ replaced by $s+3 < d_{\min}(A)$. We discuss the missing case $s+3 = d_{\min}(A)$ in section 7.

As above, let A be an invertible linear operator with minimal polynomial degree $d_{\min}(A)$ on a finite dimensional Hilbert space. Let s be a given nonnegative integer, $s+3 < d_{\min}(A)$. We assume that

(6.1) A admits an $(s+2)$-term recurrence, but A is *not* normal(s),

and derive a contradiction.

For deriving the contradiction we need some notation. Suppose that the space is decomposed into cyclic invariant subspaces of A as in (2.5). Let $\widehat{A}$ be the restriction of A to $\mathcal{K}_{d_1}(A, w_1)$, i.e., the invertible linear operator defined by

$$\widehat{A} \;:\; \mathcal{K}_{d_1}(A, w_1) \to \mathcal{K}_{d_1}(A, w_1)\,, \qquad v \mapsto Av \;\text{ for } v \in \mathcal{K}_{d_1}(A, w_1)\,.$$

The operator $\widehat{A}$ depends on the choice of w_1, which we consider fixed here, so $\widehat{A}$ is fixed as well. It is clear that $d_1 = d_{\min}(A) = d_{\min}(\widehat{A})$. We denote $d = d_1$ for simplicity. Now let $v_1 \in \mathcal{K}_d(\widehat{A}, w_1)$ be any initial vector of grade d, and let $v_1, \ldots, v_d$ be the corresponding orthogonal basis of $\mathcal{K}_d(\widehat{A}, v_1) = \mathcal{K}_d(\widehat{A}, w_1)$ generated by (3.1)–(3.3). Then the matrix representation of the operator $\widehat{A}$ with respect to this particular basis is a $d \times d$ unreduced upper Hessenberg matrix H_d, which is defined by the equation

$$\widehat{A}\,[v_1, \ldots, v_d] \;=\; [v_1, \ldots, v_d]\, H_d\,. \tag{6.2}$$

The matrix formed by the first $d-1$ columns of H_d coincides with the $d \times (d-1)$ upper Hessenberg matrix generated by (3.1)–(3.3) with $\widehat{A}$ and the initial vector v_1, while the last column of H_d is given by the vector

$$h_d = \begin{bmatrix} h_{1,d} \\ \vdots \\ h_{d,d} \end{bmatrix}, \quad \text{where} \quad h_{m,d} = \frac{(\widehat{A}v_d, v_m)}{(v_m, v_m)}, \quad m = 1, \dots, d. \tag{6.3}$$

In short, $H_d = [H_{d,d-1}, h_d]$. We now proceed in two steps.

Step 1. *Show that there exists a basis for which $h_{1,d} \neq 0$.*

We first show that under assumption (6.1) there exists an initial vector $v_1 \in \mathcal{K}_d(\widehat{A}, w_1)$ of grade $d = d_{\min}(\widehat{A})$ for which the matrix representation H_d of $\widehat{A}$ has $h_{1,d} \neq 0$. Suppose not, i.e., for all $v_1 \in \mathcal{K}_d(\widehat{A}, w_1)$ of grade $d_{\min}(\widehat{A})$, we have for the resulting entry $h_{1,d}$,

$$0 = h_{1,d} = \frac{(\widehat{A}v_d, v_1)}{(v_1, v_1)} = \frac{(v_d, \widehat{A}^* v_1)}{(v_1, v_1)},$$

where $\widehat{A}^*$ is the adjoint of $\widehat{A}$. In particular, this implies that for all vectors $v_1 \in \mathcal{K}_d(\widehat{A}, w_1)$ of grade $d = d_{\min}(\widehat{A})$,

$$\widehat{A}^* v_1 \in \{v_1, \dots, \widehat{A}^{d-2} v_1\}.$$

By Lemma 4.3, $\widehat{A}$ is normal(t) for some $t \leq d_{\min}(\widehat{A}) - 2$. Therefore, A acting on any vector of grade $d_{\min}(A)$ has $d_{\min}(A)$ distinct eigenvalues and corresponding eigenvectors that are mutually orthogonal. From this it is easy to see that A is normal(t). By the sufficiency result in Theorem 3.4, A admits an optimal $(t+2)$-term recurrence. However, we have assumed in (6.1) that A admits an optimal $(s+2)$-term recurrence, so $t = s$. But then A is normal(s), which contradicts the second part of the assumption. In summary, there exists an initial vector v_1 of grade $d = d_{\min}(A)$, such that (6.2) holds with $H_d = [H_{d,d-1}, h_d]$, where $H_{d,d-1}$ is $(s+2)$-band Hessenberg (this follows from the first part of our assumption), while $h_{1,d} \neq 0$.

Step 2. *Rotation of the nonzero entry $h_{1,d}$.*

The following result is called the Rotation Lemma for reasons apparent from its proof.

LEMMA 6.1 (Rotation Lemma). *Let s, d be nonnegative integers, $s + 3 < d$. Let H_d be a $d \times d$ unreduced upper Hessenberg matrix with $h_{1,d} \neq 0$ and $H_{d,d-1}$, the matrix formed by the first $d-1$ columns of H_d, being an $(s+2)$-band Hessenberg matrix. Then there exists a unitary matrix G such that $\widetilde{H}_d \equiv G^* H_d G$ is a $d \times d$ unreduced upper Hessenberg matrix with $[\widetilde{h}_{1,d-1}, \widetilde{h}_{2,d-1}] \neq [0, 0]$.*

Proof. The main idea of this proof is to find $d-1$ (complex) Givens rotations of the form

$$G_i \equiv \begin{bmatrix} I_{d-1-i} & & & \\ & c_i & \overline{s}_i & \\ & -s_i & c_i & \\ & & & I_{i-1} \end{bmatrix}, \quad c_i^2 + |s_i|^2 = 1, \quad c_i \in \mathbb{R}, \quad i = 1, \dots, d-1, \tag{6.4}$$

which, applied symmetrically to H_d, "rotate" the nonzero entry $h_{1,d}$ to the $(d-1)$st column of the resulting matrix $\widetilde{H}_d = (G_1 \cdots G_{d-1})^* H_d (G_1 \cdots G_{d-1})$. To prove the assertion it suffices to show the following. First, $\widetilde{H}_d$ must be an unreduced upper Hessenberg matrix, and, second, at least one of its entries $\widetilde{h}_{1,d-1}, \widetilde{h}_{2,d-1}$ is nonzero. See Figure 6.1 for an illustration of this idea.

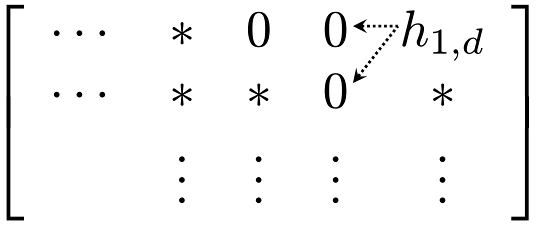


FIG. 6.1. *Graphical illustration of the Rotation Lemma. Shown is the upper-right-hand corner of $H_d = [H_{d,d-1}, h_d]$. We know that $H_{d,d-1}$ is $(s+2)$-band Hessenberg with $s+3 < d$ and that $h_{1,d} \neq 0$. We construct an orthogonal transformation G such that the matrix $\widetilde{H}_d = G^* H_d G$ remains unreduced upper Hessenberg, while the nonzero entry $h_{1,d} \neq 0$ is "rotated" to the last column of $\widetilde{H}_{d,d-1}$, so that at least one of its entries $\widetilde{h}_{1,d-1}$ and $\widetilde{h}_{2,d-1}$ is nonzero.*

Proceeding in an inductive manner, we denote $H^{(0)} \equiv H_d$. To start, choose $s_1 \in \mathbb{R} \setminus \{0\}$ and $c_1 \in \mathbb{R}$ such that $c_1^2 + s_1^2 = 1$. We have explicitly chosen real parameters s_1, c_1 since this simplifies our arguments below. These two parameters determine our first Givens rotation G_1 of the form (6.4). By construction, the matrix $H^{(1)} \equiv G_1^* H^{(0)} G_1$ is upper Hessenberg except for its entry

$$h^{(1)}_{d,d-2} = s_1 h^{(0)}_{d-1,d-2}.$$

Since $s_1 \neq 0$ and $h^{(0)}_{d-1,d-2} \neq 0$ ($H^{(0)}$ is unreduced), we have $h^{(1)}_{d,d-2} \neq 0$. The transformation by G_1 modifies only the last two rows and columns of $H^{(0)}$, so that the entries on the subdiagonal of $H^{(1)}$ satisfy $h^{(1)}_{i+1,i} = h^{(0)}_{i+1,i} \neq 0$, $i = 1, \ldots, d-3$. Next, we determine G_2 such that its application from the right to $H^{(1)}$ eliminates the nonzero entry in position $(d, d-2)$. Application of G_2^* from the left then introduces a nonzero entry in position $(d-1, d-3)$, which we will subsequently eliminate using G_3, and so forth.

In a general step $j = 2, \ldots, d-1$, suppose that $s_{j-1} \neq 0$, $h^{(j-1)}_{i+1,i} = h^{(0)}_{i+1,i} \neq 0$, $i = 1, \ldots, d-j-1$, and $h^{(j-1)}_{i+1,i} \neq 0$ for $i = d-j+2, \ldots, d-1$. Next suppose that

$$H^{(j-1)} \equiv G^*_{j-1} H^{(j-2)} G_{j-1}$$

is an upper Hessenberg matrix except for its entry

$$h^{(j-1)}_{d-j+2,d-j} = s_{j-1} h^{(0)}_{d-j+1,d-j} \neq 0.$$

The next Givens rotation G_j is (uniquely) determined to eliminate this nonzero entry, i.e., we determine c_j and s_j by the equation

$$[h^{(j-1)}_{d-j+2,d-j}, h^{(j-1)}_{d-j+2,d-j+1}] \begin{bmatrix} c_j & \overline{s}_j \\ -s_j & c_j \end{bmatrix} = [0, h^{(j)}_{d-j+2,d-j+1}]. \tag{6.5}$$

Since $h^{(j-1)}_{d-j+2,d-j} \neq 0$, it is clear that $s_j \neq 0$ and $h^{(j)}_{d-j+2,d-j+1} \neq 0$. As a result, the matrix

$$H^{(j)} \equiv G_j^* H^{(j-1)} G_j$$

is an upper Hessenberg except for its entry

$$h^{(j)}_{d-j+1,d-j-1} = s_j h^{(0)}_{d-j,d-j-1} \neq 0.$$

The unitary transformation determined by G_j modifies only $(d-j)$th and $(d-j+1)$st rows and columns of $H^{(j-1)}$. Therefore, the subdiagonal entries of $H^{(j)}$ satisfy $h^{(j)}_{i+1,i} = h^{(0)}_{i+1,i} \neq 0$ for $i = 1, \ldots, d-j-2$, and, since $h^{(j)}_{d-j+2,d-j+1} \neq 0$, we have shown inductively that indeed $h^{(j)}_{i+1,i} \neq 0$ for $i = d-j+1, \ldots, d-1$. In the end, we receive the unitary matrix $G = G_1 \cdots G_{d-1}$ and the upper Hessenberg matrix $H^{(d-1)} = G^* H^{(0)} G$ with $h^{(d-1)}_{i+1,i} \neq 0$ for $i = 2, \ldots, d-1$. To complete the proof we need to show that the initial parameters s_1, c_1 can be chosen so that, first, $h^{(d-1)}_{2,1} \neq 0$ ($H^{(d-1)}$ is unreduced), and, second, $[\, h^{(d-1)}_{1,d-1}, h^{(d-1)}_{2,d-1} \,] \neq [0, 0]$.

First, if $h^{(d-1)}_{2,1} = 0$, then we must have $h^{(d-1)}_{1,1} \neq 0$, for if otherwise $H^{(d-1)}$ would be singular. From $H^{(d-1)} = G^* H^{(0)} G$ we receive $H^{(0)} G = G H^{(d-1)}$, and thus the first column of G is an eigenvector of $H^{(0)}$ corresponding to the eigenvalue $h^{(d-1)}_{1,1}$. Note that the first column of G depends on our choice of s_1, while the matrix $H^{(0)}$ is fixed and has at most d linearly independent eigenvectors. Apparently, the case $h^{(d-1)}_{2,1} = 0$ happens only for a finite number of values of s_1 (if any); almost every initial choice of s_1 will yield $h^{(d-1)}_{2,1} \neq 0$.

Second, we have assumed that the first $d-1$ columns of $H^{(0)}$ form an unreduced $(s+2)$-band Hessenberg matrix with $s+3 < d$, and therefore $h^{(0)}_{1,d-2} = h^{(0)}_{1,d-1} = 0$ (see Figure 6.1). Denote the entries of the (lower Hessenberg) matrix G by $g_{i,j}$. It is easy to see that $g_{d,d-1} = -c_2 s_1$. Again consider the matrix equation $H^{(0)} G = G H^{(d-1)}$. Comparing the entries in position $(1, d-1)$ on both sides shows that

$$-c_2 s_1 h^{(0)}_{1,d} \;=\; g_{1,1} h^{(d-1)}_{1,d-1} + g_{1,2} h^{(d-1)}_{1,d-1}\,, \tag{6.6}$$

where $h^{(0)}_{1,d} \neq 0$ and $s_1 \neq 0$. Therefore, to show that $[h^{(d-1)}_{1,d-1}, h^{(d-1)}_{2,d-1}] \neq [0, 0]$, it suffices to show that $c_2 \neq 0$. For c_2 it holds that (cf. (6.5))

$$h^{(1)}_{d,d-2} c_2 - h^{(1)}_{d,d-1} s_2 \;=\; 0.$$

We know that $h^{(1)}_{d,d-2} \neq 0 \neq s_2$. Thus, $c_2 = 0$ if and only if $h^{(1)}_{d,d-1} = 0$, which holds if and only if

$$c_1 s_1 h^{(0)}_{d-1,d-1} + c_1^2 h^{(0)}_{d,d-1} - s_1^2 h^{(0)}_{d-1,d} - c_1 s_1 h^{(0)}_{d,d} \;=\; 0.$$

We write $s_1 = \sin(\theta)$, $c_1 = \cos(\theta)$ and apply standard identities for trigonometric functions to see that the above equation is equivalent with

$$\left(h^{(0)}_{d-1,d-1} - h^{(0)}_{d,d}\right)\sin(2\theta) + \left(h^{(0)}_{d,d-1} + h^{(0)}_{d-1,d}\right)\cos(2\theta) + \left(h^{(0)}_{d,d-1} - h^{(0)}_{d-1,d}\right) \;=\; 0\,.$$

The left-hand side in this equation is a nontrivial trigonometric polynomial of degree two, which has at most two roots in the interval $[0, 2\pi)$. Consequently, for almost all choices of s_1 we receive $c_2 \neq 0$, giving a nonzero right-hand side in (6.6). Hence, for almost all choices of s_1, we must have $[h^{(d-1)}_{1,d-1}, h^{(d-1)}_{2,d-1}] \neq [0, 0]$. □

We can now derive the contradiction to (6.1). Consider the relation (6.2), where H_d is of the form assumed in the Lemma 6.1. Without loss of generality we may assume that the columns of V_d are normalized (normalization does not alter the nonzero pattern of H_d). By Lemma 6.1, there exists a unitary matrix G such that $\widetilde{H}_d = G^* H_d G$ is unreduced upper Hessenberg with either $\widetilde{h}_{1,d-1}$ or $\widetilde{h}_{2,d-1}$ nonzero. Then (6.2) is equivalent with

$$\widehat{A}(V_d G) \;=\; (V_d G)\widetilde{H}_d\,. \tag{6.7}$$

Denote the entries of G by $g_{i,j}$, and let $V_d G \equiv [y_1, \ldots, y_d]$. Then, since the basis $v_1, \ldots, v_d$ is orthonormal and the matrix G is unitary, the basis $y_1, \ldots, y_d$ is orthonormal,

$$(y_i, y_j) \;=\; \left(\sum_{k=1}^{d} v_k g_{k,i}, \sum_{k=1}^{d} v_k g_{k,j} \right) \;=\; \sum_{k=1}^{d} \overline{g}_{k,j} g_{k,i} \;=\; \delta_{i,j}\,,$$

where $\delta_{i,j}$ is the Kronecker delta. By (6.7), the vectors $y_1, \ldots, y_d$ form the unique (up to scaling) basis of $\mathcal{K}_d(\widehat{A}, y_1)$ generated by (3.1)–(3.3) with $\widehat{A}$ and starting vector y_1. But since $[\widetilde{h}_{1,d-1}, \widetilde{h}_{2,d-1}] \neq [0, 0]$, we see that $\widehat{A}$ (and hence A) admits for the given y_1 an optimal recurrence of length at least $d - 1$. Since we have assumed that A admits an optimal $(s+2)$-term recurrence, we must have $d - 1 \leq s + 2$, or, equivalently, $d = d_{\min}(A) \leq s + 3$. This is a contradiction since $s + 3 < d_{\min}(A)$.

As claimed at the beginning of this section, we now have shown Theorem 3.5, with the assumption $s + 2 < d_{\min}(A)$ replaced by $s + 3 < d_{\min}(A)$.

7. Concluding discussion. In this section we discuss our rather theoretical analysis above.

1. *Matrix formulation and the Faber–Manteuffel theorem.*

When formulated in terms of matrices rather than linear operators, Theorems 3.4 and 3.5 make up the Faber–Manteuffel theorem [4] in the formulation given in [12, section 2]. We state this result here for completeness.

THEOREM 7.1. *Let A be an $N \times N$ nonsingular matrix with minimal polynomial degree $d_{\min}(A)$. Let B be an $N \times N$ Hermitian positive definite matrix, and let s be a nonnegative integer, $s + 2 < d_{\min}(A)$. Then A admits for the given B an optimal $(s + 2)$-term recurrence if and only if A is B-normal(s).*

In this formulation, the Hilbert space from Theorems 3.4 and 3.5 is $\mathbb{C}^N$, equipped with the inner product generated by the Hermitian positive definite matrix B. (In case A is real, we consider B to be real as well, and the adjoint A^* is the regular transpose A^T.) The matrix A is B-normal(s) if its B-adjoint, i.e., the matrix $A^+ \equiv B^{-1}A^*B$, is a polynomial of degree s in A, and s is the smallest degree for which this is true. A complete characterization of the matrices A and B for which A is B-normal(s) is given in [12, section 3].

In this paper we have chosen the linear operator rather than the matrix formulation, because it appears to be a natural generalization. Moreover, both proofs we have given use the restriction of the linear operator A to certain cyclic invariant subspaces. In the matrix formulation, such restrictions lead to nonsquare as well as square but singular matrices. This involves a more complicated notation, which obstructs rather than helps the theoretical understanding. For instance, the restriction $\widehat{A}$ of a nonsingular $N \times N$ matrix A to a cyclic invariant subspace of A with (orthonormal) basis $v_1, \ldots, v_d$ can be represented as $\widehat{A} = VHV^*$, where $V = [v_1, \ldots, v_d]$ and H is a $d \times d$ nonsingular matrix. If $d < N$, $\widehat{A}$ is a singular $N \times N$ matrix (more precisely, $\widehat{A}$ has rank $d < N$). Any vector w in the cyclic invariant subspace can be represented as $w = V\omega$, where ω is a vector of length d containing the coefficients of w in the basis, so that $Aw = \widehat{A}w = VH\omega$, where VH is a (nonsquare) matrix of size $N \times d$. On the other hand, in the linear operator formulation, $\widehat{A}$ is invertible, and we may simply write $\widehat{A}w$ for the application of $\widehat{A}$ to any vector w in the space.

2. *On the strategies of the two different proofs of Theorem* 3.5.

The two different proofs of Theorem 3.5 given in this paper (with the second one excluding the case $s + 3 = d_{\min}(A)$; see below) follow two different strategies.

The first proof, given in section 5, is based on vectors of grade $s+2$, and works its way up to vectors of full grade $d_{\min}(A)$. This general strategy is similar to the one in the original paper of Faber and Manteuffel [4]. The details of our proof here, however, are quite different from [4]. In particular, simple arguments about the number of roots of certain polynomials (particularly in Lemmas 4.1 and 4.3) have replaced the continuity and topology arguments in the proof of [4]. We therefore consider this a simpler proof than the one given in [4].

The second proof, given in section 6, works immediately with vectors of full grade $d_{\min}(A)$. We consider this approach more elementary than our first proof. We assume that the assertion of Theorem 3.5 is false, i.e., that A admits an optimal $(s+2)$-term recurrence but is not normal(s). We show that if A is not normal(s), there must exist at least one initial vector v_1 of full grade $d = d_{\min}(A)$, for which the corresponding matrix H_d has a nonzero entry above its sth superdiagonal. If this nonzero entry already is in $H_{d,d-1}$, we are done. However, we cannot guarantee this, and therefore we need the Rotation Lemma to rotate a nonzero from the dth column of H_d into the $(d-1)$st column. This shows that A cannot admit an optimal $(s+2)$-term recurrence, contradicting our initial assumption.

3. *The Rotation Lemma and the missing case $s+3 = d_{\min}(A)$.*

In the Rotation Lemma we rotate the nonzero entry $h_{1,d}$, where $d = d_{\min}(A)$, to give $\widetilde{h}_{1,d-1} \neq 0$ *or* $\widetilde{h}_{2,d-1} \neq 0$; see Figure 6.1. Therefore, the matrix $\widetilde{H}_{d,d-1}$ is *at least* $(d-1)$-band Hessenberg. The shortest possible optimal recurrence that A may admit hence is of length $d-1$, or $s+2$ for $s = d-3$. The assumption that A admits an optimal recurrence of length $s+3 < d_{\min}(A)$ then leads to a contradiction.

To prove also the missing case $s+3 = d_{\min}(A)$, we need to guarantee that there exists a choice of s_1 so that $\widetilde{h}_{1,d-1} \neq 0$, giving a d-band Hessenberg matrix $\widetilde{H}_{d,d-1}$. Since Theorem 3.5 also holds for the case $s+3 = d_{\min}(A)$, we know that such s_1 *must exist*, but we were unable to prove the existence without using Theorem 3.5. Note, however, that in practical applications we are interested in recurrences of length $s+2 \ll d_{\min}(A)$. Therefore the missing case of the Rotation Lemma is only of rather theoretical interest.

We point out that the construction given in the Rotation Lemma, namely, the structure-preserving unitary transformation of an upper Hessenberg matrix, may be of interest beyond its application in our current context. To state this idea in a more general way, we introduce some notation. Let Ω_d be the set of the $d \times d$ unreduced upper Hessenberg matrices, and let $\Omega_d(s+2)$ be the subset consisting of the $(s+2)$-band Hessenberg matrices (these are unreduced by assumption; cf. Definition 3.1). Consider a *fixed* $H \in \Omega_d$, and define the set

$$\mathcal{R}_H \equiv \{\, G^*HG \in \Omega_d \ : \ G \text{ is unitary} \,\}.$$

Hence $\mathcal{R}_H$ is the set of all unitary transformations of H that are unreduced upper Hessenberg. Note that since $H \in \mathcal{R}_H$, the set $\mathcal{R}_H$ is nonempty. Using the Rotation Lemma (for $s+3 < d$) and Theorem 3.5 (for $s+3 = d$) the following result can be proved.

THEOREM 7.2. *Let s, d be given nonnegative integers, $s+2 < d$. For any $H \in \Omega_d$, the following assertions are equivalent:*

(1) *H is I-normal(s), i.e., $H^* = p(H)$ for a polynomial of (smallest possible) degree s;*

(2) $\mathcal{R}_H \subset \Omega_d(s+2)$.

This result means that an unreduced upper Hessenberg matrix H is I-normal(s) if and only if H is $(s+2)$-band Hessenberg, and all unitary transformations that preserve the unreduced upper Hessenberg structure of H also preserve the $(s+2)$-band structure of H.

4. *What distinguishes Theorem* 3.5 *from other results about normal operators.*

Theorem 3.5 gives a *necessary* condition when an operator A is normal (of some degree s). This condition is also *sufficient*, as shown by Theorem 3.4. Hence this condition might be taken as a *definition* of normality, and it might be included among the numerous equivalent definitions in [8, 3]. We believe, however, that the nature of the result distinguishes it from the many other equivalent ones. This distinction is clear from the second proof given in section 6.

Consider the linear operator A, and any cyclic invariant subspace $\mathcal{K}_d(A, v_1)$. Then the matrix representation of A with respect to the orthogonal basis $v_1, \dots, v_d$ of $\mathcal{K}_d(A, v_1)$ generated by (3.1)–(3.3) is a $d \times d$ unreduced upper Hessenberg matrix H_d (cf. (6.2), where this is shown for the restriction of A to $\mathcal{K}_d(A, v_1)$). Typically, equivalent results for normality are derived using knowledge of *the whole matrix*, H_d in this case. But Theorem 3.5 is based only on knowledge of *a part of the matrix*, namely, the first $d-1$ columns of H_d. Our experience in this area shows that this difference also is the reason why Theorem 3.5 is rather difficult to prove, particularly when compared with other results about normal matrices or operators.

Acknowledgments. We thank Tom Manteuffel and an anonymous referee for suggestions that helped us to improve the presentation of the results.

REFERENCES

[1] W. E. Arnoldi, *The principle of minimized iteration in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 17–29.

[2] E. W. Cheney, *Introduction to Approximation Theory*, McGraw-Hill, New York, 1966.

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

[4] V. Faber and T. Manteuffel, *Necessary and sufficient conditions for the existence of a conjugate gradient method*, SIAM J. Numer. Anal., 21 (1984), pp. 352–362.

[5] D. K. Faddeev and V. N. Faddeeva, *Computational Methods of Linear Algebra*, W. H. Freeman, San Francisco, 1963.

[6] F. R. Gantmacher, *The Theory of Matrices.* Vols. 1, 2, Chelsea, New York, 1959.

[7] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, Frontiers in Appl. Math. 17, SIAM, Philadelphia, 1997.

[8] R. Grone, C. R. Johnson, E. M. de Sá, and H. Wolkowicz, *Normal matrices*, Linear Algebra Appl., 87 (1987), pp. 213–225.

[9] K. Hoffman and R. Kunze, *Linear Algebra*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1971.

[10] A. Kleppner, *The cyclic decomposition theorem*, Integral Equations Oper. Theory, 25 (1996), pp. 490–495.

[11] J. Liesen and P. E. Saylor, *Orthogonal Hessenberg reduction and orthogonal Krylov subspace bases*, SIAM J. Numer. Anal., 42 (2005), pp. 2148–2158 (electronic).

[12] J. Liesen and Z. Strakoš, *On optimal short recurrences for generating orthogonal Krylov subspace bases*, SIAM Rev., to appear.

[13] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed., SIAM, Philadelphia, 2003.

[14] H. Zha and Z. Zhang, *The Arnoldi process, short recurrences and displacement ranks*, Linear Algebra Appl., 249 (1996), pp. 169–188.

Numer Algor (2009) 51:133–142
DOI 10.1007/s11075-008-9242-3

ORIGINAL PAPER

On orthogonal reduction to Hessenberg form with small bandwidth

V. Faber · J. Liesen · P. Tichý

Received: 12 June 2008 / Accepted: 24 September 2008 /
Published online: 30 October 2008

Abstract Numerous algorithms in numerical linear algebra are based on the reduction of a given matrix A to a more convenient form. One of the most useful types of such reduction is the orthogonal reduction to (upper) Hessenberg form. This reduction can be computed by the Arnoldi algorithm. When A is Hermitian, the resulting upper Hessenberg matrix is tridiagonal, which is a significant computational advantage. In this paper we study necessary and sufficient conditions on A so that the orthogonal Hessenberg reduction yields a Hessenberg matrix with small bandwidth. This includes the orthogonal reduction to tridiagonal form as a special case. Orthogonality here is meant with respect to some given but unspecified inner product. While the main result is already implied by the Faber-Manteuffel theorem on short recurrences for orthogonalizing Krylov sequences (see Liesen and Strakoš, SIAM Rev 50:485–503, 2008), we consider it useful to present a new, less

The work of J. Liesen was supported by the Emmy Noether- and the Heisenberg-Program of the Deutsche Forschungsgemeinschaft.

The work of P. Tichý was supported by the Emmy Noether-Program of the Deutsche Forschungsgemeinschaft and by the GAAS grant IAA100300802.

V. Faber
Carnation, WA, USA
e-mail: vance.faber@gmail.com

J. Liesen ( )
Institute of Mathematics, Technical University of Berlin, Straße des 17. Juni 136,
10623 Berlin, Germany
e-mail: liesen@math.tu-berlin.de

P. Tichý
Institute of Computer Science, Academy of Sciences of the Czech Republic,
Pod vodárenskou věží 2, 18207 Prague, Czech Republic
e-mail: tichy@cs.cas.cz

technical proof. Our proof utilizes the idea of a “minimal counterexample”, which is standard in combinatorial optimization, but rarely used in the context of linear algebra.

Keywords Reduction to Hessenberg form · Krylov subspace methods · Arnoldi method · Lanczos method

Mathematics Subject Classifications (2000) 65F10 · 65F25

1 Introduction

Many applications in engineering and science lead to linear algebraic problems involving a very large matrix $A \in \mathbb{C}^{N\times N}$. A common approach to solve such problems is to reduce A to a matrix that requires significantly less storage, or that is well suited for further processing. Algebraically, such reduction amounts to finding a more convenient basis for representing A.

One of the most useful types of such reduction is the *orthogonal reduction to (upper) Hessenberg form*, which is used, for example, in modern implementations of the QR method for solving eigenvalue problems (see [10] for a recent survey), and in the GMRES method for solving linear algebraic systems [9]. A standard method for computing this reduction is the Arnoldi algorithm [1]. Given a matrix A, a hermitian positive definite (HPD) matrix $B \in \mathbb{C}^{N\times N}$ defining the B-inner product $\langle x, y\rangle_B \equiv y^* Bx$, and an initial vector $v_1 \in \mathbb{C}^N$, the Arnoldi algorithm generates a B-orthogonal basis for the (maximal) Krylov subspace of A and v_1.

More precisely, let d be the grade of v_1 with respect to A, i.e., the smallest possible degree of a polynomial p that satisfies $p(A)v_1 = 0$. Then the Arnoldi algorithm sequentially generates vectors $v_1, \ldots, v_d$, such that

$$\operatorname{span}\{v_1, \ldots, v_n\} = \operatorname{span}\{v_1, \ldots, A^{n-1}v_1\} \equiv \mathcal{K}_n(A, v_1), \quad n = 1, \ldots, d, \tag{1.1}$$

$$\langle v_i, v_j\rangle_B = 0, \qquad i \neq j, \quad i, j = 1, \ldots, d. \tag{1.2}$$

This is achieved by the following steps: For $n = 1, 2, \ldots,$

$$v_{n+1} = Av_n - \sum_{m=1}^{n} h_{m,n} v_m, \quad \text{where}$$

$$h_{m,n} = \frac{\langle Av_n, v_m\rangle_B}{\langle v_m, v_m\rangle_B}, \quad \text{for } m = 1, \ldots, n.$$

If $v_{n+1} = 0$ then stop.

Here we have stated the classical Gram-Schmidt variant of the Arnoldi algorithm. For notational convenience, the basis vectors are not normalized. Other implementations are often preferable from a numerical point of view. In this paper, however, we assume exact arithmetic only, and do not consider differences in the finite precision behavior of different implementations.

Collecting the basis vectors in a matrix V_n, and the recurrence coefficients $h_{i,j}$ in a matrix H_n, the Arnoldi algorithm can be written in the following matrix form,

$$AV_n = V_n H_n + v_{n+1} e_n^T \,, \quad n = 1, \ldots, d. \tag{1.3}$$

Here e_n is the n-th column of the identity matrix I_n and H_n is an $n \times n$ unreduced upper Hessenberg matrix given by

$$H_n = \begin{bmatrix} h_{11} & \cdots & h_{1,n-1} & h_{1,n} \\ 1 & \ddots & \vdots & \vdots \\ & \ddots & h_{n-1,n-1} & h_{n-1,n} \\ & & 1 & h_{n,n} \end{bmatrix}. \tag{1.4}$$

The B-orthogonality of the basis vectors means that $V_n^* B V_n$ is an invertible $n \times n$ diagonal matrix, $n = 1, \ldots, d$.

If d is the grade of v_1 with respect to A, then $\mathcal{K}_d(A, v_1)$ is A-invariant, and v_{d+1} must be the zero vector, so that the Arnoldi algorithm terminates at step $n = d$. Hence, at the d-th iteration step the relation (1.3) becomes

$$AV_d = V_d H_d \,. \tag{1.5}$$

Here H_d can be interpreted as the matrix representation of the linear operator A restricted to the A-invariant subspace $\mathcal{K}_d(A, v_1)$. Or, H_d can be interpreted as a *reduction of A to upper Hessenberg form*. For more about the theory and different implementations of the Arnoldi algorithm, we refer to [8, Chapter 6.3].

Now suppose that A is self-adjoint with respect to the B-inner product, i.e. that $\langle Ax, y \rangle_B = \langle x, Ay \rangle_B$ for all vectors $x, y \in \mathbb{C}^N$. This holds if and only if $A^* B = BA$, or, equivalently, the B-adjoint $A^+ \equiv B^{-1} A^* B$ satisfies $A^+ = A$. Denote $D \equiv V_d^* B V_d$. Since $A^+ = A$ we obtain, cf. (1.5),

$$H_d^* D = H_d^* V_d^* B V_d = V_d^* A^* B V_d = V_d^* B A^+ V_d = V_d^* B A V_d = D H_d \,.$$

Since D is diagonal, the upper Hessenberg matrix H_d must be tridiagonal. In other words, when A is self-adjoint with respect to the B-inner product, the Arnoldi algorithm B-orthogonally reduces A to tridiagonal form. In the special case $B = I$ and thus $A^+ = A^*$, the algorithm for computing this reduction is known as the Hermitian Lanczos algorithm [5].

Obviously, a reduction to tridiagonal form is very convenient from a numerical point of view. It is therefore of great interest to study necessary and sufficient conditions on A so that there exists an HPD matrix B for which A can be orthogonally reduced to tridiagonal form or, more generally, to banded upper Hessenberg form with small bandwidth. Apart from trivial cases, the main necessary and sufficient condition on A is that there exists an HPD matrix B for which the B-adjoint A^+ is a low degree polynomial in A. As described in [7], this result is implied by the Faber-Manteuffel theorem on the existence of short recurrences for generating orthogonal Krylov subspace bases (in particular, see [7, Fig. 2.2]). Therefore, the question whether a given matrix is orthogonally reducible to banded Hessenberg form with low bandwidth has been completely answered. A separate proof of this result has been attempted in [6], but, as described in [7], that proof is based on less rigorous definitions and applies to nonderogatory matrices A only.

The purpose of this paper is to give a new proof of the necessary and sufficient conditions for orthogonal reducibility to upper Hessenberg form with small bandwidth. After recalling the sufficiency result from [7, Theorem 2.13], we first prove the necessity result for nonderogatory matrices (similarly as in [6], but starting from more rigorous definitions). We then show the general case inductively using a “minimal counterexample” argument. This is a standard argument in combinatorial optimization, but we have rarely seen this idea applied in linear algebra. Here we show that the smallest matrix giving a counterexample for the general case must be nonderogatory. Since we know from the first step of the proof that the result holds for nonderogatory matrices, no counterexample can possibly exist.

Reducibility to banded Hessenberg form with small bandwidth, particularly tridiagonal form, is a key property in many applications. Nevertheless, we are not aware that any complete proof of the necessary and sufficient conditions, that is independent of the technically more complicated result of Faber and Manteuffel, has appeared in the literature before. We point out that unlike the proofs of the Faber-Manteuffel theorem in [2, 3] (also cf. [11] for a related proof), our proof here is entirely based on linear algebra arguments. Furthermore, we believe that the general idea of our proof is of interest in its own right, which is a main reason for writing this paper.

2 Main definitions and sufficient conditions

Suppose that $A \in \mathbb{C}^{N\times N}$ is a given matrix, $B \in \mathbb{C}^{N\times N}$ is a given HPD matrix, and $v_1 \in \mathbb{C}^N$ is a given initial vector. (When A is real, we only consider real HPD matrices B and real initial vectors v_1.) We denote the degree of the minimal polynomial of A by $d_{\min}(A)$.

Consider the corresponding Hessenberg reduction of A as in (1.5). The Krylov subspace basis vectors $v_1, \ldots, v_d$ in this reduction are defined uniquely up to scaling by the conditions (1.1)–(1.2). This means that any other set of basis vectors $\widehat{v}_1, \ldots, \widehat{v}_d$ that also satisfies (1.1)–(1.2) is given by $\widehat{v}_n = \sigma_n v_n$ for some (nonzero) scalars $\sigma_1, \ldots, \sigma_n$. In matrix form, this can be written as $\widehat{V}_d = V_d S_d$, where $S_d = \operatorname{diag}(\sigma_1, \ldots, \sigma_d)$. Hence for this other basis, A satisfies the identity

$$A\widehat{V}_d = \widehat{V}_d \widehat{H}_d,$$

where $\widehat{H}_d = S_d^{-1} H_d S_d$. Clearly, the nonzero patterns of H_d and $\widehat{H}_d$ coincide. In particular, the upper bandwidth of H_d is independent of the algorithm that is used to compute the orthogonal reduction to Hessenberg form.

In this paper we are mostly interested in this upper bandwidth. We say that H_d is $(s+2)$-*band Hessenberg*, when the s-th superdiagonal of H_d contains at least one nonzero entry, and all entries above the s-th superdiagonal are zero. (Here the diagonal of H_d is considered the 0-th superdiagonal.) We can now rigorously define the concept of reducibility to banded upper Hessenberg form. We use the same definition as in [7, Definition 2.11], with the exception that here we do not require A to be nonsingular.

Definition 2.1 Let $A \in \mathbb{C}^{N\times N}$, let $B \in \mathbb{C}^{N\times N}$ be an HPD matrix, and let s be a nonnegative integer, $s+2 \leq d_{\min}(A)$.

(1) If for an initial vector v_1 the matrix H_d in (1.5) is $(s+2)$-band Hessenberg, then we say that A is reducible for the given B and v_1 to $(s+2)$-band Hessenberg form.
(2) If A is reducible for the given B and any initial vector v_1 to at most $(s+2)$-band Hessenberg form, while it is reducible for the given B and at least one v_1 to $(s+2)$-band Hessenberg form, then we say that A is reducible for the given B to $(s+2)$-band Hessenberg form.

Let us briefly explain why we assume $s+2 \leq d_{\min}(A)$ in this definition. First, by this assumption we exclude the trivial case $d_{\min}(A) \leq 1$, in which each initial vector v_1 is an eigenvector of A. Second, the grade d of any initial vector v_1 is at most $d_{\min}(A)$, and hence the corresponding Hessenberg matrix H_d in (1.5) has at most $d_{\min}(A)+1$ nonzero bands. Consequently, for all nonnegative intergers s with $s+2 > d_{\min}(A)$, the question whether H_d is $(s+2)$-band Hessenberg is uninteresting, since in this case the upper triangle of H_d is allowed to be completely full.

Note that by this definition the integer s is *uniquely determined*. This means that when A is reducible for the given B to $(s+2)$-band Hessenberg form, then A is *not* reducible for this B to $(t+2)$-band Hessenberg form for any $t \neq s$.

Definition 2.2 Let $A \in \mathbb{C}^{N\times N}$, and let $B \in \mathbb{C}^{N\times N}$ be HPD. Suppose that

$$A^{+} \equiv B^{-1}A^{*}B = p_s(A)\,, \tag{2.1}$$

where p_s is a polynomial of the smallest possible degree s having this property. Then A is called normal of degree s with respect to B, or, shortly, B-normal(s).

Using this definition, it is possible to prove the following sufficiency result for reducibility to $(s+2)$-band Hessenberg form; see [7, Theorem 2.13] (also cf. [4] for an analysis of the sufficient conditions in case $B = I$).

Theorem 2.3 *Let $A \in \mathbb{C}^{N\times N}$, let $B \in \mathbb{C}^{N\times N}$ be an HPD matrix, and let s be a nonnegative integer, $s+2 \leq d_{\min}(A)$. If A is B-normal(s), then A is reducible for the given B to $(s+2)$-band Hessenberg form.*

Our statement of the sufficiency result is a little bit different from the one in [7, Theorem 2.13]. Here we assume that $s+2 \leq d_{\min}(A)$, while [7, Theorem 2.13] assumes $s+2 < d_{\min}(A)$. The assumption in [7] is made for notational consistency in that paper; extending the result to the case $s+2 = d_{\min}(A)$ is straightforward. Furthermore, we have formulated the result for general matrices A, while in [7] it is assumed that A is nonsingular. The extension to the singular case is easy.

3 Necessary conditions

In this section we prove the reverse direction of Theorem 2.3, i.e., we show that if A is reducible for the given B to $(s+2)$-band Hessenberg form, where $s+2 \leq d_{\min}(A)$, then A is B-normal(s). Our proof is based on three technical lemmas.

In the first lemma, we adopt [2, Lemma 4.3] to the notation used in this paper, and we generalize the assertion to include the case of singular A.

Lemma 3.1 *Let $A \in \mathbb{C}^{N\times N}$, let $B \in \mathbb{C}^{N\times N}$ be an HPD matrix, and let s be a nonnegative integer, $s+2 \leq d_{\min}(A)$. The matrix A is B-normal(s) if and only if,*

$$A^+ v \in \mathcal{K}_{s+1}(A, v) \quad \textit{for all vectors } v \textit{ of grade } d_{\min}(A), \tag{3.1}$$

and there exists a vector v such that $A^+ v \notin \mathcal{K}_s(A, v)$.

Proof Let A be B-normal(s). Then for each v, $A^+ v = p_s(A)v \in \mathcal{K}_{s+1}(A, v)$. Moreover, since s is the smallest degree of a polynomial for which $A^+ = p_s(A)$, there must exist a vector v such that $A^+ v \notin \mathcal{K}_s(A, v)$.

In the proof of the other direction we first suppose that A is nonsingular. Then by [2, Lemma 4.3], (3.1) implies that A is B-normal(t) for some $t \leq s$. Since there exists a vector v such that $A^+ v \notin \mathcal{K}_s(A, v)$, we must have $t \geq s$, and thus $t = s$.

Now suppose that A is singular. Then there exists a scalar $\mu \in \mathbb{C}$ such that $C \equiv A + \mu I$ is nonsingular. Clearly, $d_{\min}(A) = d_{\min}(C)$. Furthermore, note that for any vector v of grade $d_{\min}(A)$, we have $\mathcal{K}_{s+1}(A, v) = \mathcal{K}_{s+1}(C, v)$. Moreover, since

$$A^+ = B^{-1}A^*B = B^{-1}C^*B - \overline{\mu}I = C^+ - \overline{\mu}I,$$

$A^+ v \in \mathcal{K}_{s+1}(A, v)$ holds if and only if $C^+ v \in \mathcal{K}_{s+1}(C, v)$. Hence, if the singular matrix A satisfies the assertion, then the nonsingular matrix $C = A + \mu I$ satisfies the assertion as well, so that C must be B-normal(s). But $C^+ = p_s(C)$ implies that $A^+ = q_s(A)$, where q_s is a polynomial of (smallest possible) degree s. Hence A is B-normal(s) as well, which finishes the proof. □

In the next lemma we prove the necessity result for nonderogatory matrices A (see also [6, pp. 2156–2157] for a similar argument).

Lemma 3.2 *Let $A \in \mathbb{C}^{N\times N}$ be a nonderogatory matrix, i.e., $d_{\min}(A) = N$. Let $B \in \mathbb{C}^{N\times N}$ be an HPD matrix, and let s be a nonnegative integer, $s+2 \leq d_{\min}(A)$. If A is reducible for the given B to $(s+2)$-band Hessenberg form, then A is B-normal(s).*

Proof We prove the assertion by contradiction. Suppose that A is reducible for the given B to $(s+2)$-band Hessenberg form, but that A is *not* B-normal(s). By Lemma 3.1, there either exists an integer $t < s$ such that $A^+v_1 \in \mathcal{K}_{t+1}(A, v_1)$ for all vectors v_1, or there exists a vector v_1 of grade $d_{\min}(A) = N$ such that $A^+v_1 \notin \mathcal{K}_{s+1}(A, v_1)$.

In the first case, one can easily show that the matrix A is reducible to (at most) $(t+2)$-band Hessenberg form, which is a contradiction since $t < s$.

In the second case, consider a vector v_1 of grade N such that $A^+v_1 \notin \mathcal{K}_{s+1}(A, v_1)$. Since v_1 is of full grade, we know that there exist scalars $\beta_1, \ldots, \beta_N \in \mathbb{C}$, such that

$$A^+v_1 = \sum_{j=1}^{N} \beta_j v_j ,$$

where $v_1, \ldots, v_N$ is the B-orthogonal basis of $\mathcal{K}_N(A, v_1)$ generated by the Arnoldi algorithm. By assumption, at least one β_j, $s+2 \leq j \leq N$, is nonzero. If this nonzero scalar is β_k, then the entry $h_{1,k}$ of H_N satisfies

$$h_{1,k} = \frac{\langle Av_k, v_1\rangle_B}{\langle v_1, v_1\rangle_B} = \frac{\langle v_k, A^+v_1\rangle_B}{\langle v_1, v_1\rangle_B} = \overline{\beta}_k \frac{\langle v_k, v_k\rangle_B}{\langle v_1, v_1\rangle_B} \neq 0 .$$

But since $k \geq s+2$, this means that H_d is *not* $(s+2)$-band Hessenberg, which contradicts our assumption that A is reducible to $(s+2)$-band Hessenberg form. □

We next show that the "minimal counterexample" of a matrix A that is reducible for the given B to $(s+2)$-band Hessenberg form but that is *not* B-normal(s) must be nonderogatory.

Lemma 3.3 *Suppose that s is a given nonnegative integer. Let A be a square matrix of smallest possible dimension N and with $d_{\min}(A) \geq s+2$ such that the following holds: There exists HPD matrix $B \in \mathbb{C}^{N\times N}$ such that*

1. *A is reducible for the given B to $(s+2)$-band Hessenberg form,*
2. *A is not B-normal(s).*

Then A is nonderogatory (i.e. $d_{\min}(A) = N$).

Proof Suppose that A is a matrix that satisfies the assumptions, and that B is the corresponding HPD matrix for which A is reducible to $(s+2)$-band Hessenberg form, but with respect to which A is not normal of degree s.

Let the Jordan normal form of A be given by $A = WJW^{-1}$, where $J = J_1 \oplus \cdots \oplus J_k$ with eigenvalues $\lambda_1, \ldots, \lambda_k$, and corresponding invariant subspaces of dimensions $s_1, \ldots, s_k$, respectively. If $k = 1$, then A is nonderogatory and we are done. Hence we may assume that $k > 1$.

Suppose that v_1 is any initial vector of grade d with respect to A and consider the corresponding Hessenberg reduction (1.5) using the B-inner product. Using the Jordan normal form of A, it is easy to see that this Hessenberg reduction is equivalent with

$$J\widehat{V}_d = \widehat{V}_d H_d , \qquad \widehat{V}_d^* \widehat{B} \widehat{V}_d \quad \text{diagonal} , \tag{3.2}$$

where $\widehat{V}_d \equiv W^{-1} V_d$ and $\widehat{B} \equiv W^* B W$, which is HPD. Note that the Hessenberg matrices in the Hessenberg reduction of A and in (3.2) coincide. Since A is reducible for the given B to $(s+2)$-band Hessenberg form, J is reducible for the given $\widehat{B}$ to $(s+2)$-band Hessenberg form (and vice versa).

It suffices to show that J is nonderogatory. Suppose not. Then there are two Jordan blocks, say J_1 and J_2 with $s_1 \leq s_2$, that correspond to the same eigenvalue (i.e. $\lambda_1 = \lambda_2$). Define the $(N - s_1) \times (N - s_1)$ matrix $\widetilde{J} \equiv J_2 \oplus \cdots \oplus J_k$, which satisfies $d_{\min}(\widetilde{J}) = d_{\min}(J) \geq s + 2$. Now define an inner product $[\cdot, \cdot]$ on $\mathbb{C}^{N-s_1} \times \mathbb{C}^{N-s_1}$ by

$$[x, y] \equiv \langle 0_{s_1} \oplus x, 0_{s_1} \oplus y \rangle_{\widehat{B}} . \tag{3.3}$$

Here 0_{s_1} denotes the zero vector of length s_1. This inner product is generated by an HPD matrix $\widetilde{B}$, $[x, y] = y^* \widetilde{B} x$ for all vectors x and y. Using the standard basis vectors and the definition of $[\cdot, \cdot]$ it is easy to show that $\widetilde{B}$ is the $(N - s_1) \times (N - s_1)$ trailing principal submatrix of $\widehat{B}$ (using MATLAB notation, $\widetilde{B} = \widehat{B}(1 + s_1 : N, 1 + s_1 : N)$).

If y_1 is any initial vector of grade d with respect to $\widetilde{J}$, then $v_1 = 0_{s_1} \oplus y_1$ is of grade d with respect to J. By construction, the corresponding Hessenberg reductions of $\widetilde{J}$ and J using the $\widetilde{B}$- and $\widehat{B}$-inner products, respectively, lead to the same unreduced upper Hessenberg matrix H_d. Consequently, the matrix $\widetilde{J}$ is reducible for $\widetilde{B}$ to $(s+2)$-band Hessenberg form.

Since $N - s_1 < N$, our initial assumption implies that the matrix $\widetilde{J}$ is $\widetilde{B}$-normal(s). Then [7, Theorem 3.1] shows: First, $\widetilde{J}$ is diagonalizable and hence diagonal, in particular $s_2 = 1$. Second, assuming that the eigenvalues of $\widetilde{J}$ are ordered so that the same eigenvalues form a single block, the HPD matrix $\widetilde{B}$ is block diagonal with block sizes corresponding to those of $\widetilde{J}$. Third, there exists a polynomial p_s of smallest possible degree s such that $p_s(\widetilde{J}) = \widetilde{J}^*$ (i.e., $p_s(\lambda_j) = \overline{\lambda}_j$ for all eigenvalues λ_j of A).

Consequently, J is diagonal with the first two eigenvalues equal, and $p_s(J) = J^*$, where p_s is a polynomial of smallest possible degree with this property. Moreover, $\widehat{B}$ is HPD and block diagonal with block sizes corresponding to those of J, except for possibly its first row and column. For simplicity of the presentation, we assume that $\widehat{B}$ is diagonal except for its first row and column;

the argument for the *block* diagonal case is more technical but mathematically analogous. Then $\widehat{B}$ has the nonzero structure

$$\widehat{B} = \begin{bmatrix} \star & \star & \cdots & \star \\ \star & \star & & \\ \vdots & & \ddots & \\ \star & & & \star \end{bmatrix}.$$

Now we reverse the roles of J_1 and J_2 and repeat the whole construction. More specifically, we denote the columns of the matrix W (from the Jordan decomposition of A) by $w_1, \ldots, w_N$. Then $A = WJW^{-1} = W_1 J W_1^{-1}$, where $W_1 \equiv [w_2, w_1, w_3, \ldots, w_N]$. Here we have used that $J_1 = J_2$ and that J is diagonal. Repeating the above construction yields a matrix $B_1 = W_1^* B W_1$, which is of the same form as $\widehat{B}$, i.e.

$$B_1 = W_1^* B W_1 = \begin{bmatrix} \star & \star & \cdots & \star \\ \star & \star & & \\ \vdots & & \ddots & \\ \star & & & \star \end{bmatrix}.$$

In particular, by comparing the second row on both sides of this equation, we see that

$$w_1^* B [w_2, w_1, w_3, \ldots, w_N] = [\star,\ \star,\ 0,\ \cdots,\ 0]\ .$$

Then the first row of $\widehat{B}$ is given by $w_1^* B W = [\star,\ \star,\ 0,\ \cdots,\ 0]$, which shows that indeed $\widehat{B}$ is block diagonal with block sizes corresponding to those of J. Hence the $N \times N$ matrix J is $\widehat{B}$-normal(s), which contradicts our assumption and completes the proof. □

In the following theorem we state the main result of this paper. The sufficiency part has already been stated in Theorem 2.3 above and is repeated here for completeness.

Theorem 3.4 *Let $A \in \mathbb{C}^{N \times N}$, let $B \in \mathbb{C}^{N \times N}$ be an HPD matrix, and let s be a nonnegative integer, $s + 2 \leq d_{\min}(A)$. The matrix A is B-normal(s) if and only if A is reducible for the given B to $(s+2)$-band Hessenberg form.*

Proof We only have to show that if A is reducible for the given B to $(s+2)$-band Hessenberg form, then A is B-normal(s). By Lemma 3.2, this statement is true for nonderogatory matrices A. However, by Lemma 3.3, the minimal counterexample is nonderogatory. Hence there is no minimal counterexample, so that the assertion must hold. □

References

1. Arnoldi, W.E.: The principle of minimized iteration in the solution of the matrix eigenvalue problem. Q. Appl. Math. **9**, 17–29 (1951)
2. Faber, V., Liesen, J., Tichý, P.: The Faber-Manteuffel theorem for linear operators. SIAM J. Numer. Anal. **46**, 1323–1337 (2008)
3. Faber V., Manteuffel, T.: Necessary and sufficient conditions for the existence of a conjugate gradient method. SIAM J. Numer. Anal. **21**, 352–362 (1984)
4. Huckle, T.: The Arnoldi method for normal matrices. SIAM J. Matrix Anal. Appl. **15**, 479–489 (1994)
5. Lanczos, C.: An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. J. Res. Natl. Bur. Stand. **45**, 255–282 (1950)
6. Liesen, J., Saylor, P.E.: Orthogonal Hessenberg reduction and orthogonal Krylov subspace bases. SIAM J. Numer. Anal. **42**, 2148–2158 (2005)
7. Liesen, J., Strakoš, Z.: On optimal short recurrences for generating orthogonal krylov subspace bases. SIAM Rev. **50**, 485–503 (2008)
8. Saad, Y.: Iterative Methods for Sparse Linear Systems, 2nd edn. Society for Industrial and Applied Mathematics, Philadelphia (2003)
9. Saad, Y., Schultz, M.H.: GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. SIAM J. Sci. Statist. Comput. **7**, 856–869 (1986)
10. Watkins, D.S.: The QR algorithm revisited. SIAM Rev. **50**, 133–145 (2008)
11. Zha, H., Zhang, Z.: The Arnoldi process, short recursions, and displacement ranks. Linear Algebra Appl. **249**, 169–188 (1996)

SIAM J. MATRIX ANAL. APPL.
Vol. 31, No. 4, pp. 2205–2221

ON CHEBYSHEV POLYNOMIALS OF MATRICES*

VANCE FABER[†], JÖRG LIESEN[‡], AND PETR TICHÝ[§]

Abstract. The mth Chebyshev polynomial of a square matrix A is the monic polynomial that minimizes the matrix 2-norm of $p(A)$ over all monic polynomials $p(z)$ of degree m. This polynomial is uniquely defined if m is less than the degree of the minimal polynomial of A. We study general properties of Chebyshev polynomials of matrices, which in some cases turn out to be generalizations of well-known properties of Chebyshev polynomials of compact sets in the complex plane. We also derive explicit formulas of the Chebyshev polynomials of certain classes of matrices, and explore the relation between Chebyshev polynomials of one of these matrix classes and Chebyshev polynomials of lemniscatic regions in the complex plane.

Key words. matrix approximation problems, Chebyshev polynomials, complex approximation theory, Krylov subspace methods, Arnoldi's method

AMS subject classifications. 41A10, 15A60, 65F10

DOI. 10.1137/090779486

1. Introduction.

Let $A \in \mathbb{C}^{n\times n}$ be a given matrix, let $m \geq 1$ be a given integer, and let $\mathcal{M}_m$ denote the set of complex *monic* polynomials of degree m. We consider the approximation problem

$$\min_{p\in\mathcal{M}_m} \|p(A)\|, \tag{1.1}$$

where $\|\cdot\|$ denotes the matrix 2-norm (or spectral norm). As shown by Greenbaum and Trefethen [11, Theorem 2] (also cf. [13, Theorem 2.2]), problem (1.1) has a uniquely defined solution when m is smaller than $d(A)$, the degree of the minimal polynomial of A. This is a nontrivial result since the matrix 2-norm is not strictly convex, and approximation problems in such norms are in general not guaranteed to have a unique solution; see [13, pp. 853–854] for more details and an example. In this paper we assume that $m < d(A)$, which is necessary and sufficient so that the value of (1.1) is positive, and we denote the unique solution of (1.1) by $T_m^A(z)$. Note that if $A \in \mathbb{R}^{n\times n}$, then the Chebyshev polynomials of A have real coefficients, and hence in this case it suffices to consider only real monic polynomials in (1.1).

It is clear that (1.1) is unitarily invariant, i.e., that $T_m^A(z) = T_m^{U^*AU}(z)$ for any unitary matrix $U \in \mathbb{C}^{n\times n}$. In particular, if the matrix A is normal, i.e., unitarily diagonalizable, then

$$\min_{p\in\mathcal{M}_m} \|p(A)\| = \min_{p\in\mathcal{M}_m} \max_{\lambda\in\Lambda(A)} |p(\lambda)|,$$

[*] Received by the editors December 8, 2009; accepted for publication (in revised form) by M. H. Gutknecht April 5, 2010; published electronically June 24, 2010.
http://www.siam.org/journals/simax/31-4/77948.html

[†] Cloudpak, Seattle, WA 98109 (vance.faber@gmail.com).

[‡] Institute of Mathematics, Technical University of Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany (liesen@math.tu-berlin.de). The work of this author was supported by the Heisenberg Program of the Deutsche Forschungsgemeinschaft.

[§] Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 2, 18207 Prague, Czech Republic (tichy@cs.cas.cz). The work of this author is a part of the Institutional Research Plan AV0Z10300504 and was supported by GAAS grant IAA100300802 and by project M100300901 of the institutional support of the Academy of Sciences of the Czech Republic.

where $\Lambda(A)$ denotes the set of the eigenvalues of A. The (uniquely defined) mth degree monic polynomial that deviates least from zero on a compact set Ω in the complex plane is called the *mth Chebyshev polynomial*[1] *of the set* Ω. We denote this polynomial by $T_m^{\Omega}(z)$.

The last equation shows that for a normal matrix A the matrix approximation problem (1.1) is equal to the scalar approximation problem of finding $T_m^{\Lambda(A)}(z)$, and in fact $T_m^A(z) = T_m^{\Lambda(A)}(z)$. Because of these relations, problem (1.1) can be considered a generalization of a classical problem of mathematics from scalars to matrices. As a consequence, Greenbaum and Trefethen [11] as well as Toh and Trefethen [24] have called the solution $T_m^A(z)$ of (1.1) the *mth Chebyshev polynomial of the matrix* A (regardless of A being normal or not).

A motivation for studying problem (1.1) and the Chebyshev polynomials of matrices comes from their connection to Krylov subspace methods, and in particular the Arnoldi method for approximating eigenvalues of matrices [2]. In a nutshell, after m steps of this method a relation of the form $AV_m = V_m H_m + r_m e_m^T$ is computed, where $H_m \in \mathbb{C}^{m\times m}$ is an upper Hessenberg matrix, $r_m \in \mathbb{C}^n$ is the mth "residual" vector, e_m is the mth canonical basis vector of $\mathbb{C}^m$, and the columns of $V_m \in \mathbb{C}^{n\times m}$ form an orthonormal basis of the Krylov subspace $\mathcal{K}_m(A, v_1) = \text{span}\{v_1, Av_1, \ldots, A^{m-1}v_1\}$. The vector $v_1 \in \mathbb{C}^n$ is an arbitrary (nonzero) initial vector. The eigenvalues of H_m are used as approximations for the eigenvalues of A. Note that $r_m = 0$ if and only if the columns of V_m span an invariant subspace of A, and if this holds, then each eigenvalue of H_m is an eigenvalue of A.

As shown by Saad [15, Theorem 5.1], the characteristic polynomial φ_m of H_m satisfies

$$\|\varphi_m(A)v_1\| \;=\; \min_{p\in\mathcal{M}_m} \|p(A)v_1\|. \tag{1.2}$$

An interpretation of this result is that the characteristic polynomial of H_m solves the Chebyshev approximation problem for A *with respect to the given starting vector* v_1. Saad pointed out that (1.2) "seems to be the only known optimality property that is satisfied by the [Arnoldi] approximation process in the nonsymmetric case" [16, p. 171]. To learn more about this property, Greenbaum and Trefethen [11, p. 362] suggested "to disentangle [the] matrix essence of the process from the distracting effects of the initial vector," and hence study the "idealized" problem (1.1) instead of (1.2). They referred to the solution of (1.1) as the *mth ideal Arnoldi polynomial of A* (in addition to the name mth Chebyshev polynomial of A).

Greenbaum and Trefethen [11] seem to be the first who studied existence and uniqueness of Chebyshev polynomials of matrices. Toh and Trefethen [24] derived an algorithm for computing these polynomials based on semidefinite programming; see also Toh's Ph.D. thesis [21, Chapter 2]. This algorithm is now part of the SDPT3 Toolbox [23]. The paper [24] as well as [21] and [25, Chapter 29] give numerous computed examples for the norms, roots, and coefficients of Chebyshev polynomials of matrices. It is shown numerically that the lemniscates of these polynomials tend to approximate pseudospectra of A. In addition, Toh has shown that the zeros of $T_m^A(z)$ are contained in the field of values of A [21, Theorem 5.10]. This result is

[1] Pafnuti Lvovich Chebyshev (1821–1894) determined the polynomials $T_m^{\Omega}(z)$ of $\Omega = [-a, a]$ (a real interval symmetric to zero) in his 1859 paper [5], which laid the foundations of modern approximation theory. We recommend Steffens' book [18] to readers who are interested in the historical development of the subject.

part of his interesting analysis of Chebyshev polynomials of linear operators in infinite dimensional Hilbert spaces [21, Chapter 5]. The first explicit solutions for the problem (1.1) for a nonnormal matrix A we are aware of have been given in [13, Theorem 3.4]. It is shown there that $T_m^A(z) = (z-\lambda)^m$ if $A = J_\lambda$, a Jordan block with eigenvalue $\lambda \in \mathbb{C}$. Note that in this case the Chebyshev polynomials of A are independent of the size of A.

The above remarks show that problem (1.1) is a mathematically interesting generalization of the classical Chebyshev problem, which has an important application in the area of iterative methods. Yet, our survey of the literature indicates that there has been little theoretical work on Chebyshev polynomials of matrices (in particular when compared with the substantial work on Chebyshev polynomials for compact sets). The main motivation for writing this paper was to extend the existing theory of Chebyshev polynomials of matrices. Therefore we considered a number of known properties of Chebyshev polynomials of compact sets, and tried to find matrix analogues. Among these are the behavior of $T_m^A(z)$ under shifts and scaling of A, a matrix analogue of the "alternation property," as well as conditions on A so that $T_m^A(z)$ is even or odd (section 2). We also give further explicit examples of Chebyshev polynomials of some classes of matrices (section 3). For a class of block Toeplitz matrices, we explore the relation between their Chebyshev polynomials and Chebyshev polynomials of lemniscatic regions in the complex plane (section 4).

All computations in this paper have been performed using MATLAB [20]. For computing Chebyshev polynomials of matrices we have used the DSDP software package for semidefinite programming [3] and its MATLAB interface.

2. General results.

In this section we state and prove results on the Chebyshev polynomials of a general matrix A. In later sections we will apply these results to some specific examples.

2.1. Chebyshev polynomials of shifted and scaled matrices.

In the following we will write a complex (monic) polynomial of degree m as a function of the variable z and its coefficients. More precisely, for $x = [x_0, \ldots, x_{m-1}]^T \in \mathbb{C}^m$ we write

$$p(z;x) \;\equiv\; z^m - \sum_{j=0}^{m-1} x_j\, z^j \;\in\; \mathcal{M}_m. \tag{2.1}$$

Let two complex numbers, α and β, be given, and define $\delta \equiv \beta - \alpha$. Then

$$\begin{aligned} p(\beta+z;x) = p((\beta-\alpha)+(\alpha+z);x) &= (\delta+(\alpha+z))^m - \sum_{j=0}^{m-1} x_j(\delta+(\alpha+z))^j \\ &= \sum_{j=0}^{m}\binom{m}{j}\delta^{m-j}(\alpha+z)^j \;-\; \sum_{j=0}^{m-1} x_j \sum_{\ell=0}^{j}\binom{j}{\ell}\delta^{j-\ell}(\alpha+z)^\ell \\ &= (\alpha+z)^m + \sum_{j=0}^{m-1}\left(\binom{m}{j}\delta^{m-j}(\alpha+z)^j - x_j\sum_{\ell=0}^{j}\binom{j}{\ell}\delta^{j-\ell}(\alpha+z)^\ell\right) \qquad (2.2)\\ &= (\alpha+z)^m - \sum_{j=0}^{m-1}\left(\sum_{\ell=j}^{m-1}\binom{\ell}{j}\delta^{\ell-j}x_\ell - \binom{m}{j}\delta^{m-j}\right)(\alpha+z)^j \\ &\equiv (\alpha+z)^m - \sum_{j=0}^{m-1} y_j(\alpha+z)^j \equiv p(\alpha+z;y). \end{aligned}$$

A closer examination of (2.2) shows that the two vectors y and x in the identity $p(\alpha + z; y) = p(\beta + z; x)$ are related by

$$\begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{m-1} \end{bmatrix} = \begin{bmatrix} \tbinom{0}{0}\delta^0 & \tbinom{1}{0}\delta^1 & \tbinom{2}{0}\delta^2 & \cdots & \tbinom{m-1}{0}\delta^{m-1} \\ & \tbinom{1}{1}\delta^0 & \tbinom{2}{1}\delta^1 & \cdots & \tbinom{m-2}{1}\delta^{m-2} \\ & & \ddots & & \vdots \\ & & & & \tbinom{m-1}{m-1}\delta^0 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{m-1} \end{bmatrix} - \begin{bmatrix} \tbinom{m}{0}\delta^m \\ \tbinom{m}{1}\delta^{m-1} \\ \vdots \\ \tbinom{m}{m-1}\delta^1 \end{bmatrix}.$$

We can write this as

$$y = h_\delta(x), \quad \text{where} \quad h_\delta(x) \equiv M_\delta x - v_\delta. \tag{2.3}$$

The matrix $M_\delta \in \mathbb{C}^{m \times m}$ is an invertible upper triangular matrix; all its diagonal elements are equal to 1. Thus, for any $\delta \in \mathbb{C}$,

$$h_\delta \,:\, x \mapsto M_\delta x - v_\delta$$

is an invertible affine linear transformation on $\mathbb{C}^m$. Note that if $\delta = 0$, then $M_\delta = I$ (the identity matrix) and $v_\delta = 0$, so that $y = x$.

The above derivation can be repeated with αI, βI, and A replacing α, β, and z, respectively. This yields the following result.

LEMMA 2.1. *Let $A \in \mathbb{C}^{n \times n}$, $x \in \mathbb{C}^m$, $\alpha \in \mathbb{C}$, and $\beta \in \mathbb{C}$ be given. Then for any monic polynomial p of the form* (2.1)*,*

$$p(\beta I + A; x) = p(\alpha I + A; h_\delta(x)), \tag{2.4}$$

where $\delta \equiv \beta - \alpha$, and h_δ is defined as in (2.3)*.*

The assertion of this lemma is an ingredient in the proof of the following theorem.

THEOREM 2.2. *Let $A \in \mathbb{C}^{n \times n}$, $\alpha \in \mathbb{C}$, and a positive integer $m < d(A)$ be given. Denote by $T_m^A(z) = p(z; x_*)$ the mth Chebyshev polynomial of A. Then the following hold:*

$$\min_{p \in \mathcal{M}_m} \|p(A + \alpha I)\| = \min_{p \in \mathcal{M}_m} \|p(A)\|, \qquad T_m^{A+\alpha I}(z) = p(z; h_{-\alpha}(x_*)), \tag{2.5}$$

where $h_{-\alpha}$ is defined as in (2.3)*, and*

$$\min_{p \in \mathcal{M}_m} \|p(\alpha A)\| = |\alpha|^m \min_{p \in \mathcal{M}_m} \|p(A)\|, \qquad T_m^{\alpha A}(z) = p(z; D_\alpha x_*), \tag{2.6}$$

where $D_\alpha \equiv \operatorname{diag}(\alpha^m, \alpha^{m-1}, \ldots, \alpha)$.

Proof. We first prove (2.5). Equation (2.4) with $\beta = 0$ shows that $p(A; x) = p(A + \alpha I; h_{-\alpha}(x))$ holds for any $x \in \mathbb{C}^m$. This yields

$$\begin{aligned} \min_{p \in \mathcal{M}_m} \|p(A + \alpha I)\| &= \min_{x \in \mathbb{C}^m} \|p(A + \alpha I; x)\| = \min_{x \in \mathbb{C}^m} \|p(A + \alpha I; h_{-\alpha}(x))\| \\ &= \min_{x \in \mathbb{C}^m} \|p(A; x)\| = \min_{p \in \mathcal{M}_m} \|p(A)\| \end{aligned}$$

(here we have used that the transformation $h_{-\alpha}$ is invertible). To see that the polynomial $p(z; h_{-\alpha}(x_*))$ is indeed the mth Chebyshev polynomial of $A + \alpha I$, we note that

$$\|p(A + \alpha I; h_{-\alpha}(x_*))\| = \|p(A; x_*)\| = \min_{p \in \mathcal{M}_m} \|p(A)\| = \min_{p \in \mathcal{M}_m} \|p(A + \alpha I)\|.$$

The equations in (2.6) are trivial if $\alpha = 0$, so we can assume that $\alpha \neq 0$. Then the matrix D_α is invertible, and a straightforward computation yields

$$\begin{aligned}\min_{p\in\mathcal{M}_m} \|p(\alpha A)\| &= \min_{x\in\mathbb{C}^m} \|p(\alpha A; x)\| = |\alpha|^m \min_{x\in\mathbb{C}^m} \|p(A; D_\alpha^{-1}x)\| = |\alpha|^m \min_{x\in\mathbb{C}^m} \|p(A; x)\| \\ &= |\alpha|^m \min_{p\in\mathcal{M}_m} \|p(A)\|.\end{aligned}$$

Furthermore,

$$\|p(\alpha A; D_\alpha x_*)\| = |\alpha|^m \, \|p(A; x_*)\| = |\alpha|^m \min_{p\in\mathcal{M}_m} \|p(A)\| = \min_{p\in\mathcal{M}_m} \|p(\alpha A)\|,$$

so that $p(z; D_\alpha x_*)$ is the mth Chebyshev polynomial of the matrix αA. □

The fact that the “true” Arnoldi approximation problem, i.e., the right-hand side of (1.2), is translation invariant has been mentioned previously in [11, p. 361]. Hence the translation invariance of problem (1.1) shown in (2.5) is not surprising. The underlying reason is that the monic polynomials are normalized “at infinity.”

The result for the scaled matrices in (2.6), which also may be expected, has an important consequence that is easily overlooked: Suppose that for some given $A \in \mathbb{C}^{n\times n}$ we have computed the sequence of norms of problem (1.1), i.e., the quantities

$$\|T_1^A(A)\|, \;\; \|T_2^A(A)\|, \;\; \|T_3^A(A)\|, \ldots.$$

If we scale A by $\alpha \in \mathbb{C}$, then the norms of the Chebyshev approximation problem *for the scaled matrix* αA are given by

$$|\alpha| \, \|T_1^A(A)\|, \;\; |\alpha|^2 \, \|T_2^A(A)\|, \;\; |\alpha|^3 \, \|T_3^A(A)\|, \ldots.$$

A suitable scaling can therefore turn any given sequence of norms for the problem with A into a strictly monotonically decreasing (or, if we prefer, increasing) sequence for the problem with αA. For example, the matrix

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

yields

$$\|T_0^A(A)\| = 1, \;\; \|T_1^A(A)\| \approx 11.4077, \;\; \|T_2^A(A)\| = 9;$$

cf. [25, p. 280] (note that by definition $T_0^A(z) \equiv 1$ for any matrix A). The corresponding norms for the scaled matrices $\frac{1}{12} \cdot A$ and $12 \cdot A$ are then (approximately) given by

$$1, \;\; 0.95064, \;\; 0.0625, \quad \text{and} \quad 1, \;\; 136.8924, \;\; 1296,$$

respectively. In general we expect that the behavior of an iterative method for solving linear systems or for approximating eigenvalues is invariant under scaling of the given matrix. In particular, by looking at the sequence of norms of problem (1.1) *alone* we cannot determine how fast a method “converges.” In practice, we always have to measure “convergence” in some *relative* (rather than absolute) sense. Note that the quantity $\min_{p\in\mathcal{M}_m} \|p(A)\|/\|A^m\|$ is independent of a scaling of the matrix A, and hence in our context it may give relevant information. We have not explored this topic further.

2.2. Alternation property for block-diagonal matrices. It is well known that Chebyshev polynomials of compact sets Ω are characterized by an *alternation property*. For example, if $\Omega = [a, b]$ is a finite real interval, then $p(z) \in \mathcal{M}_m$ is the unique Chebyshev polynomial of degree m on Ω if and only if $p(z)$ assumes its extreme values $\pm \max_{z\in\Omega} |p(z)|$ with successively alternating signs on at least $m + 1$ points (the "alternation points") in Ω; see, e.g., [4, section 7.5]. There exist generalizations of this classical result to complex as well as to finite sets Ω; see, e.g., [6, Chapter 3] and [4, section 7.5]. The following is a generalization to block-diagonal matrices.

THEOREM 2.3. *Consider a block-diagonal matrix $A = \operatorname{diag}(A_1, \dots, A_h)$, let $k \equiv \max_{1\leq j\leq h} d(A_j)$, and let ℓ be a given positive integer such that $k \cdot \ell < d(A)$. Then the matrix $T^A_{k\cdot\ell}(A) = \operatorname{diag}(T^A_{k\cdot\ell}(A_1), \dots, T^A_{k\cdot\ell}(A_h))$ has at least $\ell + 1$ diagonal blocks $T^A_{k\cdot\ell}(A_j)$ with norm equal to $\|T^A_{k\cdot\ell}(A)\|$.*

Proof. The assumption that $k \cdot \ell < d(A)$ implies that $T^A_{k\cdot\ell}(z)$ is uniquely defined. For simplicity of notation we denote $B = T^A_{k\cdot\ell}(A)$ and $B_j \equiv T^A_{k\cdot\ell}(A_j)$, $j = 1, \dots, h$. Without loss of generality we can assume that $\|B\| = \|B_1\| \geq \cdots \geq \|B_h\|$.

Suppose that the assertion is false. Then there exists an integer i, $1 \leq i \leq \ell$, so that $\|B\| = \|B_1\| = \cdots = \|B_i\| > \|B_{i+1}\|$. Let $\delta \equiv \|B\| - \|B_{i+1}\| > 0$, and let $q_j(z) \in \mathcal{M}_k$ be a polynomial with $q_j(A_j) = 0$, $1 \leq j \leq h$. Define the polynomial

$$t(z) \equiv \prod_{j=1}^{\ell} q_j(z) \;\in\; \mathcal{M}_{k\cdot\ell}.$$

Let ϵ be a positive real number with

$$\epsilon < \frac{\delta}{\delta + \max\limits_{1\leq j\leq h} \|t(A_j)\|}.$$

Then $0 < \epsilon < 1$, and hence

$$r_\epsilon(z) \equiv (1-\epsilon)\, T^A_{k\cdot\ell}(z) + \epsilon\, t(z) \;\in\; \mathcal{M}_{k\cdot\ell}.$$

Note that $\|r_\epsilon(A)\| = \max_{1\leq j\leq h} \|r_\epsilon(A_j)\|$. For $1 \leq j \leq i$, we have

$$\|r_\epsilon(A_j)\| = (1-\epsilon)\,\|B_j\| = (1-\epsilon)\,\|B\| < \|B\|.$$

For $i + 1 \leq j \leq h$, we have

$$\begin{aligned} \|r_\epsilon(A_j)\| &= \|(1-\epsilon)\, B_j + \epsilon\, t(A_j)\| \\ &\leq (1-\epsilon)\,\|B_j\| + \epsilon\,\|t(A_j)\| \\ &\leq (1-\epsilon)\,\|B_{i+1}\| + \epsilon\,\|t(A_j)\| \\ &= (1-\epsilon)\,(\|B\| - \delta) + \epsilon\,\|t(A_j)\| \\ &= (1-\epsilon)\,\|B\| + \epsilon\,(\delta + \|t(A_j)\|) - \delta. \end{aligned}$$

Since $\epsilon\,(\delta + \|t(A_j)\|) - \delta < 0$ by the definition of ϵ, we see that $\|r_\epsilon(A_j)\| < \|B\|$ for $i+1 \leq j \leq h$. But this means that $\|r_\epsilon(A)\| < \|B\|$, which contradicts the minimality of the Chebyshev polynomial of A. □

The numerical results shown in Table 1 illustrate this theorem. We have used a block-diagonal matrix A with four Jordan blocks of size 3×3 on its diagonal, so that $k = 3$. Theorem 2.3 then guarantees that $T^A_{3\ell}(A)$, $\ell = 1, 2, 3$, has at least $\ell + 1$ diagonal blocks with the same maximal norm. This is clearly confirmed for $\ell = 1$ and

TABLE 1

Numerical illustration of Theorem 2.3*: Here* $A = \mathrm{diag}(A_1, A_2, A_3, A_4)$*, where each* $A_j = J_{\lambda_j}$ *is a* 3×3 *Jordan block. The four eigenvalues are* -3, -0.5, 0.5, 0.75.

m	$\|T_m^A(A_1)\|$	$\|T_m^A(A_2)\|$	$\|T_m^A(A_3)\|$	$\|T_m^A(A_4)\|$
1	$\underline{2.6396}$	1.4620	2.3970	$\underline{2.6396}$
2	$\underline{4.1555}$	$\underline{4.1555}$	3.6828	$\underline{4.1555}$
3	$\underline{9.0629}$	5.6303	7.6858	$\underline{9.0629}$
4	$\underline{14.0251}$	$\underline{14.0251}$	11.8397	$\underline{14.0251}$
5	$\underline{22.3872}$	20.7801	17.6382	$\underline{22.3872}$
6	$\underline{22.6857}$	$\underline{22.6857}$	20.3948	$\underline{22.6857}$

$\ell = 2$ (it also holds for $\ell = 3$). For these ℓ we observe that *exactly* $\ell + 1$ diagonal blocks achieve the maximal norm. Hence in general the lower bound of $\ell + 1$ blocks attaining the maximal norm in step $m = k \cdot \ell$ cannot be improved. In addition, we see in this experiment that the number of diagonal blocks with the same maximal norm is not necessarily a monotonically increasing function of the degree of the Chebyshev polynomial.

Now consider the matrix

$$A \;=\; \mathrm{diag}(A_1, A_2) \;=\; \left[\begin{array}{cc|cc} 1 & 1 & & \\ & 1 & & \\ \hline & & -1 & 1 \\ & & & -1 \end{array}\right].$$

Then for $p(z) = z^2 - \alpha z - \beta \in \mathcal{M}_2$ we get

$$p(A) \;=\; \left[\begin{array}{cc|cc} 1-(\alpha+\beta) & 2-\alpha & & \\ & 1-(\alpha+\beta) & & \\ \hline & & 1-(\alpha+\beta) & -2-\alpha \\ & & & 1-(\alpha+\beta) \end{array}\right].$$

For $\alpha = 0$ and *any* $\beta \in \mathbb{R}$ we will have $\|p(A)\| = \|p(A_1)\| = \|p(A_2)\|$. Hence there are infinitely many polynomials $p \in \mathcal{M}_2$ for which the two diagonal blocks have the same maximal norm. One of these polynomials is the Chebyshev polynomial $T_2^A(z) = z^2+1$. As shown by this example, the condition in Theorem 2.3 on a polynomial $p \in \mathcal{M}_{k \cdot \ell}$ that at least $\ell + 1$ diagonal blocks of $p(A)$ have equal maximal norm is in general necessary but *not* sufficient so that $p(z) = T_{k \cdot \ell}^A(z)$.

Finally, as a special case of Theorem 2.3 consider a matrix $A = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$ with distinct diagonal elements $\lambda_1, \dots, \lambda_n \in \mathbb{C}$. If $m < n$, then there are at least $m + 1$ diagonal elements λ_j with $|T_m^A(\lambda_j)| = \|T_m^A(A)\| = \max_{1 \le i \le n} |T_m^A(\lambda_i)|$. Note that $T_m^A(z)$ in this case is equal to the mth Chebyshev polynomial of the finite set $\{\lambda_1, \dots, \lambda_n\} \subset \mathbb{C}$. This shows that the Chebyshev polynomial of degree m of a set in the complex plane consisting of $n \ge m + 1$ points attains its maximum value at least at $m + 1$ points.

2.3. Chebyshev polynomials with known zero coefficients. In this section we study properties of a matrix A so that its Chebyshev polynomials have known zero coefficients. An extreme case in this respect is when $T_m^A(z) = z^m$, i.e., when all coefficients of $T_m^A(z)$, except the leading one, are zero. This happens if and only if

$$\|A^m\| \;=\; \min_{p \in \mathcal{M}_m} \|p(A)\|.$$

Equivalently, this says that the zero matrix is the best approximation of A^m from the linear space span$\{I, A, \ldots, A^{m-1}\}$ (with respect to the matrix 2-norm). To characterize this property, we recall that the dual norm to the matrix 2-norm $\|\cdot\|$ is the trace norm (also called energy norm or c_1-norm),

$$|||\, M\, ||| \equiv \sum_{j=1}^{r} \sigma_j(M)\,, \tag{2.7}$$

where $\sigma_1(M), \ldots, \sigma_r(M)$ denote the singular values of the matrix $M \in \mathbb{C}^{n\times n}$ with rank$(M) = r$. For $X \in \mathbb{C}^{n\times n}$ and $Y \in \mathbb{C}^{n\times n}$ we define the inner product $\langle X, Y\rangle \equiv$ trace(Y^*X). We can now adapt a result of Zię̧tak [27, p. 173] to our context and notation.

Theorem 2.4. *Let $A \in \mathbb{C}^{n\times n}$ and a positive integer $m < d(A)$ be given. Then*

$$T_m^A(z) = z^m$$

if and only if there exists a matrix $Z \in \mathbb{C}^{n\times n}$ with $|||\, Z\, ||| = 1$ such that

$$\langle Z, A^k\rangle = 0, \;\; k = 0, \ldots, m-1, \qquad \textit{and} \qquad \mathrm{Re}\, \langle Z, A^m\rangle = \|A^m\|\,. \tag{2.8}$$

As shown in [13, Theorem 3.4], the mth Chebyshev polynomial of a Jordan block J_λ with eigenvalue $\lambda \in \mathbb{C}$ is given by $(z-\lambda)^m$. In particular, the mth Chebyshev polynomial of J_0 is of the form z^m. A more general class of matrices with $T_m^A(z) = z^m$ is studied in section 3.1 below.

It is well known that the Chebyshev polynomials of real intervals that are symmetric with respect to the origin are alternating between even and odd, i.e., every second coefficient (starting from the highest one) of $T_m^{[-a,a]}(z)$ is zero, which means that $T_m^{[-a,a]}(z) = (-1)^m T_m^{[-a,a]}(-z)$. We next give an analogue of this result for Chebyshev polynomials of matrices.

Theorem 2.5. *Let $A \in \mathbb{C}^{n\times n}$ and a positive integer $m < d(A)$ be given. If there exists a unitary matrix P such that either $P^*AP = -A$, or $P^*AP = -A^T$, then*

$$T_m^A(z) = (-1)^m T_m^A(-z). \tag{2.9}$$

Proof. We prove the assertion only in case $P^*AP = -A$; the other case is similar. If this relation holds for a unitary matrix P, then

$$\|(-1)^m T_m^A(-A)\| = \|T_m^A(P^*AP)\| = \|P^*T_m^A(A)P\| = \|T_m^A(A)\| = \min_{p\in\mathcal{M}_m} \|p(A)\|,$$

and the result follows from the uniqueness of the mth Chebyshev polynomial of A. ▯

As a special case consider a normal matrix A and its unitary diagonalization $U^*AU = D$. Then $T_m^A(z) = T_m^D(z)$, so we may consider only the Chebyshev polynomial of the diagonal matrix D. Since $D = D^T$, the conditions in Theorem 2.5 are satisfied if and only if there exists a unitary matrix P such that $P^*DP = -D$. But this means that the set of the diagonal elements of D (i.e., the eigenvalues of A) must be symmetric with respect to the origin (i.e., if λ_j is an eigenvalue, $-\lambda_j$ is an eigenvalue as well). Therefore, whenever a discrete set $\Omega \subset \mathbb{C}$ is symmetric with respect to the origin, the Chebyshev polynomial $T_m^\Omega(z)$ is even (odd) if m is even (odd).

As an example of a nonnormal matrix, consider

$$A = \begin{bmatrix} 1 & \epsilon & & \\ & -1 & 1/\epsilon & \\ & & 1 & \epsilon \\ & & & -1 \end{bmatrix}, \quad \epsilon > 0\,,$$

which has been used by Toh [22] in his analysis of the convergence of the GMRES method. He has shown that $P^T AP = -A$, where

$$P = \begin{bmatrix} & & & 1 \\ & & -1 & \\ & 1 & & \\ -1 & & & \end{bmatrix}$$

is an orthogonal matrix.

For another example consider

$$C = \begin{bmatrix} J_\lambda & \\ & J_{-\lambda} \end{bmatrix}, \qquad J_\lambda,\, J_{-\lambda} \in \mathbb{C}^{n\times n}, \quad \lambda \in \mathbb{C}\,. \tag{2.10}$$

It is easily seen that

$$J_{-\lambda} = -I^{\pm} J_\lambda I^{\pm}, \quad \text{where} \quad I^{\pm} = \mathrm{diag}(1, -1, 1, \ldots, (-1)^{n-1}) \in \mathbb{R}^{n\times n}. \tag{2.11}$$

Using the symmetric and orthogonal matrices

$$P = \begin{bmatrix} & I \\ I & \end{bmatrix}, \qquad Q = \begin{bmatrix} I^{\pm} & \\ & I^{\pm} \end{bmatrix},$$

we receive $QPCPQ = -C$.

The identity (2.11) implies that

$$\|T_m^C(J_{-\lambda})\| \;=\; \|T_m^C(-I^{\pm} J_\lambda I^{\pm})\| \;=\; \|T_m^C(J_\lambda)\|,$$

i.e., the Chebyshev polynomials of C attain the same norm on each of the two diagonal blocks. In general, we can shift and rotate any matrix consisting of two Jordan blocks of the same size and with respective eigenvalues $\lambda, \mu \in \mathbb{C}$ into the form (2.10). It then can be shown that the Chebyshev polynomials $T_m^A(z)$ of $A = \mathrm{diag}(J_\lambda, J_\mu)$ satisfy the "norm balancing property" $\|T_m^A(J_\lambda)\| = \|T_m^A(J_\mu)\|$. The proof of this property is rather technical and we skip it for brevity.

2.4. Linear Chebyshev polynomials.

In this section we consider the linear Chebyshev problem

$$\min_{\alpha\in\mathbb{C}} \|A - \alpha I\|.$$

Work related to this problem has been done by Friedland [8], who characterized solutions of the problem $\min_{\alpha\in\mathbb{R}} \|A - \alpha B\|$, where A and B are two complex, and possibly rectangular matrices. This problem in general does not have a unique solution. More recently, Afanasjew et al. [1] have studied the restarted Arnoldi method with restart length equal to 1. The analysis of this method involves approximation problems of the form $\min_{\alpha\in\mathbb{C}} \|(A - \alpha I)v_1\|$ (cf. (1.2)), whose unique solution is $\alpha = v_1^* A v_1$.

THEOREM 2.6. *Let $A \in \mathbb{C}^{n \times n}$ be any (nonzero) matrix, and denote by $\Sigma(A)$ the span of the right singular vectors of A corresponding to the maximal singular value of A. Then $T_1^A(z) = z$ if and only if there exists a vector $w \in \Sigma(A)$ with $w^*Aw = 0$.*

Proof. If $T_1^A(z) = z$, then $\|A\| = \min_{\alpha \in \mathbb{C}} \|A - \alpha I\|$. According to a result of Greenbaum and Gurvits [10, Theorem 2.5], there exists a unit norm vector $w \in \mathbb{C}^n$, so that[2]

$$\min_{\alpha \in \mathbb{C}} \|A - \alpha I\| = \min_{\alpha \in \mathbb{C}} \|(A - \alpha I)w\|.$$

The unique solution of the problem on the right-hand side is $\alpha_* = w^*Aw$. Our assumption now implies that $w^*Aw = 0$, and the equations above yield $\|A\| = \|Aw\|$, which shows that $w \in \Sigma(A)$.

On the other hand, suppose that there exists a vector $w \in \Sigma(A)$ such that $w^*Aw = 0$. Without loss of generality we can assume that $\|w\| = 1$. Then

$$\|A\| \geq \min_{\alpha \in \mathbb{C}} \|A - \alpha I\| \geq \min_{\alpha \in \mathbb{C}} \|Aw - \alpha w\| = \min_{\alpha \in \mathbb{C}} (\|Aw\| + \|\alpha w\|) = \|Aw\|.$$

In the first equality we have used that $w^*Aw = 0$, i.e., that the vectors w and Aw are orthogonal. The assumption $w \in \Sigma(A)$ implies that $\|Aw\| = \|A\|$, and thus equality must hold throughout the above relations. In particular, $\|A\| = \min_{\alpha \in \mathbb{C}} \|A - \alpha I\|$, and hence $T_1^A(z) = z$ follows from the uniqueness of the solution. □

An immediate consequence of this result is that if zero is outside the field of values of A, then $\|T_1^A(A)\| < \|A\|$. Note that this also follows from [21, Theorem 5.10], which states that the zeros of $T_m^A(z)$ are contained in the field of values of A.

We will now study the relation between Theorem 2.4 for $m = 1$ and Theorem 2.6. Let $w \in \Sigma(A)$ and let $u \in \mathbb{C}^n$ be a corresponding left singular vector, so that $Aw = \|A\|u$ and $A^*u = \|A\|w$. Then the condition $w^*Aw = 0$ in Theorem 2.6 implies that $w^*u = 0$. We may assume that $\|w\| = \|u\| = 1$. Then the rank-one matrix $Z \equiv uw^*$ satisfies $|||Z||| = 1$,

$$0 = w^*u = \sum_{i=1}^{n} \overline{w}_i u_i = \text{trace}(Z) = \langle Z, I \rangle = \langle Z, A^0 \rangle,$$

and

$$\langle Z, A \rangle = \text{trace}(A^*uw^*) = \|A\| \, \text{trace}(ww^*) = \|A\| \sum_{i=1}^{n} w_i \overline{w}_i = \|A\|.$$

Hence Theorem 2.6 shows that $T_1^A(z) = z$ if and only if there exists a rank-one matrix Z satisfying the conditions (2.8).

Above we have already mentioned that $T_1^A(z) = z$ holds when A is a Jordan block with eigenvalue zero. It is easily seen that, in the notation of Theorem 2.6, the vector w in this case is given by the last canonical basis vector. Furthermore, $T_1^A(z) = z$ holds for any matrix A that satisfies the conditions of Theorem 2.5, i.e., $P^*AP = -A$ or $P^*AP = -A^T$ for some unitary matrix P.

[2] Greenbaum and Gurvits have stated this result for real matrices only, but since its proof mainly involves singular value decompositions of matrices, a generalization to the complex case is straightforward.

An interesting special case of Theorem 2.5 arises when the matrix A is normal, so that

$$\min_{\alpha\in\mathbb{C}} \|A-\alpha I\| = \min_{\alpha\in\mathbb{C}} \max_{\lambda_i\in\Lambda(A)} |\lambda_i-\alpha|.$$

It is well known that the unique solution α_* of this problem is given by the center of the (closed) disk of smallest radius in the complex plane that contains all the complex numbers $\lambda_1,\ldots,\lambda_n$.[3]

For nonnormal matrices this characterization of α_* is not true in general. For example, if

$$A = \left[\begin{array}{c|c} J_1 & \\ \hline & -1 \end{array}\right], \quad J_1\in\mathbb{R}^{4\times 4},$$

then the smallest circle that encloses all eigenvalues of A is centered at zero, but the solution of $\min_{\alpha\in\mathbb{C}}\|A-\alpha I\|$ is given by $\alpha_*\approx -0.4545$, and we have $\|T_1^A(A)\|\approx 1.4545 < \|A\|\approx 1.8794$.

3. Special classes of matrices.

In this section we apply our previous general results to Chebyshev polynomials of special classes of matrices.

3.1. Perturbed Jordan blocks.

Our first class consists of perturbed Jordan blocks of the form

$$A = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ \nu & & & 0 \end{bmatrix} = \nu(J_0^T)^{n-1} + J_0 \ \in\ \mathbb{C}^{n\times n}, \tag{3.1}$$

where $\nu\in\mathbb{C}$ is a complex parameter. Matrices of this form have recently been studied by Greenbaum in her analysis of upper and lower bounds for the norms of matrix functions [9]. Note that for $\nu=0$ the matrix A is a Jordan block with eigenvalue zero (and hence A is not diagonalizable), while for $\nu=1$ the matrix A is unitary (and hence unitarily diagonalizable), and has the nth roots of unity as its eigenvalues. We have $d(A)=n$ for any $\nu\in\mathbb{C}$.

THEOREM 3.1. *If A is as in* (3.1), *where $\nu\in\mathbb{C}$ is given, then, for $1\le m\le n-1$,*

$$A^m = \nu(J_0^T)^{n-m} + J_0^m, \quad \|A^m\| = \max\{1,|\nu|\}, \quad \textit{and} \quad T_m^A(z) = z^m.$$

Proof. For simplicity of notation we use $J=J_0$ in this proof. Consider an integer s, $0\le s\le n-2$. Then a simple computation yields

$$\begin{aligned}(J^T)^{n-1}J^s + J(J^T)^{n-s} &= (J^T)^{n-(s+1)}\,(J^T)^sJ^s + JJ^T(J^T)^{n-(s+1)}\\ &= (J^T)^{n-(s+1)}\,\mathrm{diag}\big(\underbrace{0,\ldots,0}_{s},1,\ldots,1\big)\\ &\quad + \mathrm{diag}\,(1,\ldots,1,0)\,(J^T)^{n-(s+1)}\\ &= (J^T)^{n-(s+1)}.\end{aligned} \tag{3.2}$$

[3] The problem of finding this disk, which is uniquely determined either by two or by three of the numbers, was first posed by Sylvester in [19]. This "paper" consists solely of the following sentence: "It is required to find the least circle which shall contain a given set of points in a plane."

We prove the first identity inductively. For $m = 1$ the statement is trivial. Suppose now that the assertion is true for some m, $1 \leq m \leq n-2$. Then

$$\begin{aligned} A^{m+1} &= (\nu (J^T)^{n-1} + J)\,(\nu (J^T)^{n-m} + J^m) \\ &= \nu^2 (J^T)^{2n-m-1} + \nu((J^T)^{n-1} J^m + J(J^T)^{n-m}) + J^{m+1} \\ &= \nu (J^T)^{n-(m+1)} + J^{m+1}, \end{aligned}$$

where in the last equality we have used (3.2).

To prove the second identity it is sufficient to realize that each row and column of A^m contains at most one nonzero entry, either ν or 1. Therefore, $\|A^m\| = \max\{1, |\nu|\}$.

Finally, note that the matrices $I, A, \ldots, A^{n-1}$ have nonoverlapping nonzero patterns. Therefore, for any $p \in \mathcal{M}_m$, $1 \leq m \leq n-1$, at least one entry of $p(A)$ is 1 and at least one entry is ν, so $\|p(A)\| \geq \max\{1, |\nu|\}$. On the other hand, we know that $\|A^m\| = \max\{1, |\nu|\}$, and uniqueness of $T_m^A(z)$ implies that $T_m^A(z) = z^m$. □

3.2. Special bidiagonal matrices. Let positive integers ℓ and h, and ℓ complex numbers $\lambda_1, \ldots, \lambda_\ell$ (not necessarily distinct) be given. We consider the matrices

$$D = \begin{bmatrix} \lambda_1 & 1 & & \\ & \lambda_2 & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_\ell \end{bmatrix} \in \mathbb{C}^{\ell \times \ell}, \quad E = (J_0^T)^{\ell-1} \in \mathbb{R}^{\ell \times \ell}, \tag{3.3}$$

and form the block Toeplitz matrix

$$B = \begin{bmatrix} D & E & & \\ & D & \ddots & \\ & & \ddots & E \\ & & & D \end{bmatrix} \in \mathbb{C}^{\ell \cdot h \times \ell \cdot h}. \tag{3.4}$$

Matrices of the form (3.4) have been used by Reichel and Trefethen [14], who related the pseudospectra of these matrices to their symbol $f_B(z) = D + zE$. Chebyshev polynomials for examples of such matrices have been studied numerically in [21, 24, 25] (cf. our examples following Theorem 3.3).

LEMMA 3.2. *In the notation established above,* $\chi_D(B) = J_0^\ell$, *where* $\chi_D(z) = (z-\lambda_1) \cdot \ldots \cdot (z-\lambda_\ell)$ *is the characteristic polynomial of* D.

Proof. Let $e_1, \ldots, e_{\ell \cdot h}$ denote the canonical basis vectors of $\mathbb{C}^{\ell \cdot h}$, and let $e_0 = e_{-1} = \cdots = e_{-\ell+1} = 0$. It then suffices to show that $\chi_D(B) e_j = e_{j-\ell}$ for $j = 1, 2, \ldots, \ell \cdot h$, or, equivalently, that

$$\chi_D(B) e_{k \cdot \ell + j} = e_{(k-1)\cdot \ell + j}, \quad k = 0, 1, \ldots, h-1, \quad j = 1, 2, \ldots, \ell. \tag{3.5}$$

To prove these relations, note that

$$\chi_D(B) = (B - \lambda_1 I) \cdot \ldots \cdot (B - \lambda_\ell I),$$

where the factors on the right-hand side commute. Consider a fixed j between 1 and ℓ. Then it follows directly from the structure of the matrix $B - \lambda_j I$ that

$$(B - \lambda_j I)\, e_{k \cdot \ell + j} \;=\; e_{k \cdot \ell + j - 1}, \quad k = 0, 1, \ldots, h-1.$$

Consequently, for $k = 0, 1, \dots, h-1$, and $j = 1, 2, \dots, \ell$,

$$\begin{aligned}\chi_D(B)\, e_{k\cdot\ell+j} &= (B-\lambda_{j+1}I)\cdot\ldots\cdot(B-\lambda_\ell I)\cdot(B-\lambda_1 I)\cdot\ldots\cdot(B-\lambda_j I)\, e_{k\cdot\ell+j}\\ &= (B-\lambda_{j+1}I)\cdot\ldots\cdot(B-\lambda_\ell I)\, e_{k\cdot\ell}\\ &= (B-\lambda_{j+1}I)\cdot\ldots\cdot(B-\lambda_\ell I)\, e_{(k-1)\cdot\ell+\ell}\\ &= e_{(k-1)\cdot\ell+j},\end{aligned}$$

which is what we needed to show. $\square$

This lemma allows us to derive the following result on the Chebyshev polynomials of the matrix B.

THEOREM 3.3. *Let B be defined as in* (3.4)*, and let $\chi_D(z)$ be the characteristic polynomial of D. Then $T^B_{k\cdot\ell}(z) = (\chi_D(z))^k$ for $k = 1, 2, \dots, h-1$.*

Proof. Let M_{ij} denote the entry at position (i, j) of the matrix M. A well-known property of the matrix 2-norm is $\|M\| \geq \max_{i,j} |M_{ij}|$. For any $p \in \mathcal{M}_{k\cdot\ell}$ we therefore have

$$\|p(B)\| \;\geq\; \max_{i,j} |p(B)_{ij}| \;\geq\; |p(B)_{1,k\cdot\ell+1}| \;=\; 1.$$

On the other hand, Lemma 3.2 implies that

$$\|(\chi_D(B))^k\| \;=\; \|J_0^{k\cdot\ell}\| \;=\; 1.$$

Hence the polynomial $(\chi_D(z))^m$ attains the lower bound on $\|p(B)\|$ for all $p \in \mathcal{M}_{k\cdot\ell}$. The uniqueness of the Chebyshev polynomial of B now implies the result. $\square$

In case $\ell = 1$, i.e., $B = J_{\lambda_1} \in \mathbb{C}^{n\times n}$, the theorem shows that $(z-\lambda_1)^m$ is the mth Chebyshev polynomial of B, $m = 1, \dots, n-1$. As mentioned above, this result was previously shown in [13, Theorem 3.4]. The proof in that paper, however, is based on a different approach, namely a characterization of matrix approximation problems in the 2-norm obtained by Ziętak [26, 27].

As a further example consider a matrix B of the form (3.4) with

$$D = \begin{bmatrix} 1 & 1 \\ 0 & -1 \end{bmatrix}. \tag{3.6}$$

This matrix B has been studied numerically in [24, Example 6] and [21, Example 6]. The minimal polynomial of D is given by $(z-1)(z+1) = z^2 - 1$, and hence $T^B_{2k}(z) = (z^2-1)^k$ for $k = 1, 2, \dots, h-1$. However, there seems to be no simple closed formula for the Chebyshev polynomials of B of odd degree. Our numerical experiments show that these polynomials (contrary to those of even degree) depend on the size of the matrix. Table 2 shows the coefficients of $T^B_m(z)$ for $m = 1, 2, \dots, 7$ for an (8×8)-matrix B (i.e., there are four blocks D of the form (3.6) on the diagonal of B). The coefficients in the rows of the table are ordered from highest to lowest. For example, $T^B_4(z) = z^4 - 2z^2 + 1$.

It is somewhat surprising that the Chebyshev polynomials change significantly when we reorder the eigenvalues on the diagonal of B. In particular, consider

$$\widetilde{B} = \begin{bmatrix} J_1 & E \\ & J_{-1} \end{bmatrix} \in \mathbb{R}^{2\ell\times 2\ell}, \tag{3.7}$$

where $E = (J_0^T)^{\ell-1} \in \mathbb{R}^{\ell\times\ell}$. The coefficients of $T^{\widetilde{B}}_m(z)$, $m = 1, 2, \dots, 7$, for an (8×8)-matrix of the form (3.7) are shown in Table 3.

Table 2
Coefficients of $T_m^B(z)$ *for an* (8×8)*-matrix* B *of the form* (3.4) *with* D *as in* (3.6).

m								
1	1	0						
2	1	0	-1.000000					
3	1	0	0.876114	0				
4	1	0	-2.000000	0	1.000000			
5	1	0	-1.757242	0	0.830598	0		
6	1	0	-3.000000	0	3.000000	0	-1.000000	
7	1	0	-2.918688	0	2.847042	0	0.927103	0

Table 3
Coefficients of $T_m^{\widetilde{B}}(z)$ *for an* (8×8)*-matrix* $\widetilde{B}$ *of the form* (3.7).

m								
1	1	0						
2	1	0	-1.595438					
3	1	0	-1.975526	0				
4	1	0	-2.858055	0	2.463968			
5	1	0	-3.125673	0	2.608106	0		
6	1	0	-3.771773	0	4.945546	0	-1.863541	
7	1	0	-4.026082	0	5.922324	0	-3.233150	0

Table 4
Coefficients of $T_m^C(z)$ *for an* (8×8)*-matrix* C *of the form* (2.10) *with* $\lambda = 1$.

m								
1	1	0						
2	1	0	-1.763931					
3	1	0	-2.194408	0				
4	1	0	-2.896537	0	2.502774			
5	1	0	-3.349771	0	3.696082	0		
6	1	0	-3.799998	0	5.092302	0	-1.898474	
7	1	0	-4.066665	0	6.199999	0	-4.555546	0

Note that the matrices B based on (3.6) and $\widetilde{B}$ in (3.7) are similar (when they are of the same size). Another matrix similar to these two is the matrix C in (2.10) with $c = 1$. The coefficients of Chebyshev polynomials of such a matrix C of size 8×8 are shown in Table 4. It can be argued that the 2-norm condition number of the similarity transformations between B, $\widetilde{B}$, and C is of order 2^ℓ (we skip details for brevity of the presentation). Hence this transformation is far from being orthogonal, which indicates that the Chebyshev polynomials of the respective matrices can be very different—and in fact they are. We were unable to determine a closed formula for any of the nonzero coefficients of the Chebyshev polynomials of $\widetilde{B}$ and C (except, of course, for the leading one). Numerical experiments indicate that these in general depend on the sizes of the respective matrices.

In Figure 1 we show the roots of the Chebyshev polynomials of degrees $m = 5$ and $m = 7$ corresponding to the examples in Tables 2–4. Each figure contains three sets of roots. All the polynomials are odd, and therefore all of them have one root at the origin.

4. Matrices and sets in the complex plane.

4. Matrices and sets in the complex plane. In this section we explore the relation between Chebyshev polynomials of matrices and of compact sets Ω in the

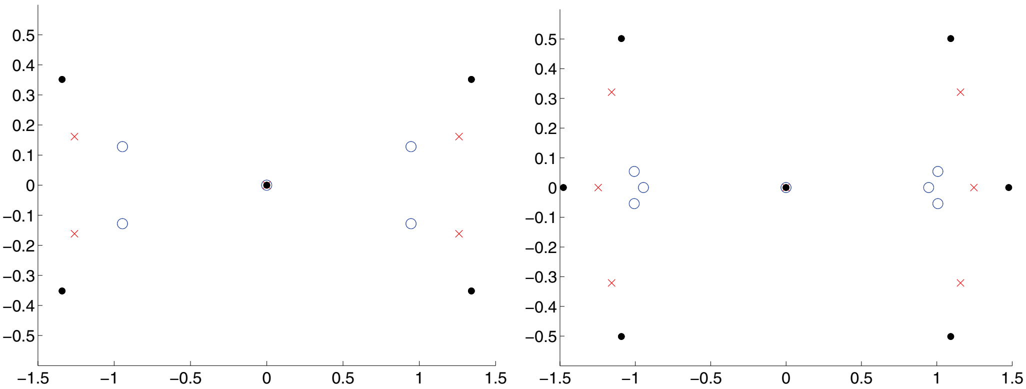


FIG. 1. *Roots of $T_m^B(z)$ (circles), $T_m^{\widetilde{B}}(z)$ (crosses), and $T_m^C(z)$ (points) of degrees $m = 5$ (left) and $m = 7$ (right) corresponding to the examples in Tables* 2–4.

complex plane. Recall that for each $m = 1, 2, \dots$ the problem

$$\min_{p \in \mathcal{M}_m} \max_{z \in \Omega} |p(z)|$$

has a unique solution $T_m^\Omega(z)$ that is called the mth Chebyshev polynomial of Ω (cf. the introduction). Similarly to the matrix case, Chebyshev polynomials of sets are known explicitly only in a few special cases. One of these cases is a disk in the complex plane centered at the point $\lambda \in \mathbb{C}$, for which the mth Chebyshev polynomial is $(z - \lambda)^m$; see, e.g., [17, p. 352]. Kamo and Borodin [12] allow us to generate more examples of Chebyshev polynomials.

THEOREM 4.1. *Let T_k^Ω be the kth Chebyshev polynomial of the infinite compact set $\Omega \subset \mathbb{C}$, let $p(z) = a_\ell z^\ell + \cdots + a_1 z + a_0$, $a_\ell \neq 0$, be a polynomial of degree ℓ, and let*

$$\Psi \;\equiv\; p^{-1}(\Omega) \;=\; \{z \in \mathbb{C} \,:\, p(z) \in \Omega\}$$

be the preimage of Ω under the polynomial map p. Then $T_{k \cdot \ell}^\Psi$, the Chebyshev polynomial of degree $m = k \cdot \ell$ of the set Ψ, is given by

$$T_m^\Psi(z) \;=\; \frac{1}{a_\ell^k}\, T_k^\Omega(p(z))\,.$$

This result has been shown also by Fischer and Peherstorfer [7, Corollary 2.2], who applied it to obtain convergence results for Krylov subspace methods. Similar ideas can be used in our context. For example, let $\mathcal{S}_A = [a, b]$ with $0 < a < b$ and $p(z) = z^2$. Then

$$\mathcal{S}_B \equiv p^{-1}(\mathcal{S}_A) = [-\sqrt{a}, -\sqrt{b}] \cup [\sqrt{a}, \sqrt{b}],$$

and Theorem 4.1 implies that $T_{2k}^{\mathcal{S}_B}(z) = T_k^{\mathcal{S}_A}(z^2)$. Such relations are useful when studying two normal matrices A and B, whose spectra are contained in the sets $\mathcal{S}_A$ and $\mathcal{S}_B$, respectively.

For an application of Theorem 4.1 that to our knowledge has not been considered before, consider a given polynomial $p = (z - \lambda_1) \cdot \ldots \cdot (z - \lambda_\ell) \in \mathcal{M}_\ell$ and the *lemniscatic region*

$$\mathcal{L}(p) \equiv \{z \in \mathbb{C} \,:\, |p(z)| \leq 1\}. \tag{4.1}$$

Note that $\mathcal{L}(p)$ is the preimage of the unit disk under the polynomial map p. Since the kth Chebyshev polynomial of the unit disk is the polynomial z^k, Theorem 4.1 implies that

$$T_{k\cdot\ell}^{\mathcal{L}(p)} = (p(z))^k.$$

Using these results and Theorem 3.3 we can now formulate the following.

THEOREM 4.2. *Let $\lambda_1, \ldots, \lambda_\ell \in \mathbb{C}$ and an integer $h > 1$ be given. Then for $p(z) = (z-\lambda_1)\cdot \ldots \cdot (z-\lambda_\ell) \in \mathcal{M}_\ell$, and each $k = 1, 2, \ldots, h-1$,*

$$(p(z))^k \;=\; T_{k\cdot\ell}^{\mathcal{L}(p)}(z) \;=\; T_{k\cdot\ell}^{B}(z),$$

where the lemniscatic region $\mathcal{L}(p)$ is defined as in (4.1)*, and the matrix B is of the form* (3.4)*. Moreover,*

$$\max_{z\in\mathcal{L}(p)} |T_{k\cdot\ell}^{\mathcal{L}(p)}(z)| \;=\; \|T_{k\cdot\ell}^{B}(B)\|.$$

This theorem connects Chebyshev polynomials of lemniscatic regions of the form (4.1) to Chebyshev polynomials of matrices B of the form (3.4). The key observation is the analogy between Theorems 3.3 and 4.1. We believe that it is possible to generate further examples along these lines.

5. Concluding remarks. We have shown that Chebyshev polynomials of matrices and Chebyshev polynomials of compact sets in the complex plane have a number of common or at least related properties. Among these are the polynomials' behavior under shifts and scalings (of matrix or set), and certain "alternation" and even/odd properties. Progress on the theory of Chebyshev polynomials of matrices can certainly be made by studying other known characteristics of their counterparts of sets in the complex plane. Furthermore, we consider it promising to further explore whether the Chebyshev polynomials of a matrix can be related to Chebyshev polynomials of a set and vice versa (see Theorem 4.2 for an example). This may give additional insight into the question of where a matrix "lives" in the complex plane.

Acknowledgments. We thank two anonymous referees for helpful comments and suggestions that improved the paper, and particulary led to a simplification of our original proof of Theorem 3.1.

REFERENCES

[1] M. Afanasjew, M. Eiermann, O. G. Ernst, and S. Güttel, *A generalization of the steepest descent method for matrix functions*, Electron. Trans. Numer. Anal., 28 (2007/08), pp. 206–222.

[2] W. E. Arnoldi, *The principle of minimized iteration in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 17–29.

[3] S. Benson, Y. Ye, and X. Zhang, *DSDP—Software for Semidefinite Programming*, Vol. 5.8, http://www.mcs.anl.gov/hs/software/DSDP (January 2006).

[4] E. K. Blum, *Numerical Analysis and Computation: Theory and Practice*, Addison–Wesley, Reading, MA, 1972.

[5] P. L. Chebyshev, *Sur les questions de minima qui se rattachent à la représentation approximative des fonctions*, Mém. Acad. Sci. St. Pétersbourg, 7 (1859), pp. 199–291.

[6] R. A. DeVore and G. G. Lorentz, *Constructive Approximation*, Grundlehren Math. Wiss. 303, Springer, New York, 1993.

[7] B. Fischer and F. Peherstorfer, *Chebyshev approximation via polynomial mappings and the convergence behaviour of Krylov subspace methods*, Electron. Trans. Numer. Anal., 12 (2001), pp. 205–215.

[8] S. Friedland, *On matrix approximation*, Proc. Amer. Math. Soc., 51 (1975), pp. 41–43.

[9] A. Greenbaum, *Upper and lower bounds on norms of functions of matrices*, Linear Algebra Appl., 430 (2009), pp. 52–65.

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

[11] A. Greenbaum and L. N. Trefethen, *GMRES/CR and Arnoldi/Lanczos as matrix approximation problems*, SIAM J. Sci. Comput., 15 (1994), pp. 359–368.

[12] S. O. Kamo and P. A. Borodin, *Chebyshev polynomials for Julia sets*, Vestnik Moskov. Univ. Ser. I Mat. Mekh., (1994), pp. 65–67.

[13] J. Liesen and P. Tichý, *On best approximations of polynomials in matrices in the matrix* 2-*norm*, SIAM J. Matrix Anal. Appl., 31 (2009), pp. 853–863.

[14] L. Reichel and L. N. Trefethen, *Eigenvalues and pseudo-eigenvalues of Toeplitz matrices*, Linear Algebra Appl., 162/164 (1992), pp. 153–185.

[15] Y. Saad, *Projection methods for solving large sparse eigenvalue problems*, in Matrix Pencils, Lecture Notes in Math. 973, B. Kågström and A. Ruhe, eds., Springer, Berlin, 1982, pp. 121–144.

[16] Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, Manchester University Press, Manchester, UK, 1992.

[17] V. I. Smirnov and N. A. Lebedev, *Functions of a Complex Variable: Constructive Theory*, translated from the Russian by Scripta Technica Ltd., MIT Press, Cambridge, MA, 1968.

[18] K.-G. Steffens, *The History of Approximation Theory: From Euler to Bernstein*, Birkhäuser, Boston, 2006.

[19] J. J. Sylvester, *A question in the geometry of situation*, Quart. J. Pure Appl. Math., 1 (1857), p. 79.

[20] The MathWorks, Inc., *MATLAB* 7.9 *(*R2009b*)*, Natick, MA, 2009.

[21] K.-C. Toh, *Matrix Approximation Problems and Nonsymmetric Iterative Methods*, Ph.D. thesis, Cornell University, Ithaca, NY, 1996.

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

[23] K. C. Toh, M. J. Todd, and R. H. Tütüncü, *SDPT*3 *version* 4.0 *(beta)—a MATLAB software for semidefinite-quadratic-linear programming*, http://www.math.nus.edu.sg/~mattohkc/sdpt3.html (February 2009).

[24] K.-C. Toh and L. N. Trefethen, *The Chebyshev polynomials of a matrix*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 400–419.

[25] L. N. Trefethen and M. Embree, *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*, Princeton University Press, Princeton, NJ, 2005.

[26] K. Ziętak, *Properties of linear approximations of matrices in the spectral norm*, Linear Algebra Appl., 183 (1993), pp. 41–60.

[27] K. Ziętak, *On approximation problems with zero-trace matrices*, Linear Algebra Appl., 247 (1996), pp. 169–183.

SIAM J. Matrix Anal. Appl.
Vol. 34, No. 4, pp. 1500–1519

PROPERTIES OF WORST-CASE GMRES*

VANCE FABER[†], JÖRG LIESEN[‡], AND PETR TICHÝ[§]

Abstract. In the convergence analysis of the GMRES method for a given matrix A, one quantity of interest is the largest possible residual norm that can be attained, at a given iteration step k, over all unit norm initial vectors. This quantity is called the worst-case GMRES residual norm for A and k. We show that the worst-case behavior of GMRES for the matrices A and A^T is the same, and we analyze properties of initial vectors for which the worst-case residual norm is attained. In particular, we prove that such vectors satisfy a certain "cross equality." We show that the worst-case GMRES polynomial may not be uniquely determined, and we consider the relation between the worst-case and the ideal GMRES approximations, giving new examples in which the inequality between the two quantities is strict at all iteration steps $k \geq 3$. Finally, we give a complete characterization of how the values of the approximation problems change in the context of worst-case and ideal GMRES for a real matrix, when one considers complex (rather than real) polynomials and initial vectors.

Key words. GMRES method, worst-case convergence, ideal GMRES, matrix approximation problems, minmax

AMS subject classifications. 65F10, 49K35, 41A52

DOI. 10.1137/13091066X

1. Introduction.

Let a nonsingular matrix $A \in \mathbb{R}^{n\times n}$ and a vector $b \in \mathbb{R}^n$ be given. Consider solving the system of linear algebraic equations $Ax = b$ with the initial guess $x_0 = 0$ using the GMRES method. This method generates a sequence of iterates $x_k \in \mathcal{K}_k(A, b) \equiv \operatorname{span}\{b, Ab, \ldots, A^{k-1}b\}$, $k = 1, 2, \ldots$, so that the corresponding kth residual $r_k \equiv b - Ax_k$ satisfies

$$\|r_k\| = \min_{p\in\pi_k} \|p(A)b\| . \tag{1.1}$$

Here $\|\cdot\|$ denotes the Euclidean norm, and π_k denotes the set of real polynomials of degree at most k and with value one at the origin; see the original paper of Saad and Schultz [14] or, e.g., the books [4, 11, 13].

The convergence analysis of GMRES deals with bounding or estimating the right-hand side of (1.1). This is a notoriously difficult problem; see, e.g., the respective chapters in [4, 11, 13]. One way to simplify this problem is to split off the right-hand-side vector b and to bound or estimate the value of the remaining polynomial matrix approximation problem only, i.e., to consider

$$\|r_k\| \leq \varphi_k(A)\,\|b\| , \quad \text{where} \quad \varphi_k(A) \equiv \min_{p\in\pi_k} \|p(A)\| . \tag{1.2}$$

Greenbaum and Trefethen nicely described the motivation for this approach in [6, pp. 361–362]. They called $\varphi_k(A)$ the *ideal GMRES* value for A and k, and the

[*] Received by the editors February 2, 2013; accepted for publication (in revised form) by D. P. O'Leary August 6, 2013; published electronically October 30, 2013.
http://www.siam.org/journals/simax/34-4/91066.html

[†] Vanco Research, Big Pine Key, FL 33043 (vance.faber@gmail.com).

[‡] Institute of Mathematics, Technical University of Berlin, 10623 Berlin, Germany (liesen@math.tu-berlin.de). The work of this author was supported by the Heisenberg Program of the Deutsche Forschungsgemeinschaft (DFG).

[§] Institute of Computer Science, Academy of Sciences of the Czech Republic, 18207 Prague, Czech Republic (tichy@cs.cas.cz). The work of this author was supported by the Grant Agency of the Czech Republic under grant P201/13-06684 S and by project M100301201 of the institutional support of the Academy of Sciences of the Czech Republic.

(uniquely determined) polynomial that attains this value the *ideal GMRES polynomial* for A and k (see [6, 12] for uniqueness proofs).

Since the majority of the existing GMRES convergence results are (upper or lower) bounds on the ideal GMRES value $\varphi_k(A)$, it is natural to ask how far this value can be from an actual kth residual norm produced by GMRES. This question was formulated by Greenbaum and Trefethen in [6, p. 366], and it can be approached by looking at the following sequence of inequalities that holds for any given $A \in \mathbb{R}^{n\times n}$, integer $k \geq 1$, and unit norm vector $b \in \mathbb{R}^n$:

$$\begin{aligned} \|r_k\| &= \min_{p\in\pi_k} \|p(A)b\| \\ &\leq \max_{\|v\|=1} \min_{p\in\pi_k} \|p(A)v\| \equiv \psi_k(A) \\ &\leq \min_{p\in\pi_k} \max_{\|v\|=1} \|p(A)v\| = \varphi_k(A). \end{aligned} \tag{1.3}$$

The value $\psi_k(A)$ introduced in (1.3) is called the *worst-case GMRES* residual norm for the given A and k. It gives an *attainable upper bound* on all possible kth GMRES residual norms for the given matrix A. A unit norm initial vector and a corresponding polynomial for which the value $\psi_k(A)$ is attained are called a *worst-case GMRES initial vector* and a *worst-case GMRES polynomial* for A and k, respectively.

Let us briefly summarize the most important previous results on worst-case and ideal GMRES (see [15, sections 1–2] for a more detailed summary). First of all, if A is singular, then $\psi_k(A) = \varphi_k(A) = 1$ for all $k \geq 1$ (to see this, simply take v as a unit norm vector in the kernel of A). Hence only nonsingular matrices A are of interest in our context. For such A, both $\psi_k(A)$ and $\varphi_k(A)$ are monotonically decreasing sequences, and $\psi_k(A) = \varphi_k(A) = 0$ for all $k \geq d(A)$, the degree of the minimal polynomial of A. Therefore, we only need to consider $1 \leq k \leq d(A) - 1$.

For a fixed k, both $\psi_k(A)$ and $\varphi_k(A)$ are continuous functions on the open set of nonsingular matrices; see [7, Theorem 3.1] or [2, Theorem 2.5]. Moreover, the equality $\psi_k(A) = \varphi_k(A)$ holds for normal matrices A and any k, as well as for $k = 1$ and any nonsingular A [5, 8]. Some nonnormal matrices A are known, however, for which $\psi_k(A) < \varphi_k(A)$, even $\psi_k(A) \ll \varphi_k(A)$, for certain k; see [2, 16].

As shown in [18], the ideal GMRES approximation problem can be formulated as a semidefinite program. Hence the ideal GMRES value $\varphi_k(A)$ and the corresponding ideal GMRES polynomial can be computed by any suitably applied semidefinite program solver. In our computations we use the MATLAB package SDPT3, version 4.0; see, e.g., [17]. On the other hand, we are not aware of any efficient algorithm for solving the worst-case GMRES approximation problem. In our experiments we use the general purpose nonlinear minimization routine `fminsearch` from MATLAB's Optimization Toolbox.

Our main goal in this paper is to contribute to the understanding of the worst-case GMRES approximation problem (1.3). In particular, we will derive special properties of worst-case GMRES initial vectors, and we will show that (in contrast to ideal GMRES), worst-case GMRES polynomials for given A and k may not be uniquely determined. Furthermore, we will give some new results on the relation between worst-case and ideal GMRES, and on the tightness of the inequality $\psi_k(A) \leq \varphi_k(A)$. Finally, we give a complete characterization of how the values of the approximation problems in the context of worst-case and ideal GMRES for a real matrix change, when one considers complex (rather than real) polynomials and initial vectors.

In this paper we do not consider quantitative estimation of the worst-case GMRES residual norm $\psi_k(A)$, and we do not study how this value depends on properties of A.

This is an important problem of great practical interest, which is largely open. For more details and a survey of the current state of the art, we refer the reader to [11, section 5.7].

2. The cross equality. In this section we generalize two results of Zavorin [19]. The first shows that $\psi_k(A) = \psi_k(A^T)$, and the second is a special property of worst-case initial vectors (they satisfy the so-called cross equality). Zavorin proved these results only for diagonalizable matrices using quite a complicated technique based on a decomposition of the corresponding Krylov matrix. Using a simple algebraic technique we prove these results for general matrices.

In our derivation we will use the following notation and basic facts about GMRES. For any given nonsingular $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$ the sequence of GMRES residual norms $\|r_k\|$, $k = 1, 2, \ldots,$ is monotonically decreasing. It terminates with $r_k = 0$ if and only if k is equal to $d(A, b)$, the degree of the minimal polynomial of b with respect to A, where always $d(A, b) \leq d(A)$. A geometric characterization of the GMRES iterate $x_k \in \mathcal{K}_k(A, b)$, which is mathematically equivalent to (1.1), is given by

$$r_k \perp A\mathcal{K}_k(A, b)\,. \tag{2.1}$$

When we need to emphasize the dependence of the kth GMRES residual r_k on A, b, and k we will write

$$r_k = \mathrm{GMRES}(A, b, k) \qquad \text{or} \qquad r_k = p_k(A)b,$$

where $p_k \in \pi_k$ is the kth GMRES polynomial of A and b, i.e., the polynomial that solves the minimization problem on the right-hand side of (1.1). As long as $r_k \neq 0$, this polynomial is uniquely determined.

Lemma 2.1. *Let $A \in \mathbb{R}^{n\times n}$ be nonsingular, let $k \geq 1$, and let $b \in \mathbb{R}^n$ be a unit norm vector such that $d(A, b) > k$. Let*

$$r_k = \mathrm{GMRES}(A, b, k), \qquad s_k = \mathrm{GMRES}\left(A^T, \frac{r_k}{\|r_k\|}, k\right).$$

Then

$$\|r_k\| \leq \|s_k\| \tag{2.2}$$

with equality if and only if

$$\frac{s_k}{\|s_k\|} = b.$$

As a consequence, if $d(A, b) > k$, then also $d(A^T, r_k) > k$.

Proof. Consider any unit norm vector b such that $1 \leq k < d(A, b)$. Then the corresponding kth GMRES residual vector $r_k = p_k(A)b$ is nonzero. The defining property (2.1) of r_k means that $\langle A^j b, r_k\rangle = 0$ for $j = 1, \ldots, k$. Hence, for any $q \in \pi_k$,

$$\|r_k\|^2 = \langle p_k(A)b, r_k\rangle = \langle b, r_k\rangle = \langle q(A)b, r_k\rangle = \langle b, q(A^T)r_k\rangle \leq \|q(A^T)r_k\|, \tag{2.3}$$

where the inequality follows from the Cauchy–Schwarz inequality and $\|b\| = 1$. Taking the minimum over all $q \in \pi_k$ in (2.3) and dividing by $\|r_k\|$ we get

$$\|r_k\| \leq \min_{q\in\pi_k} \left\| q(A^T)\frac{r_k}{\|r_k\|} \right\| = \|s_k\|.$$

Now $\|r_k\| > 0$ implies $\|s_k\| > 0$ and hence $d(A^T, r_k) > k$.

Next consider $s_k = q_k(A^T)\frac{r_k}{\|r_k\|}$ and substitute q_k for q into (2.3) to obtain

$$\|r_k\|^2 = \langle b, q_k(A^T)r_k\rangle \leq \|q_k(A^T)r_k\| = \|r_k\|\|s_k\|. \tag{2.4}$$

Therefore, $\|r_k\| = \|s_k\|$ if and only if

$$\langle b, q_k(A^T)r_k\rangle = \|q_k(A^T)r_k\|.$$

Since $\|b\| = 1$, this happens if and only if

$$b = \frac{q_k(A^T)r_k}{\|q_k(A^T)r_k\|} = \frac{q_k(A^T)r_k}{\|s_k\|\|r_k\|} = \frac{s_k}{\|s_k\|},$$

which finishes the proof. □

We now can show that the worst-case GMRES residual norms for A and A^T are identical.

THEOREM 2.2. *If $A \in \mathbb{R}^{n\times n}$ is nonsingular, then $\psi_k(A) = \psi_k(A^T)$ for all $k = 1, \ldots, d(A) - 1$.*

Proof. If b is a worst-case GMRES initial vector for A and k, $r_k = \text{GMRES}(A, b, k)$, and $s_k = \text{GMRES}(A^T, \frac{r_k}{\|r_k\|}, k)$, then, using Lemma 2.1,

$$\psi_k(A) = \|r_k\| \leq \|s_k\| \leq \psi_k(A^T). \tag{2.5}$$

Now we can reverse the roles of A and A^T to obtain the opposite inequality, i.e., $\psi_k(A^T) \leq \psi_k(A)$. □

The following theorem describes a special property of worst-case initial vectors.

THEOREM 2.3. *Let $A \in \mathbb{R}^{n\times n}$ be nonsingular, and let $1 \leq k \leq d(A) - 1$. If $b \in \mathbb{R}^n$ is a worst-case GMRES initial vector for A and k, and*

$$\begin{aligned} r_k &= p_k(A)b = \text{GMRES}(A, b, k),\\ s_k &= q_k(A^T)\frac{r_k}{\|r_k\|} = \text{GMRES}\left(A^T, \frac{r_k}{\|r_k\|}, k\right), \end{aligned}$$

then

$$\|s_k\| = \|r_k\| = \psi_k(A), \qquad b = \frac{s_k}{\psi_k(A)},$$

and

$$q_k(A^T)p_k(A)\,b = \psi_k^2(A)\,b. \tag{2.6}$$

Proof. By assumption, $\|r_k\| = \psi_k(A)$. Using Lemma 2.1 and Theorem 2.2,

$$\psi_k(A^T) = \psi_k(A) = \|r_k\| \leq \|s_k\| \leq \psi_k(A^T).$$

Therefore, $\|r_k\| = \|s_k\| = \psi_k(A)$. Using Lemma 2.1 we obtain

$$b = \frac{s_k}{\|s_k\|} = \frac{s_k}{\psi_k(A)},$$

so that $q_k(A^T)p_k(A)b = q_k(A^T)r_k = \|r_k\|s_k = \psi_k^2(A)b$. □

Equation (2.6) shows that b is an eigenvector of the matrix $q_k(A^T)p_k(A)$ with the corresponding eigenvalue $\psi_k^2(A)$. In Corollary 3.7 we will show that $q_k = p_k$, i.e., that b is a right singular vector of the matrix $p_k(A)$.

To further investigate vectors with the special property introduced in Theorem 2.3 we use the following definition.

DEFINITION 2.4. *Let $A \in \mathbb{R}^{n\times n}$ be nonsingular and let $k \geq 1$. We say that a unit norm vector $b \in \mathbb{R}^n$ with $d(A,b) > k$ satisfies the cross equality for A and k if*

$$b = \frac{s_k}{\|s_k\|}, \quad \textit{where} \quad s_k \equiv \text{GMRES}\left(A^T, \frac{r_k}{\|r_k\|}, k\right), \quad r_k \equiv \text{GMRES}(A, b, k).$$

The following algorithm is motivated by this definition. Convergence properties are shown in the theorem immediately below the algorithm statement.

Algorithm 1 (Cross iterations 1)

```
$b^{(0)} = b$,
**for** $j = 1, 2, \ldots$ **do**
    $r_k^{(j)} = \text{GMRES}(A, b^{(j-1)}, k)$
    $c^{(j-1)} = r_k^{(j)}/\|r_k^{(j)}\|$
    $s_k^{(j)} = \text{GMRES}(A^T, c^{(j-1)}, k)$
    $b^{(j)} = s_k^{(j)}/\|s_k^{(j)}\|$
**end for**
```

THEOREM 2.5. *Let $A \in \mathbb{R}^{n\times n}$ be nonsingular and let $k \geq 1$. If $b \in \mathbb{R}^n$ is any unit norm vector with $d(A,b) > k$, then the vectors generated by Algorithm* 1 *are well defined and it holds that*

$$\|r_k^{(j)}\| \leq \|s_k^{(j)}\| \leq \|r_k^{(j+1)}\| \leq \|s_k^{(j+1)}\| \leq \psi_k(A), \quad j = 1, 2, \ldots, \tag{2.7}$$

and the two sequences $\|r_k^{(j)}\|$, $j = 1, 2, \ldots$, and $\|s_k^{(j)}\|$, $j = 1, 2, \ldots$, converge to the same limit. Moreover,

$$\lim_{j\to\infty} \|b^{(j)} - b^{(j-1)}\| = 0 \quad \textit{and} \quad \lim_{j\to\infty} \|c^{(j)} - c^{(j-1)}\| = 0.$$

Proof. Using Lemma 2.1 we know that $r_k^{(1)}$ as well as $s_k^{(1)}$ are well defined and it holds that $\|r_k^{(1)}\| \leq \|s_k^{(1)}\|$. Switching the roles of A and A^T and using Lemma 2.1 again, it follows that $r_k^{(2)}$ is well defined and that $\|s_k^{(1)}\| \leq \|r_k^{(2)}\|$. Hence, (2.7) follows from Lemma 2.1 by induction.

By (2.7) the two sequences $\|r_k^{(j)}\|$ and $\|s_k^{(j)}\|$ interlace each other, are both nondecreasing, and are both bounded by $\psi_k(A)$. This implies that both sequences converge to the same limit, which does not exceed $\psi_k(A)$.

The first equality in (2.4) shows that $\|r_k^{(j)}\| = \langle b^{(j-1)}, s_k^{(j)}\rangle$. Using this fact and $b^{(j)} = s_k^{(j)}/\|s_k^{(j)}\|$ we obtain

$$\frac{1}{2}\|b^{(j)} - b^{(j-1)}\|^2 = 1 - \langle b^{(j-1)}, b^{(j)}\rangle = 1 - \langle b^{(j-1)}, s_k^{(j)}/\|s_k^{(j)}\|\rangle = 1 - \frac{\|r_k^{(j)}\|}{\|s_k^{(j)}\|}.$$

Since the sequences of norms $\|r_k^{(j)}\|$ and $\|s_k^{(j)}\|$ converge to the same limit for $j \to \infty$, their ratio converges to 1, so that $\|b^{(j)} - b^{(j-1)}\| \to 0$ for $j \to 0$.

The proof of the property for the sequence $c^{(j)}$ is analogous. □

The results in Theorem 2.5 can be interpreted as a generalization of a theorem of Forsythe from 1968 [3, Theorem 3.8] from symmetric positive definite A to general nonsingular A. As already noticed by Forsythe (for the symmetric positive definite case), there is strong numerical evidence that for each initial $b^{(0)}$ the sequence $b^{(j)}$ (resp., the sequence $c^{(j)}$) converges to a uniquely defined limit vector $\widetilde{b}$ (resp., $\widetilde{c}$). Unfortunately, we were not able to prove that this must always be the case. Such proof could be used to settle the conjecture made by Forsythe in [3, p. 66]. For a recent treatment and historical notes on this open problem we refer the reader to [1].

From the above it is clear that satisfying the cross equality represents a necessary condition for a vector $b^{(0)}$ to be a worst-case initial vector. On the other hand, we can ask whether this condition is sufficient, or, at least, whether the vectors that satisfy the cross equality are in some sense special. To investigate this question we present the following two lemmas.

Lemma 2.6. *Let $A \in \mathbb{R}^{n\times n}$ be nonsingular and let $k \geq 1$. A unit norm vector $b \in \mathbb{R}^n$ with $d(A,b) > k$ satisfies the cross equality for A and k if and only if $b \in \mathcal{K}_{k+1}(A^T, r_k)$, where $r_k = \mathrm{GMRES}(A,b,k)$. In particular, if $d(A) = n$, then each unit norm vector b with $d(A,b) = n$ satisfies the cross equality for A and $k = n-1$.*

Proof. The nonzero GMRES residual $r_k \in b + A\mathcal{K}_k(A,b) \subset \mathcal{K}_{k+1}(A,b)$ is uniquely determined by the orthogonality conditions (2.1), which can be written as

$$0 = \langle A^j b, r_k\rangle = \langle b, (A^T)^j r_k\rangle \quad \text{for} \quad j = 1,\dots,k,$$

or, equivalently,

$$b \perp A^T\mathcal{K}_k(A^T, r_k). \tag{2.8}$$

Now let $s_k = \mathrm{GMRES}(A^T, r_k/\|r_k\|, k)$. Then

$$s_k \in \frac{r_k}{\|r_k\|} + A^T\mathcal{K}_k(A^T,r_k) \subset \mathcal{K}_{k+1}(A^T,r_k), \quad s_k \perp A^T\mathcal{K}_k(A^T,r_k)\,. \tag{2.9}$$

We will now prove the equivalence. On the one hand, if b satisfies the cross equality for A and k, then $b = s_k/\|s_k\|$ and (2.9) implies that $b \in \mathcal{K}_{k+1}(A^T, r_k)$.

On the other hand, suppose that $b \in \mathcal{K}_{k+1}(A^T, r_k)$. From (2.8) it follows that also $b \perp A^T\mathcal{K}_k(A^T, r_k)$. Since $A^T\mathcal{K}_k(A^T,r_k)$ is a k-dimensional subspace of the $(k+1)$-dimensional subspace $\mathcal{K}_{k+1}(A^T,r_k)$, b has to be a multiple of s_k, i.e., $b = s_k/\|s_k\|$ or $b = -s_k/\|s_k\|$. Finally, from (2.9) we get $\langle b, s_k\rangle = \|r_k\|^{-1}\langle b, r_k\rangle = \|r_k\| > 0$. Therefore, $b = s_k/\|s_k\|$.

For $k = n-1$, we have $\mathcal{K}_{k+1}(A^T, r_k) = \mathbb{R}^n$, i.e., $b \in \mathcal{K}_{k+1}(A^T,r_k)$ is always satisfied. □

Lemma 2.7. *Let*

$$J_\lambda = \begin{bmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix} \in \mathbb{R}^{n\times n}, \quad \lambda \neq 0. \tag{2.10}$$

Then $e_n = [0,\dots,0,1]^T$ satisfies the cross equality for J_λ and every $k = 1,\dots,n-1$.

Proof. From [10, Example 2.3] we know that

$$r_k = \mathrm{GMRES}(J_\lambda, e_n, k) = \|r_k\|^2[0,\dots,0,(-\lambda)^k,(-\lambda)^{k-1},\dots,-\lambda,1]^T. \tag{2.11}$$

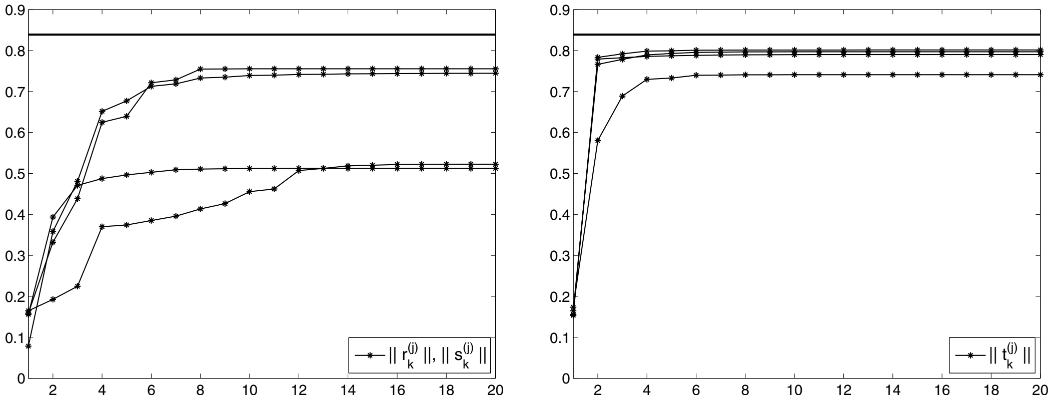


FIG. 2.1. *Cross iterations for the* 11×11 *Jordan block* J_1, $k = 5$, *and four different random initial vectors. The left part shows results for Algorithm* 1 *and the right part for Algorithm* 2. *The bold solid horizontal line represents the worst-case GMRES residual norm for* J_1.

Using Lemma 2.6, it is sufficient to show that $e_n \in \mathcal{K}_{k+1}(J_\lambda^T, r_k)$. We will look at the nonzero structure of the vectors $(J_\lambda^T)^j r_k$. First, it holds that

$$J_\lambda^T r_k = (-1)^k \|r_k\|^2 \lambda^{k+1} e_{n-k}.$$

Consequently, for $j = 1, \dots, k-1$, $(J_\lambda^T)^{j+1} r_k = (J_\lambda^T)^j (J_\lambda^T r_k)$ is a nonzero multiple of the $(n-k)$th column of $(J_\lambda^T)^j$. Hence

$$[r_k, J_\lambda^T r_k, \dots, (J_\lambda^T)^k r_k] = \left[\begin{array}{cccc} \circ & \dots & \dots & \circ \\ \vdots & & & \vdots \\ \circ & \dots & \dots & \circ \\ \bullet & \bullet & \dots & \bullet \\ \bullet & \circ & \ddots & \vdots \\ \vdots & \vdots & \ddots & \bullet \\ \bullet & \circ & \dots & \circ \end{array}\right],$$

where "$\bullet$" stands for a nonzero entry and "$\circ$" represents a zero entry. From this structure one can easily see that $e_n \in \mathcal{K}_{k+1}(J_\lambda^T, r_k)$. □

Our numerical tests predict that although e_n satisfies the cross equality for J_λ and every $k = 1, \dots, n-1$, e_n is not a worst-case GMRES initial vector for J_λ and any k. We are able to prove this statement only in special cases, for example, if $1 \le k \le n/2$ and $\lambda > 2$. In this case $\psi_k(J_\lambda) = \lambda^{-k}$ (cf. [15, Corollary 3.3]), while (2.11) shows that $r_k = \mathrm{GMRES}(J_\lambda, e_n, k)$ has the norm

$$\|r_k\| = \left(\lambda^{2k} + \sum_{j=0}^{k-1} \lambda^{2j} \right)^{-1/2} < \lambda^{-k}.$$

To give a numerical example for Algorithm 1, we consider $A = J_1 \in \mathbb{R}^{11 \times 11}$ and $k = 5$. In the left part of Figure 2.1 we plot the results of Algorithm 1 started with four random unit norm initial vectors and executed for $j = 1, 2, \dots, 10$. Each line represents one corresponding sequence $\|r_5^{(1)}\|$, $\|s_5^{(1)}\|$, $\|r_5^{(2)}\|$, $\|s_5^{(2)}\|$, $\dots$, $\|r_5^{(10)}\|$,

$\|s_5^{(10)}\|$. In each case we noted that the sequences numerically converge to uniquely defined limit vectors (cf. our remarks following the proof of Theorem 2.5). Moreover, in each case we obtain at the end a unit norm vector $b^{(10)}$ that satisfies (up to a small inaccuracy) the cross equality for J_1 and $k = 5$. We can observe that there seems to be no special structure in the norms that are attained at the end. In particular, none of the runs results in a worst-case initial vector for J_1 and $k = 5$, i.e., none of the curves attains the value $\psi_5(J_\lambda)$ that is visualized by the highest bold horizontal line in the figure.

As indicated in the left part of Figure 2.1, the sequences of residual norms generated by Algorithm 1 usually stagnate after only a few iterations. Unfortunately, this level is usually far below the worst-case level we want to reach. In order to get closer to that level, we need to disturb the process and try a different initial vector that could provide a greater GMRES residual norm. This motivates the following modification of Algorithm 1, where in each step we decide between using A or A^T to generate the next residual norm.

Algorithm 2 (Cross iterations 2)

```
$b^{(0)} = b$,
for $j = 1, 2, \dots$ do
  $v = \text{GMRES}(A, b^{(j-1)}, k)$
  $w = \text{GMRES}(A^T, b^{(j-1)}, k)$
  if $\|v\| < \|w\|$ then
    $t_k^{(j)} = w$
  else
    $t_k^{(j)} = v$
  end if
  $b^{(j)} = t_k^{(j)} / \|t_k^{(j)}\|$
end for
```

Algorithm 2 is well defined and has similar convergence properties to those stated in Theorem 2.5 for Algorithm 1. As shown in the right part of Figure 2.1, the strategy of Algorithm 2 is a little better than the one of Algorithm 1 when looking for a worst-case initial vector: It generates larger residual norms than Algorithm 1, but they are still less than the true worst-case norm. While one may use the output of Algorithm 2 as an initial point for an optimization routine like `fminsearch`, finding an efficient algorithm for computing a worst-case initial vector remains an open problem.

3. Optimization point of view.

In this section we rewrite the worst-case GMRES approximation problem (1.3) in an equivalent form in order to characterize worst-case GMRES initial vectors and the corresponding worst-case GMRES polynomials as saddle points of a certain function. This formulation will in particular be used to show that the worst-case GMRES polynomials for A and A^T are identical.

Let a nonsingular matrix $A \in \mathbb{R}^{n \times n}$ and a positive integer $k < d(A)$ be given. For vectors $c = [c_1, \dots, c_k]^T \in \mathbb{R}^k$ and $v \in \mathbb{R}^n$, we define the function

$$f(c, v) \equiv \|p(A; c)v\|^2 = v^T p(A; c)^T p(A; c)v, \tag{3.1}$$

where

$$p(z; c) = 1 - \sum_{j=1}^{k} c_j z^j.$$

Equivalently, we can express the function $f(c, v)$ using the matrix

$$K(v) \equiv [Av, A^2v, \dots, A^kv]$$

as

$$f(c, v) = \|v - K(v)c\|^2 = v^Tv - 2v^TK(v)c + c^TK(v)^TK(v)c. \tag{3.2}$$

(Here only the dependence on v is expressed in the notation $K(v)$, because A and k are both fixed.) Note that $K(v)^TK(v)$ is the Gramian matrix of the vectors $Av, A^2v, \dots, A^kv$,

$$K(v)^TK(v) = \left[v^T(A^T)^iA^jv\right]_{i,j=1,\dots,k}.$$

Next, we define the function

$$g(v) \equiv \min_{c \in \mathbb{R}^k} f(c, v),$$

which represents the kth squared GMRES residual norm for the matrix A and the initial vector v, and we denote

$$\Omega \equiv \{u \in \mathbb{R}^n : \ d(A, u) \geq k\}, \qquad \Gamma \equiv \{u \in \mathbb{R}^n : \ d(A, u) < k\}.$$

The set Γ is a closed subset, Ω is an open subset of $\mathbb{R}^n$, and $\mathbb{R}^n = \Omega \cup \Gamma$. Note that $g(v) > 0$ for all $v \in \Omega$ and $g(v) = 0$ for all $v \in \Gamma$. The following lemma is a special case of [2, Proposition 2.2] for real data and nonsingular A.

LEMMA 3.1. *In the previous notation, the function $g(v)$ is a continuous function of $v \in \mathbb{R}^n$, i.e., $g \in C^0(\mathbb{R}^n)$, and it is an infinitely differentiable function of $v \in \Omega$, i.e., $g \in C^\infty(\Omega)$. Moreover, Γ has measure zero in $\mathbb{R}^n$.*

We next characterize the minimizer of the function $f(c, v)$ as a function of v.

LEMMA 3.2. *For each given $v \in \Omega$, the problem*

$$\min_{c \in \mathbb{R}^k} f(c, v)$$

has the unique minimizer

$$c_*(v) = (K(v)^TK(v))^{-1}K(v)^Tv \in \mathbb{R}^k.$$

As a function of $v \in \Omega$, this minimizer satisfies $c_*(v) \in C^\infty(\Omega)$. Given $v \in \Omega$, $(c_*(v), v)$ is the only point in $\mathbb{R}^k \times \Omega$ with

$$\nabla_c f(c_*(v), v) = 0.$$

Proof. Since $v \in \Omega$ and A is nonsingular, the vectors $Av, A^2v, \dots, A^kv$ are linearly independent and $K(v)^TK(v)$ is symmetric and positive definite. Therefore, if $v \in \Omega$ is fixed, (3.2) is a quadratic functional in c, which attains its unique global minimum at the stationary point

$$c_*(v) = (K(v)^TK(v))^{-1}K(v)^Tv.$$

Since $K(v)^TK(v)$ is nonsingular and each entry of $(K(v)^TK(v))^{-1}$ can be expressed using Cramer's rule, the function $c_*(v)$ is a well-defined rational function of $v \in \Omega$, and thus $c_*(v) \in C^\infty(\Omega)$. Note that the vector $c_*(v)$ contains the coefficients of the kth GMRES polynomial that corresponds to the initial vector $v \in \Omega$. □

As stated in Lemma 3.1, $g(v)$ is a continuous function on $\mathbb{R}^n$, and thus it is also continuous on the unit sphere

$$S \equiv \{u \in \mathbb{R}^n : \ \|u\| = 1\}.$$

Since S is a compact set and $g(v)$ is continuous on this set, it attains its minimum and maximum on S.

We are interested in the characterization of points $(\widetilde{c}, \widetilde{v}) \in \mathbb{R}^k \times S$ such that

$$f(\widetilde{c}, \widetilde{v}) = \max_{v \in S} \min_{c \in \mathbb{R}^k} f(c, v) = \max_{v \in S} g(v). \tag{3.3}$$

This is the worst-case GMRES problem (1.3). Since $g(v) = 0$ for all $v \in \Gamma$, we have

$$\max_{v \in S} g(v) = \max_{v \in S \cap \Omega} g(v).$$

To characterize the points $(\widetilde{c}, \widetilde{v}) \in \mathbb{R}^k \times S$ that satisfy (3.3), we define for every $c \in \mathbb{R}^k$ and $v \neq 0$ the two functions

$$F(c, v) \equiv f\left(c, \frac{v}{\|v\|}\right) = \frac{f(c, v)}{v^T v}, \qquad G(v) \equiv g\left(\frac{v}{\|v\|}\right) = \frac{g(v)}{v^T v}.$$

Clearly, for any $\alpha \neq 0$, we have

$$F(c, \alpha v) = F(c, v), \qquad G(\alpha v) = G(v).$$

Lemma 3.3. *It holds that $G(v) \in C^\infty(\Omega)$. A vector $\widetilde{v} \in \Omega \cap S$ satisfies*

$$g(\widetilde{v}) \geq g(v) \quad \textit{for all} \quad v \in S$$

if and only if $\widetilde{v} \in \Omega \cap S$ satisfies

$$G(\widetilde{v}) \geq G(v) \quad \textit{for all} \quad v \in \mathbb{R}^n \backslash \{0\}.$$

Proof. Since $g(v) \in C^\infty(\Omega)$ and $0 \notin \Omega$, it holds also $G(v) \in C^\infty(\Omega)$. If $\widetilde{v} \in \Omega \cap S$ is a maximum of $G(v)$, then $\alpha\widetilde{v}$ is a maximum as well, so the equivalence is obvious. □

Theorem 3.4. *The vectors $\widetilde{c} \in \mathbb{R}^k$ and $\widetilde{v} \in S \cap \Omega$ that solve the problem*

$$\max_{v \in S} \min_{c \in \mathbb{R}^n} f(c, v)$$

satisfy

$$\nabla_c F(\widetilde{c}, \widetilde{v}) = 0, \qquad \nabla_v F(\widetilde{c}, \widetilde{v}) = 0, \tag{3.4}$$

i.e., $(\widetilde{c}, \widetilde{v})$ is a stationary point of the function $F(c, v)$.

Proof. Obviously, for any $v \in \Omega$,

$$F(c_*(v), v) = \frac{f(c_*(v), v)}{v^T v} \leq \frac{f(c, v)}{v^T v} = F(c, v) \quad \text{for all } c \in \mathbb{R}^k,$$

i.e., $c_*(v)$ also minimizes the function $F(c, v)$. Hence,

$$\nabla_c F(c_*(v), v) = 0, \qquad v \in \Omega.$$

We know that $g(v)$ attains its maximum on S at some point $\widetilde{v} \in \Omega \cap S$. Therefore, $G(v)$ attains its maximum also at $\widetilde{v}$. Since $G(v) \in C^\infty(\Omega)$, it has to hold that

$$\nabla G(\widetilde{v}) = 0.$$

Denoting $\widetilde{c} = c_*(\widetilde{v})$ and writing the function $G(v)$ as $G(v) = F(c_*(v), v)$, we get

$$\nabla G(\widetilde{v}) = 0 = \nabla_v c_*(\widetilde{v}) \nabla_c F(\widetilde{c}, \widetilde{v}) + \nabla_v F(\widetilde{c}, \widetilde{v}), \tag{3.5}$$

where $\nabla_v c_*(\widetilde{v})$ is the $n \times k$ Jacobian matrix of the function $c_*(v) : \mathbb{R}^n \to \mathbb{R}^k$ at the point $\widetilde{v}$. Here we used the standard chain rule for multivariate functions. Since $\widetilde{v} \in \Omega \cap S$, we know that $\nabla_c F(\widetilde{c}, \widetilde{v}) = 0$, and, therefore, using (3.5), $\nabla_v F(\widetilde{c}, \widetilde{v}) = 0$. □

THEOREM 3.5. *If* $(\widetilde{c}, \widetilde{v})$ *is a solution of the problem* (3.3)*, then* $\widetilde{v}$ *is a right singular vector of the matrix* $p(A; \widetilde{c})$.

Proof. Since $(\widetilde{c}, \widetilde{v})$ solves the problem (3.3), we have $0 = \nabla_v F(\widetilde{c}, \widetilde{v})$. Writing $F(c, v)$ as a Rayleigh quotient,

$$F(c, v) = \frac{v^T p(A; c)^T p(A; c) v}{v^T v},$$

we ask when $\nabla_v F(c, v) = 0$; for more details see [9, pp. 114–115]. By differentiating $F(c, v)$ with respect to v, we get

$$0 = \frac{2p(A; c)^T p(A; c) v \, \|v\|^2 - 2[v^T p(A; c)^T p(A; c) v] \, v}{(v^T v)^2}$$

and the condition $0 = \nabla_v F(\widetilde{c}, \widetilde{v})$ is equivalent to

$$p(A; \widetilde{c})^T p(A; \widetilde{c}) \widetilde{v} = F(\widetilde{c}, \widetilde{v}) \, \widetilde{v}.$$

In other words, $\widetilde{v}$ is a right singular vector of $p(A; \widetilde{c})$ and $\sigma = \sqrt{F(\widetilde{c}, \widetilde{v})}$ is the corresponding singular value. □

THEOREM 3.6. *A point* $(\widetilde{c}, \widetilde{v}) \in \mathbb{R}^k \times S$ *that solves the problem* (3.3) *is a stationary point of* $F(c, v)$ *in which the maximal value of* $F(c, v)$ *is attained.*

Proof. Using Theorem 3.4 we know that any solution $(\widetilde{c}, \widetilde{v}) \in \mathbb{R}^k \times S$ of (3.3) is a stationary point of $F(c, v)$. On the other hand, if $(\hat{c}, \hat{v}) \in \mathbb{R}^k \times S$ satisfies

$$\nabla_v F(\hat{c}, \hat{v}) = 0, \qquad \nabla_c F(\hat{c}, \hat{v}) = 0,$$

then $p(A; \hat{c})$ is the GMRES polynomial that corresponds to $\hat{v}$ and

$$F(\hat{c}, \hat{v}) = \|p(A; \hat{c}) \hat{v}\|^2 \le \|p(A; \widetilde{c}) \widetilde{v}\|^2 = F(\widetilde{c}, \widetilde{v}).$$

Hence, $(\widetilde{c}, \widetilde{v})$ is a stationary point of $F(c, v)$ in which the maximal value of $F(c, v)$ is attained. □

As a consequence of previous results we can formulate the following corollary.

COROLLARY 3.7. *With the same assumptions and notation as in Theorem* 2.3*, it holds that* $p_k = q_k$.

Proof. Using Theorems 3.5 and 3.6 we know that

$$\psi_k^2(A) b = p_k(A^T) p_k(A) b, \tag{3.6}$$

i.e., that b is a right singular vector of the matrix $p_k(A)$ that corresponds to the maximal value of $F(\widetilde{c}, \widetilde{v})$, i.e., to $\psi_k^2(A)$. From (2.6) we also know that

$$\psi_k^2(A) \, b = q_k(A^T) p_k(A) \, b, \tag{3.7}$$

where q_k is the GMRES polynomial that corresponds to A^T and the initial vector r_k. Comparing (3.6) and (3.7), and using the uniqueness of the GMRES polynomial q_k, it follows that $p_k = q_k$. □

4. Nonuniqueness of worst-case GMRES polynomials. In this section we prove that a worst-case GMRES polynomial may not be uniquely determined, and we give a numerical example for the occurrence of a nonunique case. Our results are based on Toh's parametrized family of (nonsingular) matrices

$$A = A(\omega, \varepsilon) = \begin{bmatrix} 1 & \varepsilon & & \\ & -1 & \frac{\omega}{\varepsilon} & \\ & & 1 & \varepsilon \\ & & & -1 \end{bmatrix} \in \mathbb{R}^{4\times 4}, \qquad 0 < \omega < 2, \quad 0 < \varepsilon. \tag{4.1}$$

Toh used these matrices in [16] to show that $\psi_3(A)/\varphi_3(A) \to 0$ for $\epsilon \to 0$ and each $\omega \in (0, 2)$ [16, Theorem 2.3]. In other words, he proved that the ratio of the worst-case and ideal GMRES approximations can be arbitrarily small.

THEOREM 4.1. *Let A be as in* (4.1)*. If $p_k(z)$ is a worst-case GMRES polynomial for A and k, then $p_k(-z)$ is also a worst-case GMRES polynomial for A and k.*

In particular, $p_3(z) \neq p_3(-z)$, so the worst-case GMRES polynomial for A and $k = 3$ is not uniquely determined.

Proof. Let b be any worst-case initial vector for A and k, and consider the orthogonal similarity transformation

$$A = -QA^TQ^T, \qquad Q = \begin{bmatrix} & & & 1 \\ & & -1 & \\ & 1 & & \\ -1 & & & \end{bmatrix}.$$

Then

$$p_k(A)b = Qp_k(-A^T)Q^Tb \quad \text{and} \quad \psi_k(A) = \|p_k(A)b\| = \|p_k(-A^T)w\| = \psi_k(A^T),$$

where $w = Q^Tb$. In other words, $p_k(-z)$ is a worst-case GMRES polynomial for A^T and k. Using Corollary 3.7, it is also a worst-case GMRES polynomial for A and k.

Let $p_3(z) \in \pi_3$ be any worst-case GMRES polynomial for A and $k = 3$. To show that $p_3(-z) \neq p_3(z)$ it suffices to show that $p_3(z)$ contains odd powers of z, i.e., that

$$p_3(z) \neq 1 - \beta z^2 \quad \text{for any } \beta \in \mathbb{R}. \tag{4.2}$$

Define the matrix

$$B \equiv \begin{bmatrix} 1 & 0 & \omega & 0 \\ & 1 & 0 & \omega \\ & & 1 & 0 \\ & & & 1 \end{bmatrix} = A^2.$$

From [16, Theorem 2.1] we know that the (uniquely determined) ideal GMRES polynomial for A and $k = 3$ is of the form

$$p_*(z) = 1 + (\alpha - 1)z^2, \qquad \alpha = \frac{2\omega^2}{4 + \omega^2}. \tag{4.3}$$

Therefore,

$$\min_{p\in\pi_3} \|p(A)\| = \min_{p\in\pi_1} \max_{\|v\|=1} \|p(B)v\| = \max_{\|v\|=1} \min_{p\in\pi_1} \|p(B)v\|,$$

where the last equality follows from the fact that the ideal and worst-case GMRES approximations are equal for $k = 1$ [8, 5]. If a worst-case polynomial for A and $k = 3$ is of the form $1 - \beta z^2$ for some β, then

$$\psi_3(A) = \max_{\|v\|=1} \min_{p\in\pi_3} \|p(A)v\| = \max_{\|v\|=1} \min_{p\in\pi_1} \|p(B)v\| = \min_{p\in\pi_3} \|p(A)\| = \varphi_3(A).$$

This, however, contradicts the main result by Toh that $\psi_3(A) < \varphi_3(A)$; see [16, Theorem 2.2]. ◻

To compute examples of worst-case GMRES polynomials for the Toh matrix (4.1) numerically we chose $\varepsilon = 0.1$ and $\omega = 1$, and we used the function `fminsearch` from MATLAB's Optimization Toolbox. We computed the value

$$\psi_3(A) = 0.4579$$

(we present the numerical results only to 4 digits) with the corresponding third worst-case initial vector

$$b = [-0.6376, 0.0471, 0.2188, 0.7371]^T$$

and the worst-case GMRES polynomial

$$p_3(z) = -0.025z^3 - 0.895z^2 + 0.243z + 1 = \frac{-1}{39.9}(z - 1.181)(z + 0.939)(z + 35.96).$$

One can numerically check that b is the right singular vector of $p_3(A)$ that corresponds to the second maximal singular value of $p_3(A)$. From Theorem 4.1 we know that $q_3(z) \equiv p_3(-z)$ is also a third worst-case GMRES polynomial. One can now find the corresponding worst-case initial vector leading to the polynomial q_3 using the singular value decomposition (SVD)

$$p_3(A) = USV^T,$$

where the singular values are ordered nonincreasingly on the diagonal of S. We know (by numerical observation) that b is the second column of V. We now compute the SVD of $q_3(A)$ and define the corresponding initial vector as the right singular vector that corresponds to the second maximal singular value of $q_3(A)$. It holds that

$$p_3(A^T) = p_3(A)^T = VSU^T.$$

Since $A^T = -QAQ^T$, we get $Qp_3(-A)Q^T = VSU^T$, or, equivalently,

$$q_3(A) = (Q^TV)S(Q^TU)^T.$$

So, the columns of the matrix Q^TU are right singular vectors of $q_3(A)$ and the vector Q^Tu_2, where u_2 is the second column of U, is the worst-case initial vector that gives the worst-case GMRES polynomial $q_3(z) = p_3(-z)$.

5. Ideal versus worst-case GMRES.

As mentioned above, Toh [16] as well as Faber et al. [2] have shown that worst-case GMRES and ideal GMRES are different approximation problems in the sense that there exist matrices A and iteration steps k for which $\psi_k(A) < \varphi_k(A)$. In this section we further study these two approximation problems. We start with a geometrical characterization related to the function $f(c, v)$ from (3.2).

THEOREM 5.1. *Let $A \in \mathbb{R}^{n\times n}$ be a nonsingular matrix and let $1 \le k \le d(A) - 1$. The kth ideal and worst-case GMRES approximations are equal, i.e.,*

$$\max_{v\in S}\min_{c\in\mathbb{R}^k} f(c,v) = \min_{c\in\mathbb{R}^k}\max_{v\in S} f(c,v), \tag{5.1}$$

if and only if $f(c,v)$ has a saddle point in $\mathbb{R}^k \times S$.

Proof. If $f(c,v)$ has a saddle point in $\mathbb{R}^k \times S$, then there exist vectors $\widetilde{c} \in \mathbb{R}^k$ and $\widetilde{v} \in S$ such that

$$f(\widetilde{c},v) \le f(\widetilde{c},\widetilde{v}) \le f(c,\widetilde{v}) \qquad \text{for all}\, c \in \mathbb{R}^k \text{ and all}\, v \in S.$$

The condition $f(\widetilde{c},v) \le f(\widetilde{c},\widetilde{v})$ for all $v \in S$ implies that $\widetilde{v}$ is a maximal right singular vector of the matrix $p(A;\widetilde{c})$. If $f(\widetilde{c},\widetilde{v}) \le f(c,\widetilde{v})$ for all $c \in \mathbb{R}^k$, then $p(z;\widetilde{c})$ is the GMRES polynomial that corresponds to the initial vector $\widetilde{v}$. In other words, if $f(c,v)$ has a saddle point in $\mathbb{R}^k \times S$, then there exist a polynomial $p(z;\widetilde{c})$ and a unit norm vector $\widetilde{v}$ such that $\widetilde{v}$ is a maximal right singular vector of $p(A;\widetilde{c})$ and

$$p(A;\widetilde{c})\widetilde{v} \perp A\mathcal{K}_k(A,\widetilde{v}).$$

Using [15, Lemma 2.4], the kth ideal and worst-case GMRES approximations are then equal.

On the other hand, if the condition (5.1) is satisfied, then $f(c,v)$ has a saddle point in $\mathbb{R}^k \times S$. □

In other words, the kth ideal and worst-case GMRES approximations are equal if and only if the points $(\widetilde{c},\widetilde{v}) \in \mathbb{R}^k \times S$ that solve the worst-case GMRES problem are also the saddle points of $f(c,v)$ in $\mathbb{R}^k \times S$.

We next extend the original construction of Toh [16] to obtain some further numerical examples in which $\psi_k(A) < \varphi_k(A)$. Note that the Toh matrix (4.1) is not diagonalizable. In particular, for $\omega = 1$ we have $A = X\widetilde{J}X^{-1}$, where

$$\widetilde{J} = \begin{bmatrix} 1 & 1 & & \\ & 1 & & \\ & & -1 & 1 \\ & & & -1 \end{bmatrix}, \qquad X = \begin{bmatrix} \epsilon & \epsilon & \epsilon & -\epsilon \\ -2 & -1 & 0 & 1 \\ 0 & -2\epsilon & 0 & 2\epsilon \\ 0 & 4 & 0 & 0 \end{bmatrix}.$$

One can ask whether the phenomenon $\psi_k(A) < \varphi_k(A)$ can also appear for diagonalizable matrices. The answer is yes, since both $\psi_k(A)$ and $\varphi_k(A)$ are continuous functions on the open set of nonsingular matrices; see [2, Theorems 2.5 and 2.6]. Hence one can slightly perturb the diagonal of the Toh matrix (4.1) in order to obtain a diagonalizable matrix $\widetilde{A}$ for which $\psi_k(\widetilde{A}) < \varphi_k(\widetilde{A})$.

For $\omega = 1$, the Toh matrix is an upper bidiagonal matrix with the alternating diagonal entries 1 and -1, and the alternating superdiagonal entries ϵ and ϵ^{-1}. One can consider such a matrix for any $n \ge 4$, i.e.,

$$A = \begin{bmatrix} 1 & \varepsilon & & & & \\ & -1 & \varepsilon^{-1} & & & \\ & & 1 & \varepsilon & & \\ & & & \ddots & \ddots & \\ & & & & \ddots & \varepsilon^{\pm 1} \\ & & & & & \pm 1 \end{bmatrix} \in \mathbb{R}^{n\times n},$$

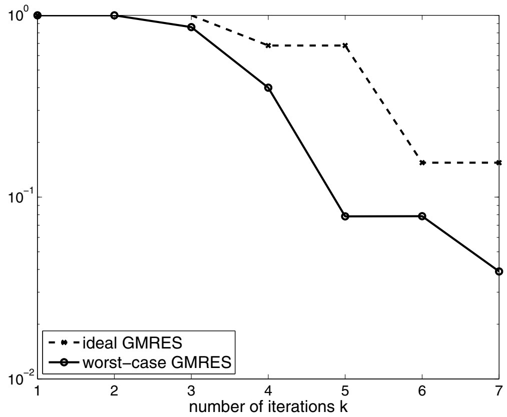


FIG. 5.1. *Ideal and worst-case GMRES can differ from step* 3 *up to step* $2n-1$.

and look at the values of $\psi_k(A)$ and $\varphi_k(A)$. If n is even, we found numerically that $\psi_k(A) = \varphi_k(A)$ for $k \neq n-1$ and $\psi_{n-1}(A) < \varphi_{n-1}(A)$. If n is odd, then our numerical experiments showed that $\psi_k(A) = \varphi_k(A)$ for $k \neq n-2$ and $\psi_{n-2}(A) < \varphi_{n-2}(A)$. Hence for all such matrices worst-case and ideal GMRES differ from each other for exactly one k.

Inspired by the Toh matrix, we define the $n \times n$ matrices (for any $n \geq 2$)

$$J_{\lambda,\varepsilon} \equiv \begin{bmatrix} \lambda & \varepsilon & & \\ & \ddots & \ddots & \\ & & \ddots & \varepsilon \\ & & & \lambda \end{bmatrix}, \qquad E_\varepsilon \equiv \begin{bmatrix} 0 & 0 & \dots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \dots & 0 \\ \varepsilon^{-1} & 0 & \dots & 0 \end{bmatrix}$$

and use them to construct the matrix

$$A = \begin{bmatrix} J_{1,\varepsilon} & \omega E_\varepsilon \\ & J_{-1,\varepsilon} \end{bmatrix} \in \mathbb{R}^{2n \times 2n}, \qquad \omega > 0.$$

One can numerically observe that here $\psi_k(A) < \varphi_k(A)$ for all steps $k = 3, \dots, 2n-1$. As an example, we plot in Figure 5.1 the ideal and worst-case GMRES convergence curves for $n = 4$, i.e., A is an 8×8 matrix, $\omega = 4$, and $\varepsilon = 0.1$. Varying the parameter ω will influence the difference between the worst-case and ideal GMRES values in these examples. Decreasing ω leads to a smaller difference, and increasing ω leads to a larger difference for large k, while the two values need not differ for some small k.

6. Ideal and worst-case GMRES for complex vectors or polynomials. We now ask whether the values of the min-max approximation (1.2) and the max-min approximation (1.3) for a matrix $A \in \mathbb{R}^{n \times n}$ can change if we allow the maximization over complex vectors and/or the minimization over complex polynomials. We will give a complete answer to this question in Theorems 6.1 and 6.3 below. In short, for the min-max approximation related to ideal GMRES the underlying fields of minimization and maximization do not matter, while for the max-min approximation related to worst-case GMRES different fields may in some cases indeed lead to different values. These results again indicate the different nature of the two approximation problems, and they complement (and in some sense complete) previous results in the literature dealing with the same question; see, in particular, [2, section 2], [8, section 3], and [20, section 4].

Let us define

$$\varphi_{k,\mathbb{K},\mathbb{F}}(A) \equiv \min_{p\in\pi_{k,\mathbb{K}}} \max_{\substack{b\in\mathbb{F}^n\\ \|b\|=1}} \|p(A)b\|, \qquad \psi_{k,\mathbb{K},\mathbb{F}}(A) \equiv \max_{\substack{b\in\mathbb{F}^n\\ \|b\|=1}} \min_{p\in\pi_{k,\mathbb{K}}} \|p(A)b\|,$$

where $\mathbb{K}$ and $\mathbb{F}$ are either the real or the complex numbers. Hence, the previously used $\varphi_k(A)$, $\psi_k(A)$, and π_k are now denoted by $\varphi_{k,\mathbb{R},\mathbb{R}}(A)$, $\psi_{k,\mathbb{R},\mathbb{R}}(A)$, and $\pi_{k,\mathbb{R}}$, respectively. We first analyze the case of $\varphi_{k,\mathbb{K},\mathbb{F}}(A)$.

THEOREM 6.1. *For a nonsingular matrix $A \in \mathbb{R}^{n\times n}$ and $1 \le k \le d(A)-1$,*

$$\varphi_{k,\mathbb{R},\mathbb{R}}(A) = \varphi_{k,\mathbb{C},\mathbb{R}}(A) = \varphi_{k,\mathbb{R},\mathbb{C}}(A) = \varphi_{k,\mathbb{C},\mathbb{C}}(A).$$

Proof. Since

$$\max_{\substack{b\in\mathbb{R}^n\\ \|b\|=1}} \|Bv\| = \|B\| = \max_{\substack{b\in\mathbb{C}^n\\ \|b\|=1}} \|Bv\|$$

holds for any real matrix $B \in \mathbb{R}^{n\times n}$, we have $\varphi_{k,\mathbb{R},\mathbb{R}}(A) = \varphi_{k,\mathbb{R},\mathbb{C}}(A)$.

Next, from $\mathbb{R} \subset \mathbb{C}$ we get immediately $\varphi_{k,\mathbb{C},\mathbb{R}}(A) \le \varphi_{k,\mathbb{R},\mathbb{R}}(A)$. On the other hand, writing $p \in \pi_{k,\mathbb{C}}$ in the form $p = p_r + \mathbf{i}\,p_i$, where $p_r \in \pi_{k,\mathbb{R}}$ and p_i is a real polynomial of degree at most k such that $p_i(0) = 0$, we get

$$\begin{aligned}\varphi^2_{k,\mathbb{C},\mathbb{R}}(A) = \min_{p\in\pi_{k,\mathbb{C}}} \max_{\substack{b\in\mathbb{R}^n\\ \|b\|=1}} \|p(A)b\|^2 &= \min_{p\in\pi_{k,\mathbb{C}}} \max_{\substack{b\in\mathbb{R}^n\\ \|b\|=1}} \left(\|p_r(A)b\|^2 + \|p_i(A)b\|^2\right)\\ &\ge \min_{p_r\in\pi_{k,\mathbb{R}}} \max_{\substack{b\in\mathbb{R}^n\\ \|b\|=1}} \|p_r(A)b\|^2 = \varphi^2_{k,\mathbb{R},\mathbb{R}}(A),\end{aligned}$$

so that $\varphi_{k,\mathbb{C},\mathbb{R}}(A) = \varphi_{k,\mathbb{R},\mathbb{R}}(A)$. Finally, from [7, Theorem 3.1] we obtain $\varphi_{k,\mathbb{R},\mathbb{R}}(A) = \varphi_{k,\mathbb{C},\mathbb{C}}(A)$. □

Since the value of $\varphi_{k,\mathbb{K},\mathbb{F}}(A)$ does not change when choosing for $\mathbb{K}$ and $\mathbb{F}$ real or complex numbers, we will again use the simple notation $\varphi_k(A)$ in the following text. The situation for the quantities corresponding to the worst-case GMRES approximation is more complicated. Our proof of this fact uses the following lemma.

LEMMA 6.2. *If $A = A(\omega,\varepsilon)$ is the Toh matrix defined in* (4.1) *and*

$$B \equiv \begin{bmatrix} A & 0\\ 0 & A\end{bmatrix}, \tag{6.1}$$

then $\psi_{3,\mathbb{R},\mathbb{R}}(B) = \varphi_3(A)$.

Proof. Using the structure of B it is easy to see that $\psi_{k,\mathbb{R},\mathbb{R}}(B) \le \varphi_k(A)$ for any k. To prove the equality, it suffices to find a real unit norm vector w with

$$\min_{p\in\pi_{3,\mathbb{R}}} \|p(B)w\| = \varphi_3(A) = \min_{p\in\pi_{3,\mathbb{R}}} \|p(A)\|. \tag{6.2}$$

The solution p_* of the ideal GMRES problem on the right-hand side of (6.2) is given by (4.3). Toh showed in [16, p. 32] that $p_*(A)$ has a twofold maximal singular value σ, and that the corresponding right and left singular vectors are given (up to a normalization) by

$$[v_1, v_2] = \begin{bmatrix} 0 & \omega\\ \omega & 0\\ 0 & -2\\ -2 & 0\end{bmatrix}, \qquad [u_1, u_2] = \begin{bmatrix} 0 & 2\\ 2 & 0\\ 0 & -\omega\\ -\omega & 0\end{bmatrix},$$

i.e., $\sigma u_1 = p_*(A)v_1$ and $\sigma u_2 = p_*(A)v_2$, where $\sigma = \|p_*(A)\|$.

Let us define

$$w \equiv \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} / \left\| \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \right\|, \qquad q(z) \equiv p_*(z).$$

Using

$$q(B) \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \sigma \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \quad \text{and} \quad \left\| \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \right\| = \left\| \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \right\|,$$

we see that $\|q(B)w\| = \sigma$. To prove (6.2) it is sufficient to show that q is the third GMRES polynomial for B and w, i.e., that q satisfies $q(B)w \perp B^j w$ for $j = 1, 2, 3$, or, equivalently,

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}^T \begin{bmatrix} A^j & 0 \\ 0 & A^j \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = u_1^T A^j v_1 + u_2^T A^j v_2 = 0, \quad j = 1, 2, 3.$$

Using linear algebra calculations we get $u_1^T A v_1 = -4\omega = -u_2^T A v_2$, and

$$0 = u_1^T A^2 v_1 = u_2^T A^2 v_2 = u_1^T A^3 v_1 = u_2^T A^3 v_2.$$

Therefore, we have found a unit norm initial vector w and the corresponding third GMRES polynomial q such that $\|q(B)w\| = \varphi_3(A)$. □

We next analyze the quantities $\psi_{k,\mathbb{K},\mathbb{F}}(A)$.

THEOREM 6.3. *For a nonsingular matrix $A \in \mathbb{R}^{n \times n}$ and $1 \leq k \leq d(A) - 1$,*

$$\psi_{k,\mathbb{R},\mathbb{R}}(A) = \psi_{k,\mathbb{C},\mathbb{R}}(A) \leq \psi_{k,\mathbb{C},\mathbb{C}}(A) \leq \psi_{k,\mathbb{R},\mathbb{C}}(A) \leq \varphi_k(A),$$

where the first and second inequalities can be strict.

Proof. For a real initial vector b, the corresponding GMRES polynomial is uniquely determined and real. This implies $\psi_{k,\mathbb{C},\mathbb{R}}(A) = \psi_{k,\mathbb{R},\mathbb{R}}(A)$. Next, from [7, Theorem 3.1] it follows that $\psi_{k,\mathbb{R},\mathbb{R}}(A) \leq \psi_{k,\mathbb{C},\mathbb{C}}(A)$. Finally, using $\mathbb{R} \subset \mathbb{C}$ we get $\psi_{k,\mathbb{C},\mathbb{C}}(A) \leq \psi_{k,\mathbb{R},\mathbb{C}}(A)$. It remains to show that the first and second inequalities can be strict, and that $\psi_{k,\mathbb{R},\mathbb{C}}(A) \leq \varphi_k(A)$.

For the first inequality, as shown in [20, section 4], there exist real matrices A and certain complex (unit norm) initial vectors b for which $\min_{p \in \pi_{k,\mathbb{C}}} \|p(A)b\| = 1$ for $k = 1, \ldots, n-1$ (complete stagnation), while such complete stagnation does not occur for any real (unit norm) initial vector. Therefore, there are matrices for which $\psi_{k,\mathbb{C},\mathbb{R}}(A) < \psi_{k,\mathbb{C},\mathbb{C}}(A)$.

To show that the second inequality can be strict, we note that for any $A \in \mathbb{R}^{n \times n}$, the corresponding matrix $B \in \mathbb{R}^{2n \times 2n}$ of the form (6.1), and $1 \leq k \leq d(A) - 1$,

$$\begin{aligned} \psi_{k,\mathbb{R},\mathbb{C}}^2(A) &= \max_{\substack{b \in \mathbb{C}^n \\ \|b\|=1}} \min_{p \in \pi_{k,\mathbb{R}}} \|p(A)b\|^2 = \max_{\substack{u,v \in \mathbb{R}^n \\ \|u\|^2 + \|v\|^2 = 1}} \min_{p \in \pi_{k,\mathbb{R}}} \|p(A)(u + \mathbf{i}\, v)\|^2 \\ &= \max_{\substack{u,v \in \mathbb{R}^n \\ \|u\|^2 + \|v\|^2 = 1}} \min_{p \in \pi_{k,\mathbb{R}}} (\|p(A)u\|^2 + \|p(A)v\|^2) \\ &= \max_{\substack{v \in \mathbb{R}^{2n} \\ \|v\|=1}} \min_{p \in \pi_{k,\mathbb{R}}} \|p(B)v\|^2 = \psi_{k,\mathbb{R},\mathbb{R}}^2(B). \end{aligned} \tag{6.3}$$

Now let A be the Toh matrix (4.1) and let $k = 3$. Toh showed in [16, Theorem 2.2] that for any unit norm $b \in \mathbb{C}^4$ and the corresponding third GMRES polynomial $p_b \in \pi_{3,\mathbb{C}}$,

$$\|p_b(A)b\| < \varphi_3(A).$$

Hence $\psi_{3,\mathbb{C},\mathbb{C}}(A) < \varphi_3(A)$. Lemma 6.2 and (6.3) imply $\varphi_3(A) = \psi_{3,\mathbb{R},\mathbb{C}}(A)$ for the Toh matrix, and, therefore, the second inequality can be strict.

Finally, since $\|p(A)\| = \|p(B)\|$ for any polynomial p, we get $\varphi_3(B) = \varphi_3(A)$, and, using (6.3), $\psi_{3,\mathbb{R},\mathbb{C}}(A) = \psi_{3,\mathbb{R},\mathbb{R}}(B) \leq \varphi_3(B) = \varphi_3(A)$. □

We do not know whether the first and second inequalities in Theorem 6.3 can be strict simultaneously, i.e., can both be strict for the same A and k. Concerning the last inequality in Theorem 6.3, we are in fact able to prove that $\psi_{k,\mathbb{R},\mathbb{C}}(A) = \varphi_k(A)$. Since the techniques used in this proof are beyond the scope of this paper, we will publish it elsewhere.

Our proof concerning the strictness of the first inequality in the previous theorem relied on a numerical example given in [20, section 4]. We will now give an alternative construction based on the nonuniqueness of the worst-case GMRES polynomial, which will lead to an example with $\psi_{k,\mathbb{R},\mathbb{R}}(A) < \psi_{k,\mathbb{R},\mathbb{C}}(A)$.

Suppose that A is a real matrix for which in a certain step k two *different* worst-case polynomials $p_b \in \pi_{k,\mathbb{R}}$ and $p_c \in \pi_{k,\mathbb{R}}$ with corresponding real unit norm initial vectors b and c exist, so that

$$\psi_{k,\mathbb{R},\mathbb{R}}(A) = \|p_b(A)b\| = \|p_c(A)c\|.$$

Note that since p_b and p_c are the uniquely determined GMRES polynomials that solve the problem (1.1) for the corresponding real initial vectors, it holds that

$$\|p_b(A)b\| < \|p(A)b\|, \qquad \|p_c(A)c\| < \|p(A)c\| \tag{6.4}$$

for any polynomial $p \in \pi_{k,\mathbb{C}} \setminus \{p_b, p_c\}$.

Writing any complex vector $w \in \mathbb{C}^n$ in the form $w = (\cos\theta)\, u + \mathbf{i}\,(\sin\theta)\, v$, with $u, v \in \mathbb{R}^n$, $\|u\| = \|v\| = 1$, we get

$$\begin{aligned}
\psi^2_{k,\mathbb{R},\mathbb{C}}(A) &= \max_{\substack{w\in\mathbb{C}^n\\ \|w\|=1}} \min_{p\in\pi_{k,\mathbb{R}}} \|p(A)b\|^2 \\
&= \max_{\substack{\theta\in\mathbb{R},u,v\in\mathbb{R}^n\\ \|u\|=\|v\|=1}} \min_{p\in\pi_{k,\mathbb{R}}} \left(\cos^2\theta\,\|p(A)u\|^2 + \sin^2\theta\,\|p(A)v\|^2\right) \\
&\geq \max_{\theta\in\mathbb{R}} \min_{p\in\pi_{k,\mathbb{R}}} \left(\cos^2\theta\|p(A)b\|^2 + \sin^2\theta\|p(A)c\|^2\right) \\
&> (\cos^2\theta)\,\psi^2_{k,\mathbb{R},\mathbb{R}}(A) + (\sin^2\theta)\,\psi^2_{k,\mathbb{R},\mathbb{R}}(A) = \psi^2_{k,\mathbb{R},\mathbb{R}}(A),
\end{aligned}$$

where the strict inequality follows from (6.4) and from the fact that $\|p(A)b\|^2$ and $\|p(A)c\|^2$ do not attain their minima for the same polynomial.

To demonstrate the strict inequality $\psi_{k,\mathbb{R},\mathbb{R}}(A) < \psi_{k,\mathbb{R},\mathbb{C}}(A)$ numerically we use the Toh matrix (4.1) with $\varepsilon = 0.1$ and $\omega = 1$, and $k = 3$. Let b and c be the corresponding two different worst-case initial vectors introduced in section 4. We vary θ from 0 to π and compute the quantities

$$\min_{p\in\pi_{3,\mathbb{R}}} \left(\cos^2\theta\,\|p(A)b\|^2 + \sin^2\theta\,\|p(A)c\|^2\right) = \min_{p\in\pi_{3,\mathbb{R}}} \|p(B)g_\theta\|^2, \tag{6.5}$$

where

$$B = \begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix} \quad \text{and} \quad g_\theta = \begin{bmatrix} (\cos\theta)b \\ (\sin\theta)c \end{bmatrix}.$$

In Figure 6.1 we can see clearly that for $\theta \notin \{0, \pi/2, \pi\}$ the value of (6.5) is strictly larger than $\psi_{3,\mathbb{R},\mathbb{R}}(A) = 0.4579$ (dashed line). Numerical computations predict that $\psi_{3,\mathbb{R},\mathbb{R}}(A) = \psi_{3,\mathbb{C},\mathbb{C}}(A)$ for the Toh matrix. Finally, Lemma 6.2 and (6.3) imply $\psi_{3,\mathbb{R},\mathbb{C}}(A) = \psi_{3,\mathbb{R},\mathbb{R}}(B) = \varphi_3(A)$ (dash-dotted line).

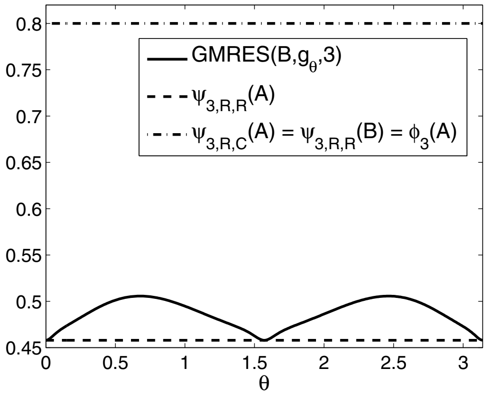


FIG. 6.1. *Illustration of the value of* (6.5) *and different quantities from Theorem* 6.3 *for the Toh matrix* $A(1.0, 0.1)$ *in* (4.1) *and* $k = 3$.

Acknowledgments. We thank Michael Eiermann and two anonymous referees for very helpful corrections and suggestions.

REFERENCES

[1] M. Afanasjew, M. Eiermann, O. G. Ernst, and S. Güttel, *A generalization of the steepest descent method for matrix functions*, Electron. Trans. Numer. Anal., 28 (2007/2008), pp. 206–222.

[2] V. Faber, W. Joubert, E. Knill, and T. Manteuffel, *Minimal residual method stronger than polynomial preconditioning*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 707–729.

[3] G. E. Forsythe, *On the asymptotic directions of the s-dimensional optimum gradient method*, Numer. Math., 11 (1968), pp. 57–76.

[4] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, Frontiers Appl. Math. 17, SIAM, Philadelphia, PA, 1997.

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

[6] A. Greenbaum and L. N. Trefethen, *GMRES/CR and Arnoldi/Lanczos as matrix approximation problems*, SIAM J. Sci. Comput., 15 (1994), pp. 359–368.

[7] W. Joubert, *On the convergence behavior of the restarted GMRES algorithm for solving nonsymmetric linear systems*, Numer. Linear Algebra Appl., 1 (1994), pp. 427–447.

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

[9] P. D. Lax, *Linear Algebra and Its Applications*, 2nd ed., Pure Appl. Math. (Hoboken), Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2007.

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

[11] J. Liesen and Z. Strakoš, *Krylov Subspace Methods. Principles and Analysis*, Oxford University Press, Oxford, 2013.

[12] J. Liesen and P. Tichý, *On best approximations of polynomials in matrices in the matrix 2-norm*, SIAM J. Matrix Anal. Appl., 31 (2009), pp. 853–863.

[13] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed., SIAM, Philadelphia, PA, 2003.

[14] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[15] P. Tichý, J. Liesen, and V. Faber, *On worst-case GMRES, ideal GMRES, and the polynomial numerical hull of a Jordan block*, Electron. Trans. Numer. Anal., 26 (2007), pp. 453–473.

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

[17] K.-C. Toh, M. J. Todd, and R. H. Tütüncü, *On the implementation and usage of SDPT*3*—a Matlab software package for semidefinite-quadratic-linear programming, version* 4.0, in Handbook on Semidefinite, Conic and Polynomial Optimization, Internat. Ser. Oper. Res. Management Sci. 166, Springer, New York, 2012, pp. 715–754.

[18] K.-C. Toh and L. N. Trefethen, *The Chebyshev polynomials of a matrix*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 400–419.

[19] I. Zavorin, *Spectral Factorization of the Krylov Matrix and Convergence of GMRES*, Tech. Report CS-TR-4309, Computer Science Department, University of Maryland, 2001.

[20] I. Zavorin, D. P. O'Leary, and H. Elman, *Complete stagnation of GMRES*, Linear Algebra Appl., 367 (2003), pp. 165–183.

BIT Numerical Mathematics **44:** 79–98, 2004.
© 2004 *Kluwer Academic Publishers. Printed in the Netherlands.*

THE WORST-CASE GMRES FOR NORMAL MATRICES*

JÖRG LIESEN[1,**] and PETR TICHÝ[1,†]

[1] *Institute of Mathematics, Technical University of Berlin,*
Straße des 17. Juni 136, 10623 Berlin, Germany.
email: liesen@math.tu-berlin.de, tichy@math.tu-berlin.de

Abstract.

We study the convergence of GMRES for linear algebraic systems with normal matrices. In particular, we explore the standard bound based on a min-max approximation problem on the discrete set of the matrix eigenvalues. This bound is sharp, i.e. it is attainable by the GMRES residual norm. The question is how to evaluate or estimate the standard bound, and if it is possible to characterize the GMRES-related quantities for which this bound is attained (worst-case GMRES). In this paper we completely characterize the worst-case GMRES-related quantities in the next-to-last iteration step and evaluate the standard bound in terms of explicit polynomials involving the matrix eigenvalues. For a general iteration step, we develop a computable lower and upper bound on the standard bound. Our bounds allow us to study the worst-case GMRES residual norm as a function of the eigenvalue distribution. For hermitian matrices the lower bound is equal to the worst-case residual norm. In addition, numerical experiments show that the lower bound is generally very tight, and support our conjecture that it is to within a factor of $4/\pi$ of the actual worst-case residual norm. Since the worst-case residual norm in each step is to within a factor of the square root of the matrix size to what is considered an "average" residual norm, our results are of relevance beyond the worst case.

AMS subject classification (2000): 15A06, 15A09, 15A18, 65F10, 65F15, 65F20, 41A10.

Key words: GMRES, evaluation of convergence, ideal GMRES, normal matrices, min-max problem.

1 Introduction.

Convergence analysis of GMRES [14] has been an active area of research since the algorithm's introduction, and numerous papers have been devoted to this subject, see, e.g., [3, Chapter 3] and [10, Section 5.2] for surveys of results. When

* Received September 2003. Accepted December 2003. Communicated by Lars Eldén.

** The work of this author was supported by the Emmy Noether-Programm of the Deutsche Forschungsgemeinschaft.

† The work of this author was supported by the Emmy Noether-Programm of the Deutsche Forschungsgemeinschaft and by the Grant Agency of Academy of Sciences of the Czech Republic under grant No. KJB1030306.

the system matrix is normal, the earliest upper bound on the GMRES residual norms (henceforth called the "standard bound") represents a certain min-max approximation problem on the set of the matrix eigenvalues [14, Proposition 4]. Being independent of the initial residual, the standard bound is in fact a bound on the "worst-case" GMRES residual norms for the given system matrix. For normal matrices the standard bound has been shown to be sharp in the sense that for each GMRES iteration step there exists an initial residual (depending on the matrix and the iteration step) for which the bound is attained [4, 8]. In addition, for normal matrices the worst-case GMRES and the "average" GMRES behavior agree to within a factor of $n^{1/2}$ ($n =$ matrix size). By average behavior we here mean that GMRES is started with an initial residual having components in the matrix eigenvectors of approximately equal size (see Section 5 for details).

The sharpness of the standard bound and its closeness to the average case sometimes lead to the impression that the GMRES convergence behavior for normal matrices is fully understood. However, two major problems still remain open. First, the solution of the min-max approximation is unknown except for special cases, and its known estimates based on only a few properties of the matrix (such as the condition number) are often misleading. Second, in many practical applications the initial residual is not "average", and a systematic study of the consequences for the GMRES convergence needs yet to be performed.

This paper is devoted to the first of the two problems, as its solution appears to be a prerequisite for studying the second. To this end it is of great interest to characterize the min-max approximation problem in terms of easily comprehensible expressions involving the matrix eigenvalues as well as to determine the initial residuals for which the standard bound is attained. Several results in this direction have been previously obtained in the literature. For (real) symmetric positive definite matrices, the initial residuals leading to the worst-case residual norm are completely characterized in [2, Section 2]. The analysis in [2] is based on classical results of approximation theory. In particular, in case of a symmetric positive definite matrix, the polynomial that solves the approximation problem on the matrix eigenvalues, i.e. the one for which the standard bound is attained, is the well-known min-max polynomial on a discrete set of real points (here the matrix eigenvalues). The result of [2] is derived in the context of the conjugate gradient method and can be applied in the GMRES context and to all complex Hermitian matrices. A special case of this result (which in particular also assumes that the eigenvalues are real) is proved in [17] by solving a constrained optimization problem using Lagrange multipliers. The related paper [18] gives necessary and sufficient conditions on the eigenvalues of normal matrices so that there exists an initial residual for which GMRES stagnates throughout the iteration (called "complete stagnation" of GMRES). For any normal matrix satisfying these conditions the authors give formulas based on the matrix eigenvalues for all initial residuals that lead to complete stagnation [18, Theorem 3.1]. The complete stagnation obviously represents a special case of worst-case GMRES convergence behavior.

General bounds on the GMRES residual norms for normal matrices that depend on the matrix eigenvalues and the initial residual are derived in [7]. The main tool in this analysis is a factorization of the Krylov matrix. Using a similar starting point as in [7] we characterize the quantities in the next-to-last GMRES iteration step for normal matrices ($(n-1)$st step in case of an n by n matrix having n distinct eigenvalues) in terms of the initial residual and explicit polynomials involving the matrix eigenvalues. We give numerical illustrations of our analytic formulas that show how GMRES behaves for different eigenvalue distributions. Based on these results we completely characterize the worst-case GMRES quantities in the next-to-last iteration step. Then we analyze the worst-case GMRES residual norm in a general iteration step and develop a lower bound on this quantity. In case of hermitian matrices our results are the same as in [2, Section 2], but with a different proof. For the general (normal) case our results complement the existing literature. We prove that our lower bound is to within a factor of (at most) the order n to the actual worst-case residual norm. Furthermore, we conjecture that this bound is much more tight (namely to within a constant factor), and give supporting numerical evidence.

The paper is organized as follows. In Section 2 we develop the basic tools needed for our general analysis in Section 3. Numerical examples studying the closeness of the lower bound to the standard bound are given in Section 4, and a concluding discussion in Section 5 closes the paper.

Throughout the paper we assume exact arithmetic.

2 Basic concepts.

In this section we define and develop the basic tools needed for our analysis. Let a linear system

$$Ax = b, \tag{2.1}$$

with a *nonsingular and normal* matrix $A \in \mathbb{C}^{n\times n}$ and $b \in \mathbb{C}^n$ be given. Furthermore, let $A = Q\Lambda Q^H$ be the eigendecomposition of A, where $Q^H Q = I$, $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, and let $L = \{\lambda_1, \ldots, \lambda_n\}$ denote the set of all eigenvalues of A. To avoid unnecessary technical complications we will assume throughout this paper that *all eigenvalues of A are distinct.*

Suppose that we solve (2.1) with GMRES [14]. Starting from an initial guess x_0, this method computes the initial residual $r_0 = b - Ax_0$ and a sequence of iterates $x_1, x_2, \ldots$, so that the ith residual $r_i \equiv b - Ax_i$ satisfies

$$\|r_i\| = \|p_i(A)r_0\| = \min_{p\in\pi_i} \|p(A)r_0\|, \tag{2.2}$$

where π_i denotes the set of polynomials of degree at most i and with value one at the origin, and $\|\cdot\|$ denotes the 2-norm. We parameterize the initial residual r_0 by

$$r_0 = Q[\varrho_1, \ldots, \varrho_n]^{\mathrm{T}}, \tag{2.3}$$

so that

$$r_i = p_i(A)r_0 = Q[p_i(\lambda_1)\varrho_1, \ldots, p_i(\lambda_n)\varrho_n]^{\mathrm{T}}, \tag{2.4}$$

and (2.2) can be written in the form

$$\|r_i\| = \min_{p\in\pi_i} \left(\sum_{j=1}^{n} |p(\lambda_j)\varrho_j|^2 \right)^{1/2} . \tag{2.5}$$

It is well-known that for each GMRES iteration step i and each initial residual r_0 with at least $i+1$ nonzero coordinates ϱ_j, there exists a *unique* polynomial $p_i \in \pi_i$ that solves (2.5). This $p_i(\lambda)$ is called the ith GMRES polynomial.

Similar to [7, 17, 18], we start with a factorization of the Krylov matrix,

$$K_{i+1} \equiv [r_0, Ar_0, \ldots, A^i r_0] \tag{2.6}$$

for some i, $0 \le i \le n-1$. We denote $D \equiv \text{diag}(\varrho_1, \ldots, \varrho_n)$, and

$$V_{i+1} \equiv \begin{bmatrix} 1 & \lambda_1 & \cdots & \lambda_1^i \\ \vdots & \vdots & & \vdots \\ 1 & \lambda_n & \cdots & \lambda_n^i \end{bmatrix} . \tag{2.7}$$

Then $K_{i+1} = QDV_{i+1}$, and the Moore–Penrose generalized inverse of K_{i+1} is given by $K_{i+1}^+ = (DV_{i+1})^+ Q^H$. If $\text{rank}(D) \ge i+1$, then K_{i+1} has full column rank, and GMRES does not terminate before the step $i+1$. In this case, as shown in [7, Theorem 2.1], see also [11, Theorem 2.1], the ith GMRES residual satisfies

$$\begin{aligned} r_i &= \|r_i\|^2 (K_{i+1}^+)^H e_1 \\ &= \|r_i\|^2 Q \left[(DV_{i+1})^+\right]^H e_1, \end{aligned} \tag{2.8}$$

where $e_1 = [1, 0, \ldots, 0]^{\mathrm{T}}$. Comparing (2.4) and (2.9) shows that

$$p_i(\lambda_j)\varrho_j = \|r_i\|^2 \left[(DV_{i+1})^+\right]_{j1}^H , \quad j = 1, \ldots, n, \tag{2.9}$$

where $\left[(DV_{i+1})^+\right]_{j1}^H$ denotes the jth entry in the first column of $\left[(DV_{i+1})^+\right]^H$. Note that (2.9) gives the complete correspondence between the ith GMRES polynomial, the ith GMRES residual norm, the coordinates of r_0 in the eigenvectors of A, and the eigenvalues of A. To understand fully the behavior of GMRES for normal matrices it would be desirable to have a general formula for the entries in the first column of $\left[(DV_{i+1})^+\right]^H$. However, such a formula is for a general value of i unknown. In the following subsection we will study the special case $i = n-1$, in which (2.9)–(2.9) can be significantly simplified.

2.1 The $(n-1)$st GMRES step.

Without loss of generality we restrict our analysis in this subsection to vectors r_0 with nonzero coordinates ϱ_j, $j = 1, \ldots, n$. In case $d \ge 1$ coordinates ϱ_j are zero, the corresponding eigencomponents do not play any role for GMRES, and hence the formulas for $i = n-1$ derived below will hold for $i = n-d-1$. When

$\varrho_j \neq 0$ for all j, GMRES terminates, i.e. computes the solution x, exactly in step n, and its residual norms satisfy

$$\|r_0\| \geq \|r_1\| \geq \cdots \geq \|r_{n-1}\| > \|r_n\| = 0. \tag{2.10}$$

In the step $i = n-1$, the Vandermonde matrix V_n is square and invertible (all eigenvalues are distinct). Then $[(DV_n)^+]^H = D^{-H}V_n^{-H}$, and (2.9) is equivalent to

$$r_{n-1} = \|r_{n-1}\|^2 Q D^{-H} V_n^{-H} e_1. \tag{2.11}$$

Formulas for the entries of an inverse Vandermonde matrix are well known, see, e.g., [6, Chapter 21.1]. In general, the jth entry in the mth column of the matrix $V_n^{-\mathrm{T}}$ is the coefficient of the jth Lagrange polynomial,

$$l_j(\lambda) \equiv \prod_{\substack{k=1\\k\neq j}}^{n} \frac{\lambda_k - \lambda}{\lambda_k - \lambda_j}, \tag{2.12}$$

corresponding to λ^{m-1}, $m = 1, \ldots, n$. Hence the first column of V_n^{-H} is given by the complex conjugates of the constant terms of the $l_j(\lambda)$, i.e.

$$V_n^{-H} e_1 = [l_1(0), \ldots, l_n(0)]^H = \left[\prod_{\substack{k=1\\k\neq 1}}^{n} \frac{\lambda_k}{\lambda_k - \lambda_1}, \ldots, \prod_{\substack{k=1\\k\neq n}}^{n} \frac{\lambda_k}{\lambda_k - \lambda_n}\right]^H. \tag{2.13}$$

The following theorem explains how the $(n-1)$st GMRES residual and iteration polynomial depend on the eigenvalue distribution of A (represented by the values $l_j(0)$) and on the initial residual r_0 (represented by the coordinates ϱ_j).

Theorem 2.1. *Suppose that GMRES is applied to the system* (2.1) *with the normal matrix $A \in \mathbb{C}^{n\times n}$ having n distinct eigenvalues, and that r_0 is parameterized by* (2.3) *with $\varrho_j \neq 0$ for all j. Then the norm of the $(n-1)$st GMRES residual r_{n-1} satisfies*

$$\|r_{n-1}\| = \left(\sum_{j=1}^{n} \left|\frac{l_j(0)}{\varrho_j}\right|^2\right)^{-1/2}, \tag{2.14}$$

and the $(n-1)$st GMRES polynomial $p_{n-1}(\lambda)$ has the form

$$p_{n-1}(\lambda) = \|r_{n-1}\|^2 \sum_{j=1}^{n} \frac{\overline{l_j(0)}}{|\varrho_j|^2} l_j(\lambda). \tag{2.15}$$

Proof. Inserting (2.13) into (2.11) yields

$$r_{n-1} = \|r_{n-1}\|^2 Q \big[l_1(0)\varrho_1^{-1}, \ldots, l_n(0)\varrho_n^{-1}\big]^H, \tag{2.16}$$

from which (2.14) follows immediately by taking norms. Next, using the property $l_j(\lambda_k) = \delta_{jk}$, the polynomial $p_{n-1}(\lambda)$ can be written as a linear combination of the Lagrange polynomials,

$$p_{n-1}(\lambda) = \sum_{j=1}^{n} p_{n-1}(\lambda_j) l_j(\lambda). \tag{2.17}$$

Equating (2.4) for $i = n - 1$ with (2.16) shows that

$$p_{n-1}(\lambda_j) = \|r_{n-1}\|^2 \frac{\overline{l_j(0)}}{|\varrho_j|^2}, \quad j = 1, \ldots, n, \tag{2.18}$$

which, inserted into (2.17), shows (2.15). □

Theorem 2.1 gives formulas for the $(n-1)$st GMRES residual and polynomial in terms of the eigenvalues of A and the coordinates of r_0 in the eigenvectors of A. The influences of both quantities are well separated in (2.14) and (2.15), so that these formulas answer all questions about the $(n-1)$st step of GMRES applied to normal matrices.

Note that the relation (2.14) implies the upper bound

$$\|r_{n-1}\| \leq \min_{1 \leq j \leq n} \left| \frac{\varrho_j}{l_j(0)} \right|. \tag{2.19}$$

The same upper bound follows from [7, Theorem 4.1] with $i = n - 1$.

Example 2.1. For numerical illustration we compute the values $|l_j(0)|$ for four different real eigenvalue distributions. Each dot in Figure 2.1 represents a data point $(\lambda_j, |l_j(0)|)$.

For the top left figure we use uniformly distributed eigenvalues in the interval $[1/20, 1]$, i.e. $\lambda_j = j/20$, for $j = 1, \ldots, 20$. We see that $|l_{10}(0)| \approx 10^5$ is the largest of the values $|l_j(0)|$. Then (2.19) implies that for any normal matrix having such eigenvalues, the GMRES residual norm in the next-to-last step will be of order 10^{-5} or smaller (note that $0 < |\varrho_j| < 1$ by assumption).

For the top right figure we use the eigenvalues of the 20 by 20 prolate matrix generated by the MATLAB command `A=gallery('prolate',20)`. Prolate matrices arise in signal processing. They are symmetric, extremely ill conditioned (here: $\lambda_1 \approx 1.76 * 10^{-14}$, $\lambda_{20} = 1 - \lambda_1$, condition number $\approx 5.69 * 10^{13}$), and their eigenvalues form two clusters that are symmetric about a certain point (here: symmetric about 0.5); see [16] for more information. In our example the cluster close to zero causes severe trouble for GMRES. None of the values $|l_j(0)|$ is larger than one, which typically (i.e., unless a very peculiar distribution of the coefficients ϱ_j is constructed) will lead to almost complete stagnation until the very last step, cf. (2.14). This represents a counterexample for the frequent assertion that in case of k (here: $k = 2$) eigenvalue clusters GMRES will essentially need only k steps for a significant reduction of the residual norm. In fact,

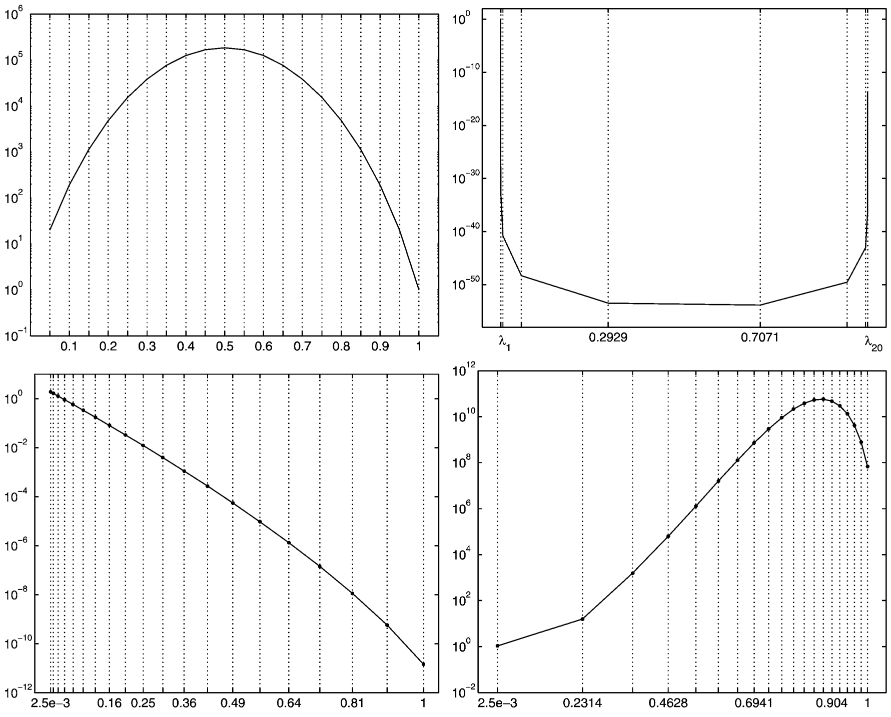


Figure 2.1: The values $|l_j(0)|$ for example eigenvalue distributions.

the location of the clusters relative to the origin and relative to each other is of great importance for the GMRES performance. This is also demonstrated in the two further examples.

The bottom left and bottom right figures show the values $|l_j(0)|$ for the eigenvalue distributions

$$\begin{aligned} \lambda_j^{(0)} &= j^2/400, \quad j = 1, \ldots, 20, \quad \text{and} \\ \lambda_j^{(1)} &= \log(j)/\log(20), \quad j = 2, \ldots, 20, \quad \lambda_1^{(1)} = 1/400, \end{aligned}$$

having clusters close to zero and one, respectively. Each normal matrix having either the $\lambda_j^{(0)}$ or the $\lambda_j^{(1)}$ as its eigenvalues has the (moderate) condition number 400. Nevertheless, the GMRES residual norms in the next-to-last step for the two eigenvalue sets may differ by several orders of magnitude. While the value of (2.14) for the eigenvalues $\lambda_j^{(0)}$ is typically close to one, it is typically of order 10^{-10} for the eigenvalues $\lambda_j^{(1)}$. This is a numerical illustration why the convergence bounds for GMRES and other Krylov subspace methods such as CG and MINRES that are based on the condition number *only* (see [3, Chapter 3.1] for an overview), can provide misleading information about the actual convergence behavior.

3 Worst-case residual norm.

In this section we study the worst-case GMRES residual norms for normal matrices. By "worst-case" we mean, for a given matrix A, the maximally attainable GMRES residual norm in every iteration step i. To make our notion precise we introduce the following definition.

Definition 3.1. *An ith worst-case GMRES residual r_i^w for $A \in \mathbb{C}^{n \times n}$ is a GMRES residual that satisfies*

$$\|r_i^w\| = \max_{\|r_0\|=1} \min_{p \in \pi_i} \|p(A) r_0\|, \quad i = 1, \ldots, n-1. \tag{3.1}$$

A few remarks concerning our definition are in place. First, the restriction that $\|r_0\| = 1$ in (3.1) is made for convenience only. If we drop this restriction, then the right-hand side of (3.1) and all subsequent formulas based on (3.1) must be multiplied by $\|r_0\|$.

Second, as indicated by the wording of the definition, worst-case residuals are not unique. For example, when $r_0^{(i)}$ yields a certain ith worst-case residual r_i^w for a given matrix A, then for all $|\alpha| = 1$, $\alpha r_0^{(i)}$ yields, for the same A, the ith GMRES residual αr_i^w. Obviously, $\|r_i^w\| = \|\alpha r_i^w\|$, so that all vectors αr_i^w are ith worst-case residuals for A.

Third, for each normal matrix $A \in \mathbb{C}^{n \times n}$ (with n distinct eigenvalues) and each GMRES iteration step $i = 1, \ldots, n-1$, there exists an ith worst-case residual r_i^w. The reasoning goes as follows. Assuming that $\|r_0\| = 1$, the standard upper bound on the GMRES residual norms [14, Proposition 4] follows easily from (2.2),

$$\|r_i\| \leq \min_{p \in \pi_i} \|p(A)\| = \min_{p \in \pi_i} \max_{\lambda_j \in L} |p(\lambda_j)|. \tag{3.2}$$

The quantity $\min_{p \in \pi_i} \|p(A)\|$ (called the "ideal GMRES" approximation [5]) is independent of r_0 and thus represents an upper bound on the worst-case GMRES residual norm for the matrix A in step i. As shown independently in [4] and [8], for each normal matrix A and each step i, there exists an initial residual $r_0^{(i)}$ so that equality holds in (3.2). Clearly, the ith GMRES residual corresponding to $r_0^{(i)}$ is an ith worst-case GMRES residual for A in the sense of Definition 3.1.

Fourth, except for special cases, there exists no *single* initial residual that leads to a worst-case GMRES residual norm $\|r_i^w\|$ in *every* step i. Typically the worst-case GMRES residual norm is in each step i achieved by a *different* initial residual $r_0^{(i)}$.

Fifth, since we assume that all eigenvalues are distinct, it holds $\|r_i^w\| > 0$ for $i = 0, \ldots, n-1$. Therefore, the initial residual $r_0^{(i)}$ corresponding to r_i^w has at least $i+1$ nonzero coordinates in the eigenvector basis.

For each subset S of the eigenvalues of A, $S \subseteq L$, we denote

$$M_i^S \equiv \min_{p \in \pi_i} \max_{\lambda_j \in S} |p(\lambda_j)|. \tag{3.3}$$

The result of [4, 8], which will play an important role in our further development, can in this notation be phrased as follows: For each normal matrix $A \in \mathbb{C}^{n \times n}$

(with n distinct eigenvalues) and each $i = 1, \ldots, n-1$, there exists a worst-case GMRES residual r_i^w with

$$\|r_i^w\| = M_i^L. \tag{3.4}$$

As outlined in the Introduction it is of great interest to find explicit formulas for the polynomials that achieve the min-max value M_i^L, and to identify the properties of the initial residuals $r_0^{(i)}$ that yield a worst-case GMRES residual in step i. In the following we will address these questions. We will first consider the iteration step $i = n-1$, and then the case of a general iteration step i.

3.1 Worst case in step $n-1$.

The following result completely characterizes the worst-case GMRES in the next-to-last iteration step.

THEOREM 3.1. *For a given normal matrix $A \in \mathbb{C}^{n\times n}$ with n distinct eigenvalues the unit norm initial residual $r_0^{(n-1)}$ yields an $(n-1)$st worst-case GMRES residual if and only if the coordinates of $r_0^{(n-1)}$ in the eigenvectors of A satisfy*

$$|\varrho_j^{(n-1)}|^2 = \frac{|l_j(0)|}{\sum_{k=1}^n |l_k(0)|}, \quad j = 1, \ldots, n. \tag{3.5}$$

The norm of the $(n-1)$st worst-case GMRES residual r_{n-1}^w is given by

$$\|r_{n-1}^w\| = \left(\sum_{k=1}^n |l_k(0)|\right)^{-1}, \tag{3.6}$$

and the corresponding worst-case GMRES polynomial $p_{n-1}^w(\lambda)$ has the form

$$p_{n-1}^w(\lambda) = \|r_{n-1}^w\| \sum_{j=1}^n \frac{\overline{l_j(0)}}{|l_j(0)|}\, l_j(\lambda). \tag{3.7}$$

Moreover,

$$|p_{n-1}^w(\lambda_j)| = \|r_{n-1}^w\| = M_{n-1}^L, \quad j = 1, \ldots, n, \tag{3.8}$$

where L denotes the set of eigenvalues of A.

PROOF. To find an $(n-1)$st worst-case GMRES residual we need to maximize the GMRES residual norm given by (2.14) under the constraint that the initial residual has unit norm. This is equivalent to solving the following constraint minimization problem for the coordinates of the initial residual in the eigenvectors of A,

$$\min_{\varrho_1^{(n-1)}\neq 0, \ldots, \varrho_n^{(n-1)}\neq 0} \sum_{j=1}^n \frac{|l_j(0)|^2}{|\varrho_j^{(n-1)}|^2}, \quad \text{where } \sum_{j=1}^n |\varrho_j^{(n-1)}|^2 = 1.$$

According to Cauchy's inequality,

$$\sum_{j=1}^n \left|\frac{l_j(0)}{\varrho_j^{(n-1)}}\right|^2 = \sum_{j=1}^n \left|\frac{l_j(0)}{\varrho_j^{(n-1)}}\right|^2 \sum_{j=1}^n |\varrho_j^{(n-1)}|^2 \geq \left(\sum_{j=1}^n |l_j(0)|\right)^2,$$

with equality if and only if

$$\xi \left| \frac{l_j(0)}{\varrho_j^{(n-1)}} \right| = |\varrho_j^{(n-1)}| \quad \Leftrightarrow \quad \xi |l_j(0)| = |\varrho_j^{(n-1)}|^2,$$

for all $j = 1, \ldots, n$ and some real ξ. The number ξ is determined from

$$\xi \sum_{k=1}^{n} |l_k(0)| = \sum_{k=1}^{n} |\varrho_k^{(n-1)}|^2 = 1 \quad \Rightarrow \quad \xi = \left(\sum_{k=1}^{n} |l_k(0)| \right)^{-1}.$$

Hence $|\varrho_j^{(n-1)}|^2$ satisfies (3.5) and the norm of the corresponding worst-case residual is given by (3.6).

Next, if we substitute $|\varrho_j^{(n-1)}|^2$ in the form (3.5) into (2.15) and use the fact that $|\varrho_j^{(n-1)}|^2 = |l_j(0)| \|r_{n-1}^w\|$, then we obtain the worst-case polynomial (3.7). Finally, since $l_j(\lambda_k) = \delta_{jk}$, the worst-case polynomial has at every eigenvalue the same absolute value as shown in the first equality in (3.8), and the second equality in (3.8) follows from (3.4) with $i = n - 1$. □

Remark 3.1. Note that the theorem gives, besides the GMRES context, the explicit solution for a general polynomial approximation problem in the complex plane. In particular, (3.6) can be derived with some effort from the results of [13, Section 3]. It can be shown that (3.6) is equivalent to

$$M_i^L = \frac{|\mathrm{det} V_n|}{\sum\limits_{j=1}^{n} |\mathrm{det} V_n^{(j)}|},$$

where $V_n^{(j)}$ denotes the $(n-1)$-by-$(n-1)$ matrix resulting from deletion of the first column and jth row of V_n. In our notation, this corresponds to the formula given in [13, Remark 3, p. 692]. The formulas (3.5) and (3.6) were derived in [17, Lemma 4.1] for real symmetric matrices. However, we are unaware that (3.5) or (3.7) have been derived before for general normal matrices.

Remark 3.2. We point out that the $(n-1)$st worst-case GMRES polynomial $p_{n-1}^w(\lambda)$ as given in (3.7) is uniquely determined, since it depends only on the uniquely determined quantities $\|r_{n-1}^w\|$ and $l_j(\lambda)$, $j = 1, \ldots, n$.

Theorem 3.1 generalizes the results of [2, Section 2] (for $i = n - 1$) from Hermitian to all normal matrices. In addition, the theorem allows to give new proofs for a number of known results. We present two examples:

1. *Complete stagnation of GMRES.* The question we ask is whether for a given normal matrix A there exists a unit norm vector $r_0^{(n-1)}$ such that GMRES completely stagnates, i.e.

$$1 = \|r_0^{(n-1)}\| = \|r_{n-1}^w\|.$$

Using Theorem 3.1 and the uniqueness of the $(n-1)$st worst-case GMRES polynomial it is easy to see that in case of complete stagnation this polynomial is given by $p^w_{n-1}(\lambda) \equiv 1$. Then (3.7) implies

$$p^w_{n-1}(\lambda_j) = \frac{\overline{l_j(0)}}{|l_j(0)|} = 1, \quad j = 1, \ldots, n.$$

In other words, complete stagnation can occur only if all $l_j(0)$, $j = 1, \ldots, n$, are real and positive. Using other means this result was previously derived in [18, Theorem 3.1].

2. *Ideal GMRES approximation.* The proofs of (3.4) in [4, 8] are based on intricate constructions. For the special case $i = n-1$ we now give a simple proof of (3.4), i.e. that

$$\max_{\|r_0\|=1} \min_{p \in \pi_i} \|p(A) r_0\| = \min_{p \in \pi_i} \|p(A)\|$$

holds for all normal matrices A. As in (3.6) and (3.7), let r^w_{n-1} and $p^w_{n-1}(\lambda)$ denote an $(n-1)$st worst-case GMRES residual and polynomial for A, respectively. Then

$$\begin{aligned} \min_{p \in \pi_{n-1}} \|p(A)\| &= \min_{p \in \pi_{n-1}} \max_{\lambda_j \in L} |p(\lambda_j)| \\ &\le \max_{\lambda_j \in L} |p^w_{n-1}(\lambda_j)| \\ &= \|r^w_{n-1}\| \\ &= \max_{\|r_0\|=1} \min_{p \in \pi_{n-1}} \|p(A) r_0\| \\ &\le \min_{p \in \pi_{n-1}} \|p(A)\|, \end{aligned}$$

so that equality must hold throughout. Note that for the last inequality we have used the standard bound (3.2).

3.2 Worst case in a general step i.

We next attempt to characterize the worst-case GMRES in a general iteration step $i < n-1$. To this end we derive a lower bound on the min-max value

$$M^L_i = \min_{p \in \pi_i} \max_{\lambda_j \in L} |p(\lambda_j)|.$$

We use the simple fact that

$$M^L_i \ge M^S_i \tag{3.9}$$

holds for any set $S \subseteq L$. For any set $S \subseteq L$ containing $i+1$ distinct elements there exists a normal $(i+1)$-by-$(i+1)$ matrix with the spectrum S. To this matrix we can apply Theorem 3.1 which completely characterizes the worst-case GMRES in step i, and in particular shows that

$$M^S_i = \left(\sum_{k=1}^{i+1} |l^S_k(0)| \right)^{-1}, \tag{3.10}$$

where $l_k^S(\lambda)$, $k = 1, \dots, i{+}1$, denotes the kth Lagrange polynomial corresponding to the elements in the set S. Using (3.9) and (3.10) it is easy to see that for the given matrix A with the spectrum L,

$$M_i^L \geq \max_{\substack{S \subseteq L \\ |S|=i+1}} M_i^S = \max_{\substack{S \subseteq L \\ |S|=i+1}} \left(\sum_{k=1}^{i+1} |l_k^S(0)| \right)^{-1} . \tag{3.11}$$

The natural question arises how close is the lower bound (3.11). In the following we will discuss this question and distinguish between two situations: Either all eigenvalues of A are real, or A has at least one non-real eigenvalue. The first case covers symmetric and hermitian matrices, the second case all other normal matrices.

3.2.1 All eigenvalues are real: (3.11) *is an equality.*

When all eigenvalues forming the set L are real, then it follows from a classical result of approximation theory that (3.11) is an equality for $i = 1, \dots, n{-}1$. This means that for each $i = 1, \dots, n-1$ there exists a set $\widehat{S} \subseteq L$ with $|\widehat{S}| = i + 1$, such that

$$M_i^L = M_i^{\hat{S}} = \left(\sum_{k=1}^{i+1} |l_k^{\hat{S}}(0)| \right)^{-1} ,$$

see, e.g., [1, Theorem 2.4 and Corollary 2.5]. In this case Theorem 3.1 can be applied to a normal $(i + 1)$-by-$(i + 1)$ matrix having the elements of $\widehat{S}$ as its eigenvalues. Then (3.5) shows that the coordinates of $r_0^{(i)}$ yielding the worst-case GMRES residual for A in step i satisfy

$$|\varrho_j^{(i)}|^2 = \frac{|l_j^{\hat{S}}(0)|}{\sum_{k=1}^{i+1} |l_k^{\hat{S}}(0)|} \quad \text{if } \lambda_j \in \widehat{S}, \qquad \varrho_j^{(i)} = 0 \quad \text{if } \lambda_j \notin \widehat{S}.$$

Since $r_0^{(i)}$ has only $i + 1$ nonzero coordinates in the eigenvectors of A, GMRES will for this initial residual have the worst-case residual norm in the step i, but then terminate in the subsequent step $i + 1$. Using a different approach, these results have been previously derived for symmetric positive definite matrices in the context of the conjugate gradient method [2].

3.2.2 At least one non-real eigenvalue: (3.11) *may be strict.*

When L contains at least one non-real eigenvalue, then (3.11) may be strict for $i = 1, \dots, n - 2$. In fact, the smallest set $S \subseteq L$ for which $M_i^L = M_i^S$ might contain as many as $2i + 1$ distinct elements in the general complex case, see, e.g., [1, Corollary 2.5]. For $|S| > i + 1$, however, the results of Theorem 3.1 cannot be used, and we are unable to express M_i^S in terms of explicit polynomials. Still, the inequality (3.11) represents a lower bound for M_i^L. Furthermore, we can find an upper bound for M_i^L using an approach similar to the proof of [7, Theorem 4.1].

THEOREM 3.2. *For any set L of n distinct complex points it holds*

$$M_i^L \leq \sqrt{(i+1)(n-i)} \max_{\substack{S \subseteq L \\ |S|=i+1}} M_i^S, \quad i=1,\dots,n-2. \tag{3.12}$$

PROOF. Consider any normal matrix $A \in \mathbb{C}^{n\times n}$ having n distinct eigenvalues forming the set L. Let $r_0^{(i)}$ denote an initial residual that yields an ith worst-case GMRES residual r_i^w and let $\varrho_j^{(i)}$, $j=1,\dots,n$, denote the coordinates of $r_0^{(i)}$ in the eigenvectors of A. The min-max value M_i^L can be written, according to (3.4) and (2.9), in the form

$$M_i^L = \|r_i^w\| = \|e_1^H (D_i V_{i+1})^+\|^{-1},$$

where $D_i \equiv \mathrm{diag}(\varrho_1^{(i)},\dots,\varrho_n^{(i)})$. Now consider $i+1$ rows of $D_i V_{i+1}$ that form a square matrix U of order $i+1$ such that $|\det(U)|$ is maximal. Then, as in the proof of [7, Theorem 4.1],

$$\|r_i^w\| \leq \sqrt{(i+1)(n-i)}\; \|U^{-H} e_1\|^{-1}. \tag{3.13}$$

The matrix U is defined by some $i+1$ eigenvalues and by corresponding coordinates $\varrho_j^{(i)}$. Denote the set of eigenvalues that define U by $\widehat{S} = \{\lambda_1^{\hat{S}},\dots,\lambda_{i+1}^{\hat{S}}\}$ and the corresponding (nonzero) coordinates by $\varrho_1^{\hat{S}},\dots,\varrho_{i+1}^{\hat{S}}$. Using (2.13), $\|r_0^{(i)}\|=1$, and Cauchy's inequality we obtain

$$\|U^{-H}e_1\|^2 = \sum_{j=1}^{i+1} \left| \frac{l_j^{\hat{S}}(0)}{\varrho_j^{\hat{S}}} \right|^2 \geq \sum_{j=1}^{i+1} \left| \frac{l_j^{\hat{S}}(0)}{\varrho_j^{\hat{S}}} \right|^2 \sum_{j=1}^{i+1} |\varrho_j^{\hat{S}}|^2 \geq \left(\sum_{j=1}^{i+1} |l_j^{\hat{S}}(0)| \right)^2,$$

i.e.

$$\|U^{-H}e_1\|^{-1} \leq \left(\sum_{j=1}^{i+1} |l_j^{\hat{S}}(0)| \right)^{-1} = M_i^{\hat{S}}. \tag{3.14}$$

Thus we have found a set $\widehat{S} \subseteq L$, $|\widehat{S}| = i+1$, such that

$$\|r_i^w\| \leq \sqrt{(i+1)(n-i)}\; M_i^{\hat{S}}. \tag{3.15}$$

Substituting in (3.15) for $M_i^{\hat{S}}$ the maximum of M_i^S over all subsets $S \subseteq L$, $|S| = i+1$, we obtain (3.12). □

Our numerical experiments with various spectra (see Section 4) show that the lower bound (3.11) is very tight and that the upper bound (3.12) represents an overestimation. In particular, we *conjecture* that there exists a small constant $C > 1$ such that

$$\max_{\substack{S \subseteq L \\ |S|=i+1}} M_i^S \leq M_i^L \leq C \max_{\substack{S \subseteq L \\ |S|=i+1}} M_i^S, \quad i=1,\dots,n-2, \tag{3.16}$$

holds for *all* sets L containing n distinct complex numbers (for $i = n-1$, (3.16) obviously holds with $C = 1$). In our numerical tests the ratio

$$\frac{M_i^L}{\max\limits_{\substack{S \subseteq L \\ |S|=i+1}} M_i^S} \tag{3.17}$$

was maximal for sets L containing n numbers uniformly distributed on the unit circle. On such sets of points, (3.17) for $i = n-2$ converges from below to $4/\pi$ as $n \to \infty$. Hence $C = 4/\pi$ is the smallest constant for which (3.16) can hold for all sets L with $|L| = n$, cf. the Appendix. On the other hand, we were unable to find a set L for which the ratio (3.17) was larger than $4/\pi$.

4 Numerical experiments.

We now study the worst-case GMRES residual norms, our lower bound (3.11), and our conjecture (3.16) with $C = 4/\pi$ for four different eigenvalue sets L. In the left part of Figures 4.1– 4.4 we plot the worst-case GMRES residual norms $\|r_i^w\|$ (bold line), and the values

$$\max_{\substack{S \subseteq L \\ |S|=i+1}} M_i^S \qquad \text{(solid line)},$$

$$\frac{4}{\pi} \max_{\substack{S \subseteq L \\ |S|=i+1}} M_i^S \qquad \text{(dashed line)}.$$

Our conjecture is that the dashed curve is an upper bound on the worst-case GMRES residual norm in every step. The right part of each figure shows the corresponding eigenvalue distributions. In the step i, we compute the values M_i^S for all subsets $S \subseteq L$, $|S| = i+1$, and determine our bounds from their maximum. This computation is quite expensive, so we consider only small sets of points ($n = 18$). The worst-case GMRES residual norm in every step is computed using the function `cheby0` of the semidefinite programming package SDPT3 [15]. Although this function may fail to converge when $\|r_i^w\|$ becomes very small (see below for details), it is the most reliable function we know for this type of computation. All experiments are performed in Matlab 6.5 Release 13 on an AMD Athlon XP 2100+ personal computer with machine precision $\varepsilon \sim 10^{-16}$.

Roots of unity. In the first numerical experiment we consider the eigenvalue set L consisting of the 18th roots of unity, i.e.

$$\lambda_k = \mathrm{e}^{\mathbf{i}\frac{2k\pi}{18}}, \quad k = 1, \ldots, 18. \tag{4.1}$$

In this case worst-case GMRES completely stagnates, cf. [18], which is confirmed by the bold line in Figure 4.1. The lower bound (3.11) closely approximates the worst-case residual norm, and the lower bound multiplied by $4/\pi$ represents an upper bound. As shown in the Appendix, see also [12], the lower bound approaches $\pi/4$ from above in the step $i = n-2$ (here: $i = 16$) when $n \to \infty$.

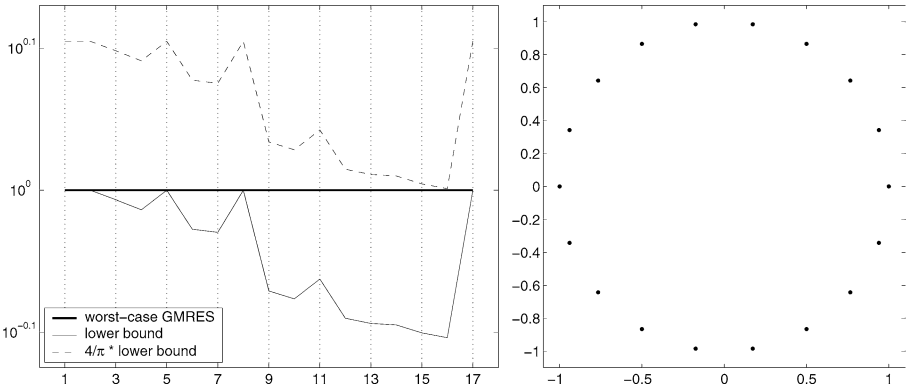


Figure 4.1: Worst-case GMRES and our bounds for roots of unity.

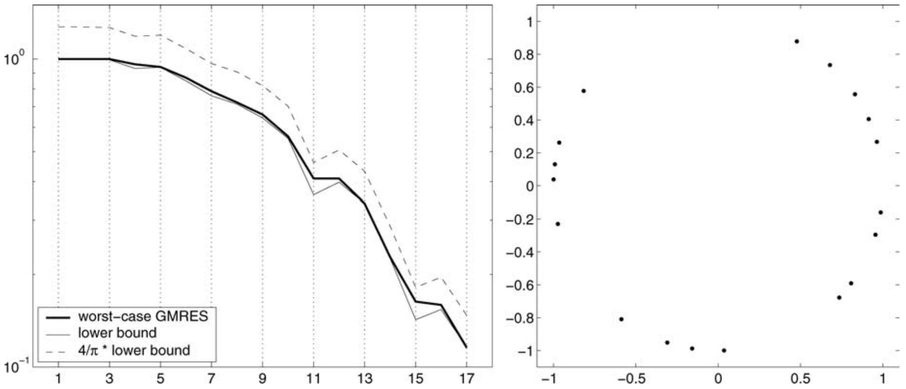


Figure 4.2: Worst-case GMRES and our bounds for random eigenvalues on the unit circle.

Hence in this step the lower bound multiplied by $4/\pi$ is proven to be a (sharp) upper bound on the worst-case GMRES. The tightness of this bound, even for the moderate $n = 18$, is clearly visible in Figure 4.1.

Random eigenvalues on the unit circle. For random eigenvalues on the unit circle (cf. the right part of Figure 4.2), the worst-case GMRES residual norms do not stagnate completely, but still converge very slowly (decreasing only about one order of magnitude until the next-to-last step). The lower bound (3.11) is very close to the worst-case residual norm, only in the iteration steps 4, 7, 11 and 15 they differ slightly. As above, the lower bound multiplied by $4/\pi$ represents an upper bound.

Random eigenvalues in the region $[0, 1] \times \mathbf{i}[0, 1]$. In this case the convergence of the worst-case residual norms is moderately fast; they decrease about 4 orders of magnitude until the next-to-last step, cf. Figure 4.3. Again the lower bound (3.11) is a good estimate (bold and soild line almost coincide), and the dashed line represents an upper bound.

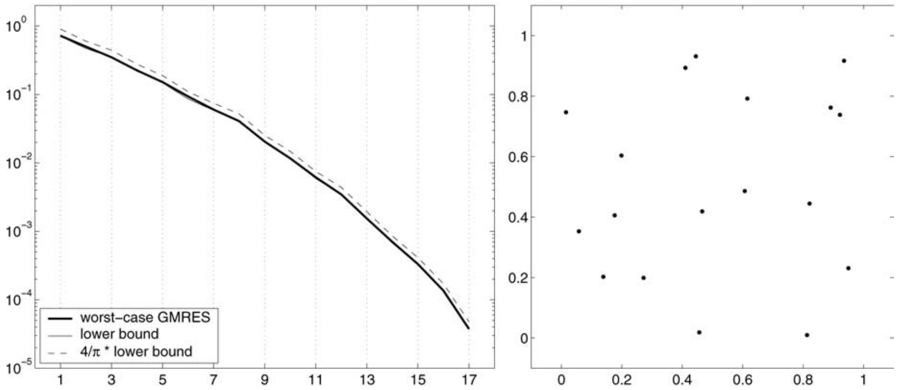


Figure 4.3: Worst-case GMRES and our bounds for random eigenvalues in the region $[0, 1] \times \mathbf{i}[0, 1]$.

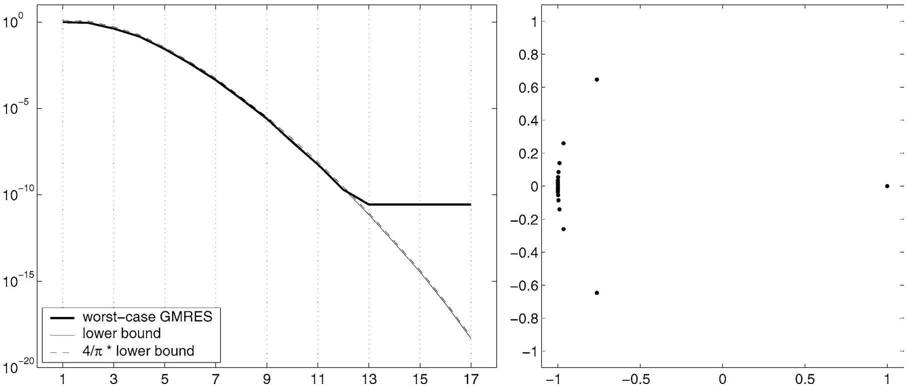


Figure 4.4: Worst-case GMRES and our bounds for the Helmert matrix.

Helmert matrix. For the last experiment we use the Helmert matrix generated by the Matlab command `gallery('orthog',18,4)`. Helmert matrices occur in a number of practical problems, for example in applied statistics [9]. Our matrix is orthogonal, and the eigenvalues cluster around -1, see the right part of Figure 4.4. The worst-case GMRES residual norm decreases quickly throughout the iteration. Until the 12th step the worst-case curve and the lower bound almost coincide, and the lower bound multiplied by $4/\pi$ represents an upper bound. However, when the worst-case residual norm drops below the level of 10^{-10} (iteration step 13 and beyond), the function `cheby0` apparently has reached its final level of accuracy and henceforth stagnates. Such stagnation (sometimes divergence) can be generally observed when the final accuracy level is reached, but we are unaware of an analysis how this level depends on the problem parameters.

In summary, the numerical experiments demonstrate that our lower bound (3.11) is very tight. Moreover, in all experiments the lower bound multiplied by $4/\pi$ represents an upper bound on the worst-case GMRES residual norms, which supports that our conjecture (3.16) with $C = 4/\pi$ is true. Note that in all experiments the bound (3.12), which contains a factor between $n^{1/2}$ and n, represents an overestimation.

5 Concluding discussion.

We conclude the paper with a discussion of our results and starting points for further work.

1. Interpretation of the lower bound (3.11). Recall that the worst-case GMRES residual norm in step i is equal to the min-max value M_i^L. This value represents the solution of an ith degree polynomial approximation problem on n distinct eigenvalues forming the set L. We bound this value from below by the same approximation problem, but on subsets of L containing exactly $i+1$ eigenvalues. The solution of each "reduced" problem (polynomial of degree i on $i+1$ distinct points) is given in Theorem 3.1.

For illustration of this process we consider the set L consisting of the nth roots of unity, cf. Figure 4.1 for $n = 18$. As shown in [18, Theorem 3.4], worst-case GMRES completely stagnates in this case, i.e. $M_{n-1}^L = 1$. Intuitively, for each $i < n-1$ there exists a subset $\widehat{S} \subset L$, $|\widehat{S}| = i+1$, that closely resembles the $(i+1)$st roots of unity. For such a set the min-max value $M_i^{\widehat{S}}$ is close to one (in orders of magnitude), which is why the lower bound (3.11) is very tight. In particular, whenever $n \bmod (i+1) = 0$, there exists a set $\widehat{S} \subset L$ consisting of exactly the $(i+1)$st roots of unity. For all such iteration steps i, the lower bound (3.11) is an equality, cf. $i = 1, 2, 5, 8$ for $n = 18$ in Figure 4.1.

Note that here, and for general sets L, M_i^S is close to M_i^L only for special sets $S \subset L$ with $|S| = i+1$. Analyzing the structure of these sets based on the eigenvalue distribution of A is a topic we plan to pursue in our future work.

2. Worst case vs. average (unbiased) case. Due to orthogonality of the eigenvectors of A, initial residuals with (approximately) equal components in all eigenvectors are often considered the "average" case, see, e.g., [5, Section 7]. We prefer to call them "unbiased" since they are not biased towards a certain eigenvector direction. For simplicity, consider any unbiased unit norm initial residual r_0^u with eigenvector components of *equal* size, i.e. $|\varrho_j^u| = n^{-1/2}$, $j = 1, \ldots, n$. Then (2.5) and (3.4) show that the ith GMRES residual r_i^u corresponding to r_0^u satisfies

$$\begin{aligned} \|r_i^w\| \geq \|r_i^u\| &= n^{-1/2} \min_{p \in \pi_i} \left(\sum_{j=1}^n |p(\lambda_j)|^2 \right)^{-1/2} \\ &\geq n^{-1/2} \min_{p \in \pi_i} \max_{1 \leq j \leq n} |p(\lambda_j)| \\ &= n^{-1/2} \|r_i^w\|. \end{aligned}$$

Since the unbiased (average) and the worst case GMRES residual norms agree up to a factor of $n^{1/2}$, our results are relevant beyond the specific analysis of the worst-case GMRES.

In practical applications the initial residual may, for example, be biased towards the eigenvalue distribution of A. Often such biased initial residuals result from choosing a nonzero initial guess x_0. The biased case depends strongly on the specific application, and a general analysis is beyond the scope of this paper.

Acknowledgements.

We thank Prof. Ludwig Elsner for pointing out a simplification in a previous version of our proof of Theorem 3.1. We also thank an anonymous referee for helpful comments.

Appendix.

PROPOSITION A.1. *The smallest constant C for which* (3.16) *can hold for all sets L containing n distinct complex numbers is $C = 4/\pi$.*

PROOF. We will show that for the set $L = \{\lambda_1, \ldots, \lambda_n\}$ defined by

$$\lambda_k = \mathrm{e}^{\mathbf{i}\frac{2k\pi}{n}}, \quad k = 1, \ldots, n, \tag{A.1}$$

the ratio (3.17) for $i = n-2$ converges from below to $C = 4/\pi$ as $n \to \infty$.

Note that all sets $S \subset L$ with $|S| = n-1$ can be obtained by rotation of the set $L - \{\lambda_1\}$. Therefore

$$\max_{\substack{S \subseteq L \\ |S| = n-1}} M_{n-2}^{S} = M_{n-2}^{L-\{\lambda_1\}} = \left(\sum_{k=2}^{n} \prod_{\substack{j=2 \\ j \neq k}}^{n} \frac{|\lambda_j|}{|\lambda_j - \lambda_k|} \right)^{-1}.$$

Substituting $|\lambda_j| = 1$ for all j, and

$$|\lambda_j - \lambda_k| = |\mathrm{e}^{\mathbf{i}\frac{2j\pi}{n}} - \mathrm{e}^{\mathbf{i}\frac{2k\pi}{n}}| = 2 \sin\left(\frac{|j-k|\pi}{n} \right),$$

shows that

$$M_{n-2}^{L-\{\lambda_1\}} = \left(\sum_{k=1}^{n-1} \frac{1}{2^{n-2}} \prod_{\substack{j=1 \\ j \neq k}}^{n-1} \frac{1}{\sin\left(\frac{j\pi}{n}\right)} \right)^{-1} = 2^{n-2} \left(\frac{\sum_{k=1}^{n-1} \sin\left(\frac{k\pi}{n}\right)}{\prod_{j=1}^{n-1} \sin\left(\frac{j\pi}{n}\right)} \right)^{-1}.$$

Using the standard formula

$$\prod_{j=1}^{n-1} \sin\left(\frac{j\pi}{n}\right) = \frac{n}{2^{n-1}},$$

we obtain

$$M_{n-2}^{L-\{\lambda_1\}} = \left[\frac{2}{n} \sum_{k=1}^{n-1} \sin\left(\frac{k\pi}{n}\right) \right]^{-1} = \frac{\pi}{2} \left[\frac{\pi}{n} \sum_{k=1}^{n-1} \sin\left(\frac{k\pi}{n}\right) \right]^{-1}. \tag{A.2}$$

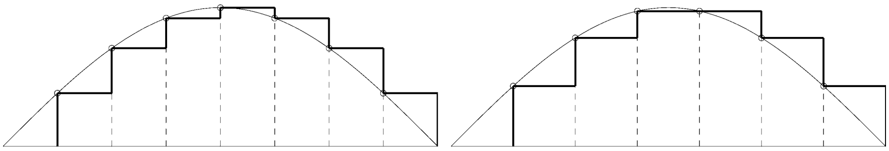


Figure A.1: The approximation of the integral for n even (left part) and n odd (right part).

Note that the expression on the right-hand side of (A.2) is an approximation of an integral,

$$\frac{\pi}{n}\sum_{k=1}^{n-1}\sin\Big(\frac{k\pi}{n}\Big) < \int_0^\pi \sin(x)\mathrm{d}x = 2, \qquad \lim_{n\to\infty}\left[\frac{\pi}{n}\sum_{k=1}^{n-1}\sin\Big(\frac{k\pi}{n}\Big)\right] = 2,$$

see Figure A.1 for a numerical illustration. Therefore,

$$M_{n-2}^{L-\{\lambda_1\}} > \frac{\pi}{4}, \qquad \lim_{n\to\infty} M_{n-2}^{L-\{\lambda_1\}} = \frac{\pi}{4}.$$

As shown in [18], complete stagnation of GMRES can occur for normal matrices having the spectrum L, and hence $M_i^L = 1$ for $i = 1, \ldots, n-1$. Therefore,

$$M_{n-2}^L < \frac{4}{\pi}\max_{\substack{S\subseteq L\\|S|=n-1}} M_i^S, \qquad \lim_{n\to\infty}\left[\frac{4}{\pi}\max_{\substack{S\subseteq L\\|S|=n-1}} M_i^S\right] = M_{n-2}^L,$$

which completes the proof. A similar result can be shown for other i; see [12] for more details. □

REFERENCES

1. R. A. DeVore and G. G. Lorentz, *Constructive Approximation*, Grundlehren Math. Wiss. 303 [Fundamental Principles of Mathematical Sciences], Springer-Verlag, Berlin, 1993.
2. A. Greenbaum, *Comparison of splittings used with the conjugate gradient algorithm*, Numer. Math., 33 (1979), pp. 181–193.
3. A. Greenbaum, *Iterative Methods for Solving Linear Systems*, Frontiers in Appl. Math. 17, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.
4. A. Greenbaum and L. Gurvits, *Max-min properties of matrix factor norms*, SIAM J. Sci. Comput., 15 (1994), pp. 348–358.
5. A. Greenbaum and L. N. Trefethen, *GMRES/CR and Arnoldi/Lanczos as matrix approximation problems*, SIAM J. Sci. Comput., 15 (1994), pp. 359–368.
6. N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996.
7. I. Ipsen, *Expressions and bounds for the GMRES residual*, BIT, 40 (2000), pp. 524–535.
8. W. Joubert, *A robust GMRES-based adaptive polynomial preconditioning algorithm for nonsymmetric linear systems*, SIAM J. Sci. Comput., 15 (1994), pp. 427–439.
9. H. Lancaster, *The Helmert matrices*, Amer. Math. Monthly, 72 (1965), pp. 4–12.
10. J. Liesen, *Construction and analysis of polynomial iterative methods for non-Hermitian systems of linear equations*, PhD thesis, Fakultät für Mathematik der Universität Bielefeld, November 1998.
11. J. Liesen, M. Rozložník, and Z. Strakoš, *Least squares residuals and minimal residual methods*, SIAM J. Sci. Comput., 23 (2002), pp. 1503–1525.

12. J. Liesen and P. Tichý, *A min-max problem on roots of unity*, Preprint 28-2003, Institute of Mathematics, Technical University of Berlin, 2003.
13. T. J. Rivlin and H. S. Shapiro, *A unified approach to certain problems of approximation and minimization*, J. Soc. Indust. Appl. Math., 9 (1961), pp. 670–699.
14. Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.
15. K. Toh, M. Todd, and R. Tütüncü, *SDPT3 – a Matlab software package for semidefinite programming, version 2.1*, the software is available on the site `http://www.math.nus.edu.sg/~mattohkc/sdpt3.html`, June 2001.
16. J. M. Varah, *The prolate matrix*, Linear Algebra Appl., 187 (1993), pp. 269–278.
17. I. Zavorin, *Spectral factorization of the Krylov matrix and convergence of GMRES*, Tech. Rep. CS-TR-4309, Computer Science Department, University of Maryland, 2001.
18. I. Zavorin, D. O'Leary, and H. Elman, *Complete stagnation of GMRES*, Linear Algebra Appl., 367 (2003), pp. 165–183.

Convergence analysis of Krylov subspace methods[†]

Jörg Liesen[*1] and **Petr Tichý**[1]

[1] Institute of Mathematics, Technical University of Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany

Received ..., revised ..., accepted ...
Published online ...

Key words Krylov subspace methods, convergence analysis.
MSC (2000) 15A06, 65F10, 41A10

One of the most powerful tools for solving large and sparse systems of linear algebraic equations is a class of iterative methods called Krylov subspace methods. Their significant advantages like low memory requirements and good approximation properties make them very popular, and they are widely used in applications throughout science and engineering. The use of the Krylov subspaces in iterative methods for linear systems is even counted among the "Top 10" algorithmic ideas of the 20th century. Convergence analysis of these methods is not only of a great theoretical importance but it can also help to answer practically relevant questions about improving the performance of these methods. As we show, the question about the convergence behavior leads to complicated nonlinear problems. Despite intense research efforts, these problems are not well understood in some cases. The goal of this survey is to summarize known convergence results for three well-known Krylov subspace methods (CG, MINRES and GMRES) and to formulate open questions in this area.

1 Introduction

Krylov subspace methods represent one of the most important classes of iterative methods for solving linear algebraic systems. Their main common ingredient are the Krylov subspaces, which are spanned by the initial residual and by vectors formed by repeated multiplication of the initial residual by the system matrix. These subspaces first appeared in a paper by the Russian scientist and navy general Aleksei Nikolaevich Krylov (1863–1945), published in 1931 [44]. Motivated by an application in naval science, Krylov was interested in analyzing oscillations of mechanical systems, and proposed a method for computing the minimal polynomial of a given matrix (see, e.g., [21, Section 42], [25, Chapter VII], or [38, Chapter 6] for detailed accounts of Krylov's method). Independently of Krylov's work, the first Krylov subspace methods for solving linear algebraic systems appeared two decades later with the publication of the conjugate gradient (CG) method for hermitian positive definite matrices by Hestenes and Stiefel [36], and the closely related methods developed by Lanczos [45, 46]. Driven by the need to solve linear systems of vastly increasing dimension and the accompanying rapid development of computational resources, these Krylov subspace methods were

[*] Corresponding author: e-mail: liesen@math.tu-berlin.de, Phone: +49 30 314 29295, Fax: +49 30 314 79706

[†] This work was supported by the Emmy Noether-Programm of the Deutsche Forschungsgemeinschaft.

used in many applications, particularly in the engineering community. In the numerical linear algebra community, the potential of Krylov subspace methods was fully recognized only after an influential paper of Reid appeared in 1971 [58]. Subsequently, numerous additional Krylov subspace methods were developed, with focus on indefinite and nonhermitian matrices. Today, the use of the Krylov subspaces in iterative methods for linear systems is counted among the "Top 10" algorithmic ideas of the 20th century [10]. One of the main reasons for this success is that the Krylov subspaces can be build up using only a function that computes the multiplication of the system matrix and a vector, so that the system matrix itself never has to be formed or stored explicitly. Hence Krylov subspace methods are particularly well suited for application to large and sparse linear systems, which today are commonplace throughout applications in science and engineering.

Mathematically, Krylov subspace methods are based on projection methods. Instead of solving the potentially very large linear system, the idea is to approximate the systems' solution from Krylov subspaces of small dimension. The goal of the convergence analysis of these methods is to *describe the convergence of this process in terms of input data of the given problem*, i.e. in dependence on properties of the system matrix, the right hand side vector and the initial guess. Understanding the convergence of Krylov subspace methods is particularly important to answer the practically relevant questions how to accelerate the convergence (in particular how to precondition the system), and how to choose potential restart parameters.

The goal of this paper is to survey the known theory of convergence of Krylov subspace methods that are based on two basic types of projection methods, namely the Galerkin (orthogonal residual (OR)) method and the minimal residual (MR) method. Both types of methods have been implemented in various commonly used algorithms. An example of the OR Krylov subspace method is the CG method [36] for hermitian positive definite matrices. Implementations of the MR Krylov subspace method are the MINRES method [56] for nonsingular hermitian indefinite matrices and the GMRES method [62] for general nonsingular matrices. The distinction between OR and MR methods made in this paper is not new. In fact it has been extensively used in the past to derive relations between the convergence quantities (e.g. error or residual norms) of different methods, see, e.g., [12, 14, 37]. Here our focus is on giving bounds for the convergence quantities of each method separately.

For normal system matrices A, the (worst-case) convergence behavior of CG, MINRES and GMRES is completely determined by the spectrum of A. The convergence analysis then reduces to analyzing a certain min-max approximation problem on the matrix eigenvalues. In the nonnormal case, however, the convergence behavior of the GMRES method may not be related to the eigenvalues at all. As a consequence, other properties of the input data must be considered to describe the convergence. Despite intense efforts to identify descriptive properties, understanding the convergence of GMRES in the general nonnormal case still remains a largely open problem.

After a brief introduction to the mathematical background of Krylov subspace methods (Section 2), we survey in Section 3 the theory of convergence of these methods. We distinguish between the normal (Section 3.1) and the nonnormal (Section 3.2) case. Section 4 contains concluding remarks. We point out that all convergence results we state in this paper were derived assuming exact arithmetic. A recent survey of the numerical stability of Krylov subspace methods that also discusses effects of finite precision arithmetic on the convergence is given in [65].

2 Krylov subspace methods

In this section we briefly describe the mathematical background of the Krylov subspace methods for solving linear algebraic systems of the form

$$Ax \;=\; b\,, \tag{1}$$

where A is a real or complex nonsingular N by N matrix, and b is a real or complex vector of length N. Suppose that x_0 is an initial guess for the solution x, and define the initial residual $r_0 = b - Ax_0$. As shown originally by Saad [59, 60] (see his book [61] for a summary), Krylov subspace methods can be derived from the following *projection method*: The nth iterate x_n, $n = 1, 2, \ldots$, is of the form

$$x_n \;\in\; x_0 + \mathcal{S}_n\,, \tag{2}$$

where $\mathcal{S}_n$ is some n-dimensional space, called the search space. Because of the n degrees of freedom, n constraints are required to make x_n unique. This is done by choosing an n-dimensional space $\mathcal{C}_n$, called the constraints space, and by requiring that the nth residual is orthogonal to that space, i.e.,

$$r_n \;=\; b - Ax_n \;\in\; r_0 + A\mathcal{S}_n\,, \qquad r_n \;\perp\; \mathcal{C}_n\,. \tag{3}$$

Orthogonality here is meant in the Euclidean inner product. A similar type of projection process appears in many areas of mathematics. As an example, consider the Petrov-Galerkin framework in the context of the finite element method for discretizing partial differential equations, see e.g. [57, Chapter 5]. There the notions of test and trial spaces correspond to search and constraints spaces in (2)–(3).

In this paper we concentrate on the projection method (2)–(3) and two basic relations between $\mathcal{S}_n$ and $\mathcal{C}_n$, that to our mind are among the most important ones:

$$\mathcal{C}_n = \mathcal{S}_n \qquad \text{(Galerkin method)}\,, \tag{4}$$

$$\mathcal{C}_n = A\mathcal{S}_n \qquad \text{(Minimal residual method)}\,. \tag{5}$$

The Galerkin and the minimal residual (MR) method are called a Krylov subspace method when the so-called Krylov subspaces $\mathcal{K}_n(A, r_0)$ are used as search spaces, i.e.

$$\mathcal{S}_n \;=\; \mathcal{K}_n(A, r_0) \;\equiv\; \text{span}\{r_0, Ar_0, \ldots, A^{n-1}r_0\}\,, \quad n = 1, 2, \ldots\,. \tag{6}$$

Using these spaces in the Galerkin method, we construct residuals $r_n = b - Ax_n$ that are orthogonal to all previous residuals $r_{n-1}, \ldots, r_0$. That is why, in the context of Krylov subspaces, the Galerkin method is often called orthogonal residual (OR) method.

There are many possible choices of Krylov subspaces and their variants (e.g. $A\mathcal{K}_n(A, r_0)$, $\mathcal{K}_n(A^H, r_0)$, $A^H\mathcal{K}_n(A^H, r_0)$, etc.) in the projection process (2)–(3). This fact certainly contributes to the overabundant supply of these methods. Also note that for each mathematical description there may be several different implementations that in exact arithmetic satisfy (2)–(3) for given search and constraint spaces, but that may differ in their finite precision behavior. Particularly comprehensive and systematic surveys of existing Krylov subspace methods can be found in [4, 9] and [14].

The Krylov subspaces form a nested sequence that ends with a subspace of maximal dimension $d = \dim \mathcal{K}_N(A, r_0)$, i.e.,

$$\mathcal{K}_1(A, r_0) \subset \cdots \subset \mathcal{K}_d(A, r_0) = \cdots = \mathcal{K}_N(A, r_0)\,.$$

The number of steps of the OR/MR Krylov subspace method is limited by the maximal Krylov subspace dimension d. We say that a projection process *breaks down* in step n if no iterate x_n exists, or if x_n is not unique. Naturally, we are interested in projection methods that ensure existence and uniqueness of their iterates x_n for each step $n \leq d$. Such *well-defined* methods terminate with the exact solution in the step d, which is called the *finite termination property*. If a method is well-defined or not, depends on the properties of the matrix A.

In general, the OR Krylov subspace method yields uniquely defined iterates for each n whenever zero is outside the *field of values* of A, which is defined as

$$\mathcal{F}(A) = \left\{ v^H A v \;:\; \|v\| = 1\,,\; v \in \mathbb{C}^N \right\}\,. \tag{7}$$

However, in this paper we limit our discussion to the OR Krylov subspace method for hermitian positive definite matrices, since only in this case the given system matrix defines a norm in which the errors are minimized (see Section 3.1.1 for details). A particular implementation in this case is the CG method [36].

The MR Krylov subspace method is well defined whenever A is nonsingular. This feature makes this method very popular, since it can be used for general matrices. The most well-known implementations are the MINRES method [56] for hermitian indefinite matrices and the GMRES method [62] for general nonsingular matrices.

Finally, note that the conditions $x_n \in x_0 + \mathcal{K}_n(A, r_0)$ and $r_n \in r_0 + A\mathcal{K}_n(A, r_0)$ imply that the error $x - x_n$ and the residual r_n can be written in the polynomial form

$$x - x_n = p_n(A)(x - x_0), \qquad r_n = p_n(A) r_0, \tag{8}$$

where p_n is a polynomial of degree at most n and with value one at the origin. For a well-defined OR/MR Krylov subspace method, the polynomial p_n is uniquely determined by the constraint conditions (3).

3 Convergence analysis

In exact arithmetic, well-defined Krylov subspace methods terminate in a finite number of steps. Therefore no limit can be formed, and terms like “convergence” or “rate of convergence” loose their classical meaning; see, e.g., [35, Chapter 9.4] for a cautioning in this direction. This situation requires approaches that are substantially different from the analysis of classical fixed point iteration methods such as Gauß-Seidel or SOR. The convergence of the latter methods has typically been described asymptotically, with the “asymptotic convergence factor” of the iteration matrix being the central concept. Surprisingly, this principal difference between the Krylov subspace methods and the classical iteration methods is still not always accepted. For example, the classical convergence *bound* for the CG method that is based on the matrix condition number (see equation (15) below) is sometimes confused with the actual convergence *behavior* of the method. Hence the actual convergence is identified with a bound based on the asymptotic convergence factor of the convex hull of the spectrum, without considering any other properties of the given data. Clearly, this approach can be very misleading in some situations.

A related difficulty in the convergence analysis is the typical requirement of finding an acceptable approximate solution x_n in $n \ll N$ steps. Therefore it is important to understand the convergence from the very beginning, i.e., in the classical terminology, to understand the "transient" behavior. This early stage of convergence, however, can depend significantly on the right hand side b and the initial guess x_0. In general, the non-existing limiting process, the relevance of the transient phase, and the dependence of this phase on b and x_0 make the convergence analysis of Krylov subspace methods a difficult nonlinear problem – although the system to be solved is linear.

We divide our discussion about the convergence of Krylov subspace methods into two parts. In the first part (Section 3.1) we consider normal system matrices A and show that in this case the spectral information is important for analyzing the convergence. The second part (Section 3.2) shows the difficulties with estimating the convergence in the nonnormal case.

3.1 Convergence analysis for normal matrices

Consider a nonsingular and *normal* matrix A, and let

$$A = V\Lambda V^H, \qquad \text{where} \qquad V^H V = I, \qquad \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_N)\,,$$

be its eigendecomposition. The orthogonality of the eigenvector basis lead s to a significant simplification in the convergence analysis of Krylov subspace methods: Considering A^n in the form $V\Lambda^n V^H$ and using (8), the errors and residuals of a Krylov subspace method satisfy

$$x - x_n = V p_n(\Lambda) V^H (x - x_0), \qquad r_n = V p_n(\Lambda) V^H r_0\,. \tag{9}$$

Because the projection property usually refers to some sort of optimality, we can expect that Krylov subspace methods for normal matrices solve some weighted polynomial minimization problem on the matrix spectrum. In the following subsections we explain that in the worst-case, the convergence speed of well-known Krylov subspace methods (CG, MINRES, GMRES) is determined by the value

$$\min_{p \in \pi_n} \max_k |p(\lambda_k)|\,, \tag{10}$$

where π_n denotes the set of polynomials of degree at most n and with value one at the origin. Note that the value (10) represents a min-max approximation problem on the discrete set of the matrix eigenvalues. The value (10) depends in a complicated (nonlinear) way on the eigenvalue distribution. Consider, for simplicity, that all eigenvalues are real and distinct. The results in [26, 51] show that there exists a subset of $n+1$ (distinct) eigenvalues $\{\mu_1, \ldots, \mu_{n+1}\} \subseteq \{\lambda_1, \ldots, \lambda_N\}$, such that

$$\min_{p \in \pi_n} \max_k |p(\lambda_k)| \;=\; \left(\sum_{j=1}^{n+1} \prod_{\substack{k=1 \\ k \neq j}}^{n+1} \frac{|\mu_k|}{|\mu_k - \mu_j|} \right)^{-1} \,. \tag{11}$$

If at least one eigenvalue of A is complex, the equality (11) does not hold in general, cf. [51]. Nevertheless, in [51] we formulate a conjecture, supported by numerical experiments and by some theoretical results, that there exist a set of $n+1$ eigenvalues such that the value on the right hand site of (11) is equal to (10) up to a factor between 1 and $4/\pi$.

Of course, except for model problems and special situations, not all eigenvalues of A are known, and hence an analysis based on (11) cannot be applied. In the following we will concentrate on the practically more relevant approach to estimate the value of (10) using only a partial knowledge of the spectrum, in particular only some set that contains all the eigenvalues (a so-called inclusion set). An inclusion set is often known a priori or can be easily estimated. We discuss the resulting convergence bounds for CG (hermitian positive definite A), MINRES (hermitian A) and GMRES (general normal A).

3.1.1 Convergence analysis for CG

Consider a *hermitian positive definite* matrix A. Each such matrix defines a norm (the so-called A-norm),

$$\|u\|_A \;=\; \left(u^H A u\right)^{\frac{1}{2}} \,, \tag{12}$$

and it is well known (see, e.g., [27]) that the OR Krylov subspace iterates x_n are in this case uniquely defined in each iterative step n and can be computed using the CG method. The CG iterates x_n satisfy

$$\|x - x_n\|_A \;=\; \min_{p\in\pi_n} \|p(A)(x - x_0)\|_A \,. \tag{13}$$

In other words, the CG method constructs an approximation x_n from the affine subspace $x_0 + \mathcal{K}_n(A, r_0)$ with minimal A-norm of the error. It can be shown that the A-norm of the error is strictly monotonically decreasing, i.e., that $\|x - x_n\|_A < \|x - x_{n-1}\|_A$ for $n = 1, \dots, d$. The A-norm of the error often has a counterpart in the underlying real-world problem. For example, when the linear system comes from finite element approximations of self-adjoint elliptic PDEs, then the A-norm of the error can be interpreted as the discretized measure of energy which is to be minimized; see, e.g., [1, 2].

A simple algebraic manipulation shows that the value (10) represents an upper bound on the relative A-norm of the error,

$$\frac{\|x - x_n\|_A}{\|x - x_0\|_A} \;\le\; \min_{p\in\pi_n} \max_k |p(\lambda_k)| \,. \tag{14}$$

This convergence bound is sharp, i.e., for each iteration step n there exist a right hand side b or an initial guess x_0 (depending on n and A) such that equality holds in (14), see [26]. In this sense, the bound (14) completely describes the *worst-case behavior* of the CG method. When the whole spectrum of A is known, one can try to determine the value of the right hand side of (14) using the formula (11). However, it is in general an open problem which subset of $n+1$ eigenvalues leads to equality in (11).

Obviously, the bound (14) depends only on the matrix eigenvalues and not on any other properties of A, b, or x_0. If a particular right hand side b is known, it is sometimes possible to incorporate the information about b into the analysis, and thus to obtain a better estimate of the actual convergence behavior.

Estimating the bound (14). Often, the largest and smallest eigenvalue (or at least estimates for them) are known. Then the classical approach is to replace the discrete set of the matrix eigenvalues by an interval containing all eigenvalues and to use Chebyshev polynomials of the first kind to estimate the min-max approximation (14). This results in the following well-known upper bound based on the condition number of A, i.e. the ratio of the largest and the smallest eigenvalue (see, e.g., [27]),

$$\frac{\|x - x_n\|_A}{\|x - x_0\|_A} \leq 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^n, \qquad \kappa = \frac{\lambda_{\max}}{\lambda_{\min}}. \tag{15}$$

We stress that there is a principal difference between the bounds (14) and (15). The bound (14) represents a min-max approximation problem on the *discrete set* $\lambda_1, \ldots, \lambda_N$, and it describes the convergence behavior in the worst-case sense. On the other hand, the bound (15) represents an estimate of the min-max approximation on the *interval* $[\lambda_{\min}, \lambda_{\max}]$ containing all eigenvalues of A. It therefore bounds the worst-case behavior for all possible eigenvalue distributions in the given interval. In other words, the bounds (14) and (15) describe different approximation problems, and thus their values can differ significantly. Clearly, the bound (15) cannot be identified with the CG convergence, and it represents an overestimate even of the worst-case behavior except for very special eigenvalue distributions in the given interval (see [50] for further discussion of this fact). The bound (15) shows, however, that a small condition number (close to 1) implies fast convergence of the CG method. This justifies the classical goal of “preconditioning”, namely to decrease the condition number of the given system matrix. On the other hand, the bound (15) does *not* show that a large condition number implies slow convergence of the CG method.

Example 3.1 Consider two example eigenvalue distributions in the interval $[1/400, 1]$. The first eigenvalue set, given by

$$\lambda_k = k^2/400, \quad k = 1, \ldots, 20, \tag{16}$$

has a cluster close to zero, whereas the second set, given by

$$\lambda_k = \log(k)/\log(20), \quad k = 2, \ldots, 20, \quad \lambda_1 = 1/400, \tag{17}$$

has a cluster close to one. Each hermitian and positive definite matrix having the eigenvalues (16) or (17) has the (moderate) condition number 400. Fig. 1 shows that the worst-case CG convergence behavior differs significantly for the eigenvalue set (16) (solid) and for the eigenvalue set (17) (dashed). Since the bound (15) (dash-dotted) represents an upper bound on the worst-case CG behavior for any eigenvalue distribution in the given interval, it cannot describe the actual CG convergence for a particular eigenvalue set like (17). □

An alternative estimate for the value (10), based on the ratio of arithmetic and geometric averages of the eigenvalues (the so-called K-condition number), was introduced by Kaporin [41]. This and other related estimates can also be found in [5, Chapter 13]. In [6], Axelsson and Kaporin propose convergence estimates for the CG method based on a generalized condition number of A, which also depends on the initial error.

Superlinear convergence of CG. In many applications it has been observed that the A-norm of the error in the CG method converges “superlinearly”, which means that speed of

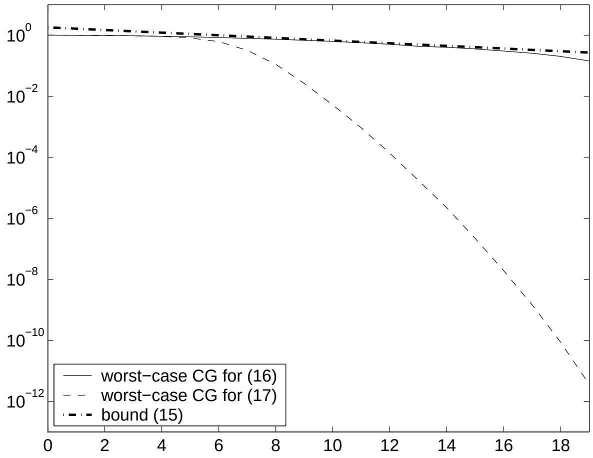


Fig. 1 For a particular eigenvalue distribution (17), the worst-case CG behavior (dashed) can significantly differ from the bound (15) (dash-dotted).

convergence increases during the iteration. Some attempts have been made to explain this behavior using the convergence of Ritz values in the Lanczos process that underlies the CG method. An intuitive explanation of the superlinear behavior, given in the early paper [11], is that when the extremal eigenvalues of A are well approximated by the Ritz values, then the CG method proceeds as if the corresponding eigenvectors were not present . This leads to a smaller "effective" condition number of A, which in turn might explain the faster convergence. This situation is discussed and analyzed, for example, in [52, 71, 69]; see [70, Chapter 5.3] for a recent summary.

The results just mentioned attempt to explain the behavior of the CG method using information that is generated during the run of the method. A different, and certainly not less interesting problem is to identify (a priori) properties of the input data A, b and x_0 that imply superlinear convergence behavior. This problem is considered in an asymptotic setting by Beckermann and Kuijlaars [7, 8]. They show that superlinear CG convergence can be observed when solving a sequence of linear systems with hermitian positive definite matrices whose eigenvalue distributions are far from an equilibrium distribution [7] (see, e.g., [22] for an introduction to these asymptotic concepts). Such favorable eigenvalue distributions occur, for example, when the system matrices come from the standard five-point finite difference discretizations of the two-dimensional Poisson equation. Another situation where superlinear convergence is observed despite an equilibrium distribution of the eigenvalues is when the components of the initial error in the eigenvector basis of the system matrices strongly vary in size [8]. In a finite dimensional setting, analytic examples for this phenomenon in the context of the discretized one-dimensional Poisson equation are given in [50].

Example 3.2 Consider the N by N tridiagonal symmetric and positive definite Toeplitz matrix $A = \mathrm{tridiag}(-1, 2, -1)$ for $N = 120$, that arises by the central finite difference approximation of the one-dimensional Poisson equation. As proved asymptotically by Beckermann and Kuijlaars [8], CG may for this model problem converge superlinearly when the initial error exhibits certain distributions of components in the eigenvector basis of A.

For particular initial errors, the superlinear convergence can in this model problem even be proved in a finite dimensional setting. In particular, consider an initial error whose components in the eigenvector basis of A are given by $\gamma \sin^{-2}(k\pi/(2N+2))$, $k = 1, \ldots, N$, where γ represents a nonzero scaling factor; cf. the solid line in the right part of Fig 2. Apparently, these components strongly vary in size, with larger components corresponding to smaller eigenvalues of A. Using the results of Naiman et al. [54], it can be shown by an elementary computation [50], that the CG errors for this initial error satisfy

$$\frac{\|x - x_n\|_A}{\|x - x_{n-1}\|_A} = \left(\frac{N-n}{N-n+3}\right)^{1/2}, \qquad n = 1, \ldots, N\,.$$

The right hand side in the above equation is a strictly decreasing function of the iteration step n, which gives an analytic proof for the superlinear CG convergence for A and this initial error. The superlinear CG convergence curve is shown as the solid line in the left part of Fig. 2. For comparison, we use an initial error with eigencomponents that are equally distributed; cf. the dashed line in the right part of Fig 2. As shown by the dashed line in the left part of Fig 2, no superlinear convergence can be observed in this case. □

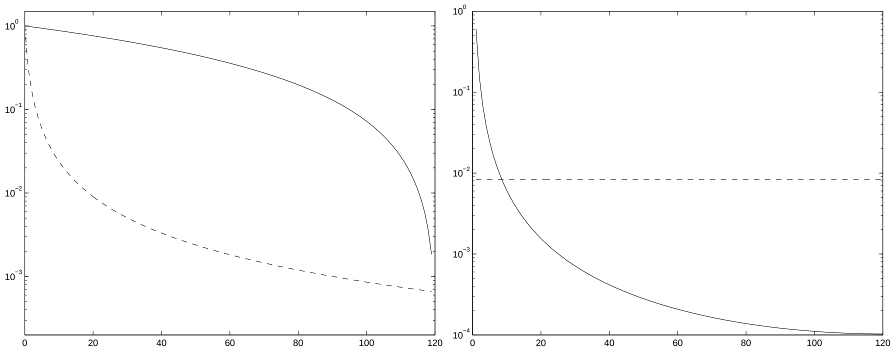


Fig. 2 CG convergence curves (left part) for two distributions of eigencomponents of the initial error (right part).

In summary, the convergence behavior of the CG method is relatively well understood, but some open problems still remain. The right approach for investigating the convergence behavior is to use all information about the eigenvalue distribution we have at our disposal. If a particular right hand side b and initial guess x_0 are given, they should be incorporated in the analysis. An example for such an approach for the model problem of the one-dimensional Poisson equation is given in [50].

3.1.2 Convergence analysis for MINRES and GMRES

In this subsection we consider nonsingular and *normal* matrices A. It is well known (see, e.g., [27]) that the iterates x_n of the MR Krylov subspace method are for any such matrix uniquely defined in each iterative step n, and that the nth residual $r_n = b - Ax_n$ satisfies

$$\|r_n\| = \min_{p\in\pi_n} \|p(A)r_0\|. \tag{18}$$

The residual norms decrease strictly monotonically whenever zero is outside the field of values of A, see [15, 33] for different proofs. However, in general no strict monotonicity is guaranteed. In fact, any (finite) nonincreasing sequence of numbers represents a convergence curve of the MR Krylov subspace residual norms applied to some linear system with a normal system matrix [3, 32, 48]. The normal matrix can even be chosen to have all its eigenvalues on the unit circle.

In the normal case, the relative residual norm of the MR Krylov subspace method can be bounded similarly as in (14),

$$\frac{\|r_n\|}{\|r_0\|} \le \min_{p\in\pi_n} \max_k |p(\lambda_k)| \tag{19}$$

and again, the bound (19) is sharp [31, 40]. In other words, the bound (19) describes the worst-case behavior of the MR Krylov subspace method. If the full spectral information is available, then the approach in [51] (cf. the discussion of formula (11)) can be used for estimating the worst-case convergence behavior. Otherwise, one can try to estimate the worst-case bound (19) similarly as in the hermitian positive definite case, i.e., by replacing the discrete spectrum by a continuous inclusion set. However, as we will see, the estimation of the min-max approximation becomes much more complicated now.

The hermitian indefinite case. When A is hermitian indefinite, the MR Krylov subspace method MINRES can be used. An estimate on the min-max approximation (19) that represents the worst-case MINRES convergence behavior, can be obtained by replacing the discrete set of the eigenvalues by the union of two intervals containing all of them and *excluding the origin*, say $I^- \cup I^+ \equiv [\lambda_{\min}, \lambda_s] \cup [\lambda_{s+1}, \lambda_{\max}]$ with $\lambda_{\min} \le \lambda_s < 0 < \lambda_{s+1} \le \lambda_{\max}$. Note that if zero would be contained in the inclusion set $I^- \cup I^+$, then the optimal min-max polynomial from π_n on this set would be the constant polynomial $p_n(z) = 1$ for all n, and the resulting convergence bounds would be useless.

When both intervals are of the same length, i.e., $\lambda_{\max} - \lambda_{s+1} = \lambda_s - \lambda_{\min}$, the following bound for the min-max value can be found,

$$\min_{p\in\pi_n} \max_k |p(\lambda_k)| \le \min_{p\in\pi_n} \max_{z\in I^-\cup I^+} |p(z)| \tag{20}$$

$$\le 2\left(\frac{\sqrt{|\lambda_{\min}\lambda_{\max}|} - \sqrt{|\lambda_s\lambda_{s+1}|}}{\sqrt{|\lambda_{\min}\lambda_{\max}|} + \sqrt{|\lambda_s\lambda_{s+1}|}}\right)^{[k/2]}, \tag{21}$$

where $[k/2]$ denotes the integer part of $k/2$, see [27, Chapter 3]. For an illustration of this bound suppose that $|\lambda_{\min}| = \lambda_{\max} = 1$ and $|\lambda_s| = \lambda_{s+1}$. Then the condition number of A is $\kappa = \lambda_{s+1}^{-1}$, and the right hand side of (21) reduces to

$$2\left(\frac{\kappa - 1}{\kappa + 1}\right)^{[k/2]}. \tag{22}$$

Apparently, (22) corresponds to the value of right hand side of (15) at step $[k/2]$ for a hermitian positive definite matrix having all its eigenvalues in the interval $[\lambda_{s+1}^2, 1]$, and thus a condition number of λ_{s+1}^{-2}. Hence the convergence bound for an indefinite matrix with condition number κ needs twice as many steps to decrease to the value of the bound for a definite matrix with condition number κ^2. Although neither of the two bounds is sharp, this clearly indicates that solving indefinite problems represents a significant challenge. In the general case when the two intervals I^- and I^+ are not of the same length, the explicit solution of the min-max approximation problem on $I^- \cup I^+$ becomes quite complicated, see, e.g., [22, Chapter 3], and no simple and explicit bound on the min-max value is known. One may of course extend the smaller interval to match the length of the larger one, and still apply (21). But this usually results in a significantly weaker convergence bound, which fails to give relevant information about the actual convergence behavior. Similar as in the case of the CG method we stress that there is a principal difference between the bounds (19) and (21). These bounds describe different approximation problems, and thus their values can differ significantly.

The general normal case. If A is a general normal matrix, the MR Krylov subspace method GMRES can be used. Again, an estimate of the right hand side of (19) can be obtained by replacing the discrete set of the eigenvalues of A by some (compact) inclusion set $\Omega \subset \mathbb{C}$ on which (nearly) optimal polynomials are explicitly known. Usually one works with connected inclusion sets, since polynomial approximation on disconnected sets is not well understood (even in the case of two disjoint intervals; see above). Because of the normalization of the polynomials at zero, the set Ω should not include the origin.

The simplest result is obtained when the spectrum of A is contained in a disk in the complex plane (that excludes the origin), say with radius $r > 0$ and center at $c \in \mathbb{C}$. Then the polynomial $p_n(z) = ((c-z)/c)^n \in \pi_n$ can be used to show that

$$\min_{p \in \pi_n} \max_k |p(\lambda_k)| \leq \left| \frac{r}{c} \right|^n .$$

In particular, a disk of small radius that is far from the origin guarantees fast convergence of the GMRES residual norms.

More refined bounds can be obtained using the convex hull $\mathcal{E}$ of an ellipse instead of a disk. For example, suppose that the spectrum is contained in an ellipse with center at $c \in \mathbb{R}$, focal distance $d > 0$ and major semi axis $a > 0$. If $0 \notin \mathcal{E}$, it can be shown that

$$\min_{p \in \pi_n} \max_k |p(\lambda_k)| \leq \frac{C_n(a/d)}{|C_n(c/d)|} \approx \left(\frac{a + \sqrt{a^2 - d^2}}{c + \sqrt{c^2 - d^2}} \right)^n ,$$

where $C_n(z)$ denotes the nth complex Chebyshev polynomial, see, e.g., [59]. We remark that, as shown by Fischer and Freund [23], the polynomials $C_n(z)/C_n(0)$ are in general not the optimal min-max polynomials from π_n on $\mathcal{E}$. However, these polynomials are asymptotically optimal and hence predict the correct rate of convergence of the min-max approximation problem on $\mathcal{E}$. For more details we refer to [61].

Of course, one would like to find a set Ω in the complex plane that yields the smallest possible upper bound on the right hand side of (19). Both a disk and the convex hull of an ellipse are convex, so one can probably improve the convergence bound by using the smallest convex set containing all the eigenvalues, i.e., the convex hull of the eigenvalues. Since A is assumed normal, this set coincides with the field of values $\mathcal{F}(A)$. Hence the bound (28) studied below in the context of nonnormal matrices can in principle be used in the normal

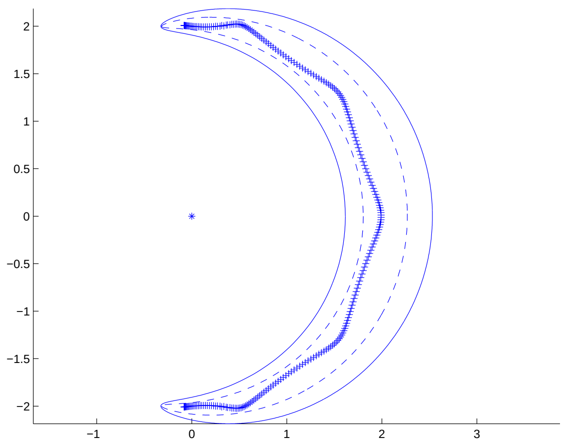


Fig. 3 Tight inclusion of the eigenvalues of the GRCAR matrix by two elements of the class of sets introduced in [43, 47].

case as well. However, all convex inclusion sets Ω are limited in their applicability by the strict requirement that $0 \notin \Omega$. In particular, if zero is inside the convex hull of the eigenvalues of A, then no convex inclusion set for these points can be used. Moreover, if the convex hull is close to the origin, then any bound derived from this set will be poor, regardless of the distance of the eigenvalues to the origin. Another difficulty with using the convex hull of the eigenvalues (or any other inclusion set bounded by a polygon) is that the boundary of this set in not smooth and hence the computation of (nearly) optimal polynomials on these sets such as the Faber polynomials is complicated, see, e.g., [64].

To overcome such difficulties, a parameterized class of non-convex sets with analytic boundaries is constructed in [47] (also see [43]), for which the Faber polynomials are explicitly known. These polynomials give rise to analytic and easily computable bounds for the min-max approximation problem; see [47] for details. Two examples of the inclusion sets are show in Fig. 3. The plus signs in this figure show the eigenvalues of the so-called Grcar matrix of order 250, generated by the MATLAB command `gallery('grcar',250,6)`. Obviously, the convex hull of these eigenvalues contains the origin (indicated by the star). On the other hand, none of the eigenvalues is particularly close to the origin, which should be exploited by the choice of the inclusion set. The boundaries of the two example inclusion sets are shown by the dashed and the solid curves.

3.2 Convergence analysis for nonnormal matrices (GMRES)

In this section we consider the case of a general nonsingular and *nonnormal* matrix A. In this general case, an MR Krylov subspace method such as GMRES yields uniquely defined iterates x_n so that the nth residual $r_n = b - Ax_n$ satisfies

$$\|r_n\| = \min_{p \in \pi_n} \|p(A)r_0\|. \tag{23}$$

Similarly to the convergence analysis for normal matrices presented above, we are interested in finding a (sharp) bound on the right hand side of (23).

Eigenvalues and convergence. If A is diagonalizable, $A = V\Lambda V^{-1}$, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$, then the following convergence bound easily follows from (23),

$$\frac{\|r_n\|}{\|r_0\|} = \min_{p\in\pi_n} \frac{\|Vp(\Lambda)V^{-1}r_0\|}{\|r_0\|} \leq \kappa(V) \min_{p\in\pi_n} \max_{k} |p(\lambda_k)| \,, \tag{24}$$

see, e.g., [62]. Here $\kappa(V) = \|V\|\,\|V^{-1}\|$ denotes the condition number of the eigenvector matrix V. A bound similar to (24) can be derived for nondiagonalizable matrices.

The bound (24) frequently is the basis for discussions of the GMRES convergence behavior. As mentioned in Section 3.1.2, this bound is sharp when A is normal. When $\kappa(V)$ is small, the right hand side of (24) typically represents a good convergence bound, and its value can be estimated using the tools described above. However, when V is far from unitary, the bound (24) may fail to provide any reasonable information about the GMRES convergence. To see this, note that when the eigenvector matrix V is ill-conditioned, then some components of the vector $V^{-1}r_0$ can be very large, potentially much larger than $\|r_0\|$. On the other hand, $\|r_n\|$ in (24) is bounded from above by $\|r_0\|$. Therefore, the linear combination $V\,[p(\Lambda)V^{-1}r_0]$ can contain a significant cancellation, which is not reflected in the minimization problem on the right hand side of (24). Apart from the fact, that the factor $\kappa(V)$ can be very large in case of ill-conditioned eigenvectors, the principal weakness of the bound (24) is that the min-max problem on the matrix eigenvalues need not have any connection with the GMRES convergence for the given nonnormal matrix. As a consequence, the curve produced by the min-max approximations on matrix eigenvalues can be substantially different from the (worst-case) GMRES convergence curve and the bound can fail to give any reasonable convergence information.

Example 3.3 For a numerical illustration consider the two N by N tridiagonal Toeplitz matrices

$$A_\lambda = \mathrm{tridiag}(-1, \lambda, -1) \quad \text{and} \quad B_\lambda = \mathrm{tridiag}(-\lambda, \lambda, -1/\lambda)\,,$$

where $\lambda \geq 2$ is a real parameter. Both A_λ and B_λ have *the same eigenvalues*, namely $\lambda - 2\cos(k\pi/(N+1))$, $k = 1, \ldots, N$. While A_λ is symmetric positive definite, B_λ is highly nonnormal (e.g. a MATLAB computation yields $\kappa(V) \approx 10^{27}$ for $N = 40$ and $\lambda = 3$). The relative GMRES residual norms for $x_0 = 0$ and the order 40 systems $A_\lambda x = [1, 0, \ldots, 0]^T$ and $B_\lambda x = [1, 0, \ldots, 0]^T$, for $\lambda = 3, 4, \ldots, 18$, are shown in Fig. 4. The relative residual norms for the systems with A_λ are plotted by solid lines (faster convergence corresponds to larger λ), and for the systems with B_λ they are plotted by dashed lines (essentially the same for all λ). We observe that the GMRES convergence speed for A_λ increases when the spectrum moves away from the origin. On the other hand, for B_λ spectral information is obviously useless for describing the GMRES convergence. In this example essentially nothing happens during the first $N - 1$ steps, and then termination occurs in the final step N. Moreover, the spectrum of B_λ gives no information about the convergence behavior after some "transient delay", which some authors attribute to the potentially large constant $\kappa(V)$ in (24). □

The above example for the matrices B_λ clearly shows that in the nonnormal case eigenvalue information is not sufficient for describing the convergence behavior of GMRES. In

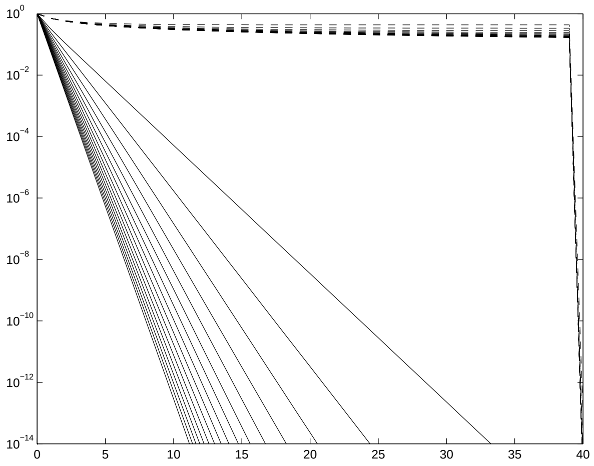


Fig. 4 Relative GMRES residual norms for the normal matrices A_λ (solid) and the nonnormal matrices B_λ (dashed) for $\lambda = 3, 4, \ldots, 18$ and $r_0 = [1, 0, \ldots, 0]^T$.

fact, in this case the eigenvalues may have nothing to do with the convergence behavior at all. As shown in [3, 32], any nonincreasing convergence curve of relative GMRES residual norms is attainable for a system matrix A having any prescribed eigenvalues. On the other hand, it needs to be stressed that from an analytic point of view the principal difficulty of nonnormality is *not* the often met belief that the convergence is slower for nonnormal than for normal matrices. This belief is incorrect because for each nonnormal matrix A there exists a normal matrix B for which the same convergence behavior can be observed (for the same initial residual r_0), cf. [33, 48]. Unfortunately, the mapping from the matrix A to the normal matrix B is highly nonlinear, and it depends strongly on r_0. Hence it is not suitable for an a priori analysis of the GMRES convergence behavior for the given A and r_0.

The idea to analyze the given nonnormal problem using a related normal problem is also used by Huhtanen and Nevanlinna [39]. They propose to split the matrix A into $A = \tilde{A} + E$, where $\tilde{A}$ is normal and E is of smallest possible rank. Using such splitting, lower bounds for the quantity $\min_{p \in \pi_n} \|p(A)\|$ (cf. (26) below) can be given in terms of certain eigenvalues of $\tilde{A}$; see [39] for details.

Worst-case GMRES analysis in the nonnormal case. It should be clear by now that in the nonnormal case the GMRES convergence behavior is *significantly more difficult to analyze* than in the normal case. A general approach to understand the worst-case GMRES convergence in the nonnormal case is to replace the complicated minimization problem (23) by another one that is easier to analyze and that, in some sense, approximates the original problem (23). Natural bounds on the GMRES residual norm arise by excluding the influence of the initial residual r_0,

$$\begin{aligned}
\frac{\|r_n\|}{\|r_0\|} &= \min_{p\in\pi_n} \frac{\|p(A)r_0\|}{\|r_0\|} && \text{(GMRES)} \\
&\leq \max_{\|v\|=1} \min_{p\in\pi_n} \|p(A)v\| && \text{(worst-case GMRES)} \qquad (25)\\
&\leq \min_{p\in\pi_n} \|p(A)\| && \text{(ideal GMRES).} \qquad (26)
\end{aligned}$$

The bound (25) corresponds to the *worst-case* GMRES behavior and represents a sharp upper bound, i.e. a bound that is attainable by the GMRES residual norm. In this sense, (25) is the best bound on $\|r_n\|/\|r_0\|$ that is independent of r_0. Despite the independence of r_0, it is not clear in general, which properties of A influence the bound (25); see, e.g., [20]. The expression (25) can be bounded by the *ideal* GMRES approximation problem (26), which was introduced by Greenbaum and Trefethen [34]. To justify the relevance of the bound (26), several researchers tried to identify cases in which (25) is equal to (26). The best known result of this type is that (25) is equal to (26) whenever A is normal [31, 40]. Despite the existence of some counterexamples [20, 67], it is still an open question whether (25) is equal or close to (26) for larger classes of nonnormal matrices. In [66] we consider this problem for a Jordan block, a representative of a nonnormal matrix, and prove equality of the expressions (25) and (26) in some steps.

A possible way to approximate the value of the matrix approximation problem (26) is to determine sets $\Omega \subset \mathbb{C}$ and $\hat{\Omega} \subset \mathbb{C}$, that are somehow associated with A, and that provide lower and upper bounds on (26),

$$c_1 \min_{p\in\pi_n} \max_{z\in\Omega} |p(z)| \;\leq\; \min_{p\in\pi_n} \|p(A)\| \;\leq\; c_2 \min_{p\in\pi_n} \max_{z\in\hat{\Omega}} |p(z)|.$$

Here c_1 and c_2 should be some (moderate size) constants depending on A and possibly on n. This approach represents a generalization of the idea for normal matrices, where the appropriate set associated with A is the spectrum of A and $c_1 = c_2 = 1$.

Trefethen [68] has suggested taking $\hat{\Omega}$ to be the *ϵ-pseudospectrum* of A,

$$\Lambda_\epsilon(A) = \left\{ z \in \mathbb{C} \,:\, \|(zI - A)^{-1}\| \geq \epsilon^{-1} \right\} .$$

Denoting by L the arc length of the boundary of $\Lambda_\epsilon(A)$, the following bound can be derived,

$$\min_{p\in\pi_n} \|p(A)\| \leq \frac{L}{2\pi\epsilon} \min_{p\in\pi_n} \max_{z\in\Lambda_\epsilon(A)} \|p(z)\| , \tag{27}$$

see, e.g., [53]. The parameter ϵ gives some flexibility, but choosing a good value can be tricky. Note that in order to make the right hand side of (27) reasonably small, one must choose ϵ large enough to make the constant $L/2\pi\epsilon$ small, but small enough to make the set $\Lambda_\epsilon(A)$ not too large. The bound (27) works well in some situations (see, e.g., [17]), but it is easy to construct examples for which no choice of ϵ gives a tight estimate of the ideal GMRES approximation problem (see, e.g., [33]).

Another approach is based on the *field of values* of A, cf. (7). Denote by $\nu(\mathcal{F}(A))$ the distance of $\mathcal{F}(A)$ from the origin, $\nu(\mathcal{F}(A)) = \min_{z\in\mathcal{F}(A)} |z|$, then

$$\min_{p\in\pi_n} \|p(A)\| \leq \left(1 - \nu(\mathcal{F}(A))\nu(\mathcal{F}(A^{-1}))\right)^{n/2} , \tag{28}$$

see, e.g., [14]. Suppose that $M = (A + A^H)/2$, the hermitian part of A, is positive definite. Then a special case of (28) is

$$\min_{p \in \pi_n} \|p(A)\| \leq \left(1 - \frac{\lambda_{\min}(M)}{\lambda_{\max}(A^H A)}\right)^{n/2} ,$$

which is one of the earliest convergence results for the MR Krylov subspace method [15, 16]. Since $\mathcal{F}(A)$ is a convex set that contains the convex hull of the eigenvalues of A, the requirement $0 \notin \mathcal{F}(A)$ makes the bound (28) useless in many situations. However, the field of values analysis can be very useful when the given linear system comes from the discretization of elliptic PDEs by the Galerkin finite element method. In such cases the coefficients of the N by N system matrix A are given by $A_{ij} = a(\phi_i, \phi_j)$, where $a(u, v)$ is the bilinear form from the weak formulation of the PDE, and $\phi_1, \ldots, \phi_N$ are the nodal basis functions. Let V_h denote the finite element space. Then a function $u_h \in V_h$ is represented by a vector $u_N \in \mathbb{R}^N$ that contains the values of u_h at the nodes of the triangulation. The matrix A satisfies $u_N^T A v_N = a(u_h, v_h)$ for all $u_h, v_h \in V_h$. These relations can be exploited to give bounds for the quantity $a(x - x_n, x - x_n) = (x - x_n)^T A(x - x_n)$, where x is the exact solution of the discretized PDE, and x_n is a Krylov subspace iterate. This leads naturally to bounds of the type (28) involving the smallest real parts of $\mathcal{F}(A)$ and $\mathcal{F}(A^{-1})$; see, e.g., [42, 63] for more details. Note that under the usual assumption that the bilinear form is coercive, the smallest real parts of $\mathcal{F}(A)$ and $\mathcal{F}(A^{-1})$ are both positive. In a more abstract setting, the field of values has been used in the convergence analysis by Eiermann [13].

A generalization of the field of values of A is the *polynomial numerical hull*, introduced by Nevanlinna [55], and defined as

$$\mathcal{H}_n(A) = \{z \in \mathbb{C} \, : \, \|p(A)\| \geq |p(z)| \text{ for all } p \in \mathcal{P}_n\} ,$$

where $\mathcal{P}_n$ denotes the set of polynomials of degree n or less. It can be shown that $\mathcal{F}(A) = \mathcal{H}_1(A)$. The set $\mathcal{H}_n(A)$ provides a lower bound on (26),

$$\min_{p \in \pi_n} \max_{z \in \mathcal{H}_n(A)} |p(z)| \, \leq \, \min_{p \in \pi_n} \|p(A)\|. \tag{29}$$

In some way, $\mathcal{H}_n(A)$ reflects the complicated relation between the polynomial of degree n and the matrix A, and provides often a very good estimate of the value of the ideal GMRES approximation (26). Greenbaum and her co-workers [19, 28, 29, 30] have obtained theoretical results about $\mathcal{H}_n(A)$ for Jordan blocks, banded triangular Toeplitz matrices and block diagonal matrices with triangular Toeplitz blocks. Clearly, for a larger applicability of the bound (29), the class of matrices for which $\mathcal{H}_n(A)$ is known needs to be extended. But in general, the determination of these sets represents a nontrivial open problem.

The bounds stated above are certainly useful to obtain a priori convergence estimates in terms of properties of A, and possibly to analyze the effectiveness of preconditioning techniques. However, the worst-case behavior of GMRES for nonnormal matrices is still not well understood. We again point out that the bound (26) is not sharp, and that it is in many situations unclear how closely the ideal GMRES approximates the worst-case GMRES. Moreover, none of the bounds stated above is able to *characterize* satisfactorily in terms of matrix properties, which approximation problem is solved by the worst-case GMRES in the nonnormal case.

The influence of the initial residual: A model problem. Users of Krylov subspace methods usually want to solve a particular linear system, and hence a worst-case analysis may be of lesser interest to them. In such context one needs to understand also how the convergence is influenced by the particular right hand side or initial residual r_0. It seems to be well known that the initial residual may have a significant influence on the GMRES convergence, in particular in the nonnormal case. However, no systematic study of this influence exists, and given the lack of understanding of even the worst-case behavior, it is unlikely that a complete understanding of the influence of r_0 on the convergence will be obtained in the near future.

In the context of discretized PDEs, r_0 is directly related to the boundary conditions and/or the source terms. It is of great importance to understand how such PDE data influences the convergence of an iterative solver like GMRES, as understanding of these relations will pave the way to efficient preconditioning techniques. Recently, this topic was addressed in an analysis of the GMRES convergence behavior for a well known convection-diffusion model problem [49], that was introduced in [24]. Here the convergence of GMRES applied to the discretized system is characterized by an initial phase of slow convergence, followed by a faster decrease of the residual norms. The length of the initial phase depends on the initial residual, which is determined by the boundary conditions (for simplicity, the source term in the PDE and the initial guess x_0 are chosen equal to zero in [49]). Typical examples for the convergence behavior are shown in Fig. 5. The GMRES convergence curves in this figure correspond to the same discretized operator but to different boundary conditions. For the considered model problem, the convergence analysis confirms an earlier conjecture of Ernst [18], that the duration of the initial phase is governed by the time it takes for boundary information to pass from the inflow boundary across the domain following the longest streamline of the velocity field. The paper [49] also discusses the question why the convergence in the second phase accelerates. Numerical results show that the speed of convergence after the initial delay

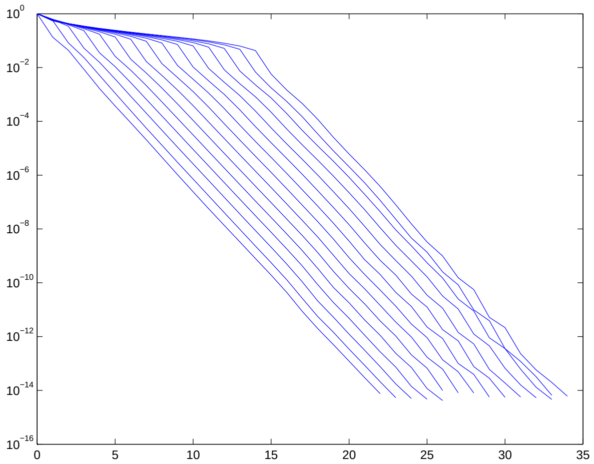


Fig. 5 Relative GMRES residual norms for the discretized convection-diffusion model problem considered in [49]. Different behavior corresponds to the same discretized operator but to different boundary conditions.

is slower for larger mesh Peclet numbers, but a complete quantitative understanding of this phenomenon remains a difficult open problem.

4 Concluding remarks

The worst-case convergence behavior of many well known Krylov subspace methods (CG, MINRES, GMRES) for normal matrices is described by the min-max approximation problem on the discrete set of the matrix eigenvalues,

$$\min_{p\in\pi_n} \max_{k} |p(\lambda_k)| \, . \tag{30}$$

In this sense, the worst-case convergence behavior is well understood. Still, for a given eigenvalue distribution the min-max value is often not known, and has to be estimated. Such estimation is of course always necessary, when only a partial information about the spectrum is known. A general approach tries to find inclusion sets for (the estimate of) the spectrum, and uses (close to) optimal polynomials on these sets to approximate the min-max value. However, this approach solves a different kind of approximation problem and can provide misleading information about the convergence.

For nonnormal matrices, the situation is even less clear. To bound the worst-case GMRES residual norm, one can use the ideal GMRES approximation

$$\min_{p\in\pi_n} \|p(A)\| \, , \tag{31}$$

that represents a matrix approximation problem. Although the value (31) need not describe GMRES worst-case behavior, it can be considered as a good approximation of the worst-case approximation in many practical cases. A general approach for approximating this value consists in finding a set in the complex plain associated with the matrix A and bounding the value (31) by the min-max approximation on this set. However, theoretical results in this field are still unsatisfactory.

Finally, it is important to note that the convergence can depend strongly on the right hand side or the initial guess so that the values (30) and (31) can overestimate the actual convergence of a Krylov subspace method.

References

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

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

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

[4] S. F. Ashby, T. A. Manteuffel, and P. E. Saylor, *A taxonomy for conjugate gradient methods*, SIAM J. Numer. Anal., 27 (1990), pp. 1542–1568.

[5] O. Axelsson, *Iterative solution methods*, Cambridge University Press, Cambridge, 1994.

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

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

[8] ———, *Superlinear CG convergence for special right-hand sides*, Electron. Trans. Numer. Anal., 14 (2002), pp. 1–19 (electronic). Orthogonal polynomials, approximation theory, and harmonic analysis (Inzel, 2000).

[9] C. G. Broyden, *A new taxonomy of conjugate gradient methods*, Comput. Math. Appl., 31 (1996), pp. 7–17. Selected topics in numerical methods (Miskolc, 1994).

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

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

[12] J. Cullum and A. Greenbaum, *Relations between Galerkin and norm-minimizing iterative methods for solving linear systems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 223–247.

[13] M. Eiermann, *Fields of values and iterative methods*, Linear Algebra Appl., 180 (1993), pp. 167–197.

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

[15] S. C. Eisenstat, H. C. Elman, and M. H. Schultz, *Variational iterative methods for nonsymmetric systems of linear equations*, SIAM J. Numer. Anal., 20 (1983), pp. 345–357.

[16] H. C. Elman, *Iterative methods for large sparse nonsymmetric systems of linear equations.*, PhD thesis, Yale University, New Haven, 1982.

[17] M. Embree, *How descriptive are GMRES convergence bounds?*, Numerical Analysis Report 99/08, Oxford University Computing Laboratory, 1999.

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

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

[20] V. Faber, W. Joubert, E. Knill, and T. Manteuffel, *Minimal residual method stronger than polynomial preconditioning*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 707–729.

[21] D. K. Faddeev and V. N. Faddeeva, *Computational methods of linear algebra*, W. H. Freeman and Co., San Francisco, 1963.

[22] B. Fischer, *Polynomial based iteration methods for symmetric linear systems*, Wiley-Teubner Series Advances in Numerical Mathematics, John Wiley & Sons Ltd., Chichester, 1996.

[23] B. Fischer and R. Freund, *Chebyshev polynomials are not always optimal*, J. Approx. Theory, 65 (1991), pp. 261–272.

[24] B. Fischer, A. Ramage, D. J. Silvester, and A. J. Wathen, *On parameter choice and iterative convergence for stabilised discretisations of advection-diffusion problems*, Computer methods in applied mechanics and engineering, 179 (1999), pp. 179–195.

[25] F. R. Gantmacher, *The theory of matrices. Vols. 1, 2*, Chelsea Publishing Co., New York, 1959.

[26] A. Greenbaum, *Comparison of splittings used with the conjugate gradient algorithm*, Numer. Math., 33 (1979), pp. 181–193.

[27] A. Greenbaum, *Iterative methods for solving linear systems*, vol. 17 of Frontiers in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

[28] ———, *Generalizations of the field of values useful in the study of polynomial functions of a matrix*, Linear Algebra Appl., 347 (2002), pp. 233–249.

[29] ———, *Card shuffling and the polynomial numerical hull of degree k*, SIAM J. Sci. Comput., 25 (2003), pp. 408–416 (electronic).

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

[31] A. Greenbaum and L. Gurvits, *Max-min properties of matrix factor norms*, SIAM J. Sci. Comput., 15 (1994), pp. 348–358. Iterative methods in numerical linear algebra (Copper Mountain Resort, CO, 1992).

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

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

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

[35] W. Hackbusch, *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.

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

[37] M. Hochbruck and C. Lubich, *Error analysis of Krylov methods in a nutshell*, SIAM J. Sci. Comput., 19 (1998), pp. 695–701 (electronic).

[38] A. S. Householder, *The theory of matrices in numerical analysis*, Dover Publications Inc., New York, 1975. Reprint of 1964 edition.

[39] M. Huhtanen and O. Nevanlinna, *Minimal decompositions and iterative methods*, Numer. Math., 86 (2000), pp. 257–281.

[40] W. Joubert, *A robust GMRES-based adaptive polynomial preconditioning algorithm for nonsymmetric linear systems*, SIAM J. Sci. Comput., 15 (1994), pp. 427–439. Iterative methods in numerical linear algebra (Copper Mountain Resort, CO, 1992).

[41] I. E. Kaporin, *New convergence results and preconditioning strategies for the conjugate gradient method*, Numer. Linear Algebra Appl., 1 (1994), pp. 179–210.

[42] A. Klawonn and G. Starke, *Block triangular preconditioners for nonsymmetric saddle point problems: field-of-values analysis*, Numer. Math., 81 (1999), pp. 577–594.

[43] T. Koch and J. Liesen, *The conformal "bratwurst" maps and associated Faber polynomials*, Numer. Math., 86 (2000), pp. 173–191.

[44] A. N. Krylov, *On the numerical solution of the equation by which the frequency of small oscillations is determined in technical problems*, Izv. Akad. Nauk SSSR Ser. Fiz.-Mat., 4 (1931), pp. 491–539. (Title translation as given in [25]).

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

[46] ———, *Solution of systems of linear equations by minimized iterations*, J. Research Nat. Bur. Standards, 49 (1952), pp. 33–53.

[47] J. Liesen, *Construction and analysis of polynomial iterative methods for non-hermitian systems of linear equations*, PhD thesis, Fakultät für Mathematik, Universität Bielefeld, 1998. http://archiv.ub.uni-bielefeld.de/disshabi/mathe.htm.

[48] J. Liesen, *Computable convergence bounds for GMRES*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 882–903 (electronic).

[49] J. Liesen and Z. Strakoš, *GMRES convergence analysis for a convection-diffusion model problem*, SIAM J. Sci. Comput., (to appear).

[50] J. Liesen and P. Tichý, *Behavior of CG and MINRES for symmetric tridiagonal Toeplitz matrices*, Preprint 34-2004, Institute of Mathematics, Technical University of Berlin, 2004.

[51] J. Liesen and P. Tichý, *The worst-case GMRES for normal matrices*, BIT Numerical Mathematics, 44 (2004), pp. 79–98.

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

[53] N. M. NACHTIGAL, S. C. REDDY, AND L. N. TREFETHEN, *How fast are nonsymmetric matrix iterations?*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 778–795. Iterative methods in numerical linear algebra (Copper Mountain, CO, 1990).

[54] A. E. NAIMAN, I. M. BABUŠKA, AND H. C. ELMAN, *A note on conjugate gradient convergence*, Numer. Math., 76 (1997), pp. 209–230.

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

[56] C. C. PAIGE AND M. A. SAUNDERS, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.

[57] A. QUARTERONI AND A. VALLI, *Numerical approximation of partial differential equations*, vol. 23 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 1994.

[58] J. K. REID, *On the method of conjugate gradients for the solution of large sparse systems of linear equations*, in Large sparse sets of linear equations (Proc. Conf., St. Catherine's Coll., Oxford, 1970), Academic Press, London, 1971, pp. 231–254.

[59] Y. SAAD, *Krylov subspace methods for solving large unsymmetric linear systems*, Math. Comp., 37 (1981), pp. 105–126.

[60] ———, *The Lanczos biorthogonalization algorithm and other oblique projection methods for solving large unsymmetric systems*, SIAM J. Numer. Anal., 19 (1982), pp. 485–506.

[61] Y. SAAD, *Iterative methods for sparse linear systems*, Society for Industrial and Applied Mathematics, Philadelphia, PA, second ed., 2003.

[62] Y. SAAD AND M. H. SCHULTZ, *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[63] G. STARKE, *Field-of-values analysis of preconditioned iterative methods for nonsymmetric elliptic problems*, Numer. Math., 78 (1997), pp. 103–117.

[64] G. STARKE AND R. S. VARGA, *A hybrid Arnoldi-Faber iterative method for nonsymmetric systems of linear equations*, Numer. Math., 64 (1993), pp. 213–240.

[65] Z. STRAKOŠ AND J. LIESEN, *On numerical stability in large scale linear algebraic computations*, Z. Angew. Math. Mech., (submitted).

[66] P. TICHÝ AND J. LIESEN, *Worst-case and ideal GMRES for a Jordan block*, submitted to Linear Algebra Appl., (2004).

[67] K.-C. TOH, *GMRES vs. ideal GMRES*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 30–36.

[68] L. N. TREFETHEN, *Approximation theory and numerical linear algebra*, in Algorithms for approximation, II (Shrivenham, 1988), Chapman and Hall, London, 1990, pp. 336–360.

[69] A. VAN DER SLUIS AND H. A. VAN DER VORST, *The rate of convergence of conjugate gradients*, Numer. Math., 48 (1986), pp. 543–560.

[70] H. A. VAN DER VORST, *Iterative Krylov methods for large linear systems*, vol. 13 of Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge, 2003.

[71] R. WINTHER, *Some superlinear convergence results for the conjugate gradient method*, SIAM J. Numer. Anal., 17 (1980), pp. 14–17.

Electronic Transactions on Numerical Analysis.
Volume 20, pp. 180-197, 2005.
Copyright © 2005, Kent State University.
ISSN 1068-9613.

ETNA
Kent State University
etna@mcs.kent.edu

ON THE WORST-CASE CONVERGENCE OF MR AND CG FOR SYMMETRIC POSITIVE DEFINITE TRIDIAGONAL TOEPLITZ MATRICES[*]

JÖRG LIESEN[†] AND PETR TICHÝ[†]

Abstract. We study the convergence of the minimal residual (MR) and the conjugate gradient (CG) method when applied to linear algebraic systems with symmetric positive definite tridiagonal Toeplitz matrices. Such systems arise, for example, from the discretization of one-dimensional reaction-diffusion equations with Dirichlet boundary conditions. Based on our previous results in [J. Liesen and P. Tichý, *BIT*, 44 (2004), pp. 79–98], we concentrate on the next-to-last iteration step, and determine the initial residuals and initial errors for the MR and CG method, respectively, that lead to the slowest possible convergence. By this we mean that the methods have made the least possible progress in the next-to-last iteration step. Using these worst-case initial vectors, we discuss which source term and boundary condition in the underlying reaction-diffusion equation are the worst in the sense that they lead to the worst-case initial vectors for the MR and CG methods. Moreover, we determine (or very tightly estimate) the worst-case convergence quantities in the next-to-last step, and compare these to the convergence quantities obtained from average (or unbiased) initial vectors. The spectral structure of the considered matrices allows us to apply our worst-case results for the next-to-last step to derive worst-case bounds also for other iteration steps. We present a comparison of the worst-case convergence quantities with the classical convergence bound based on the condition number of A, and finally we discuss the MR and CG convergence for the special case of the one-dimensional Poisson equation with Dirichlet boundary conditions.

Key words. Krylov subspace methods, conjugate gradient method, minimal residual method, convergence analysis, tridiagonal Toeplitz matrices, Poisson equation

AMS subject classifications. 15A09, 65F10, 65F20

1. Introduction.

This paper is concerned with the convergence analysis of Krylov subspace methods for solving linear algebraic systems of the form

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

with a *symmetric positive definite* matrix $A \in \mathbb{R}^{n\times n}$, and a right hand side vector $b \in \mathbb{R}^n$. We obviously assume $n > 1$. Starting from an initial guess x_0, Krylov subspace methods compute the initial residual $r_0 = b - Ax_0$, and a sequence of approximate solutions (iterates) $x_1, x_2, \ldots$, such that the ith residual $r_i = b - Ax_i$ and the ith error $e_i = x - x_i$ are of the form

$$r_i = p_i(A)r_0\,, \quad e_i = p_i(A)e_0\,, \quad p_i \in \pi_i\,,$$

where π_i denotes the set of polynomials of degree at most i and with value one at the origin. Two choices of conditions for determining the polynomials p_i have emerged as de facto standards.

In the minimal residual (MR) Krylov subspace method, the polynomial is chosen so that the Euclidean norm ($\|y\| = (y^Ty)^{1/2}$) of the residuals is minimized,

$$\|r_i\| = \min_{p\in\pi_i} \|p(A)r_0\| \qquad \text{(MR)}. \tag{1.2}$$

There are several algorithms for implementing the MR method that try to exploit as much as possible from the properties of A. Examples are the conjugate residual (CR) method [18]

[*] Received January 15, 2005. Accepted for publication July 11, 2005. Recommended by D. Szyld.

[†] Institute of Mathematics, Technical University of Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany. (liesen@math.tu-berlin.de and tichy@math.tu-berlin.de). This research was supported by the Emmy Noether-Programm of the Deutsche Forschungsgemeinschaft. The work of the second author was also supported by the Grant Agency of Academy of Sciences of the Czech Republic under grant No. KJB1030306.

for symmetric positive definite A, the minimal residual (MINRES) method [17] symmetric and nonsingular A, and the generalized minimal residual (GMRES) method [19] for general nonsingular A.

In the orthogonal residual Krylov subspace method, the ith iterate x_i is determined such that the ith residual r_i is orthogonal to all previous residuals $r_0, \ldots, r_{i-1}$. A particular implementation for symmetric positive definite matrices A is the conjugate gradient (CG) method [8]. The symmetric positive definite matrix A defines a norm (A-norm, $\|y\|_A = (y^T A y)^{1/2}$) in which the errors are minimized,

$$\|e_i\|_A \;=\; \min_{p \in \pi_i} \|p(A) e_0\|_A \qquad \text{(CG).} \tag{1.3}$$

The standard approach to analyze (1.2) and (1.3) is to exclude the influence of r_0 and e_0, and hence to consider the *worst-case convergence* instead of the convergence for the particular initial vectors. It is well known [4, 6, 9] that the (attainable) worst-case convergence quantities are given by

$$\max_{r_0 \neq 0} \min_{p \in \pi_i} \frac{\|p(A) r_0\|}{\|r_0\|} \;=\; \max_{e_0 \neq 0} \min_{p \in \pi_i} \frac{\|p(A) e_0\|_A}{\|e_0\|_A} \;=\; \min_{p \in \pi_i} \max_k |p(\lambda_k)|\,, \tag{1.4}$$

where λ_k, $k = 1, \ldots, n$, are the eigenvalues of A. The rightmost term in (1.4) depends in a nonlinear way on the eigenvalue distribution, and no explicit solution for this min-max approximation problem is known in general. Therefore, to analyze the worst-case convergence of the MR and CG methods one needs to *estimate* this min-max value. Such estimation can be based either on a suitable superset of the eigenvalues, or a suitable subset, where the first choice leads to an upper and the second to a lower bound on the worst-case convergence.

The standard choice of a superset of the discrete set of matrix eigenvalues is their convex hull $[\lambda_{\min}, \lambda_{\max}]$. Using scaled and shifted Chebyshev polynomials of the first kind on this interval, one can show the classical bound

$$\min_{p \in \pi_i} \max_k |p(\lambda_k)| \;\leq\; 2 \left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \right)^i , \tag{1.5}$$

where $\kappa(A) = \lambda_{\max}/\lambda_{\min}$ is the condition number of A; see, e.g., [5, Theorem 3.1.1]. Because of (1.4), the term on the right hand side of (1.5) represents a bound on the relative residual norm $\|r_i\|/\|r_0\|$ for MR and the relative A-norm of the error $\|e_i\|_A/\|e_0\|_A$ for CG for each initial residual r_0 and each initial error e_0, respectively. The bound (1.5) is particularly useful in practical applications when only partial information about the spectrum of A is available or can be estimated. But one should be aware that this bound is obtained from a different kind of approximation problem than the one solved by the MR and CG methods (worst-case rather than for a specific r_0 or e_0, and continuous rather than discrete), and hence that it might provide misleading information about the actual convergence of these methods; see [12] for more details and references.

To obtain a lower bound on the worst-case convergence one can in principle consider any subset of the eigenvalues. As shown in [4, 13], for each subset of exactly $i+1$ distinct eigenvalues $\{\mu_1, \ldots, \mu_{i+1}\} \subseteq \{\lambda_1, \ldots, \lambda_n\}$,

$$\min_{p \in \pi_i} \max_k |p(\lambda_k)| \;\geq\; \min_{p \in \pi_i} \max_k |p(\mu_k)| \;=\; \left(\sum_{j=1}^{i+1} \prod_{\substack{l=1 \\ l \neq j}}^{i+1} \frac{|\mu_l|}{|\mu_l - \mu_j|} \right)^{-1} . \tag{1.6}$$

Apparently, for each choice of $i+1$ distinct eigenvalues $\mu_1, \ldots, \mu_{i+1}$, the right hand side of (1.6) represents an *explicit* lower bound on the worst-case convergence quantities. Moreover, in our case of real eigenvalues, there exists a subset of $i+1$ eigenvalues, for which the lower bound (1.6) is attained. Therefore, if the subset of $i+1$ eigenvalues is properly chosen, one can obtain a very good convergence estimate. Since this estimate of the worst-case convergence requires precise knowledge about at least some eigenvalues of A, its main use is in the analysis of model problems, where the eigenvalues are known explicitly.

In this paper we consider such a class of model problems, namely the linear systems with symmetric positive definite tridiagonal Toeplitz matrices A. Such systems arise, for example, in the discretization of one-dimensional reaction-diffusion equations. We focus on the *slowest possible convergence* of the MR and CG methods. By this we mean the situation when the worst-case convergence quantity is attained in the next-to-last iteration step. For this step the only possible subset $\{\mu_1, \ldots, \mu_{i+1}\}$ of the eigenvalues of A to be chosen in (1.6) is the set of all distinct eigenvalues of A, so that the solution of the min-max approximation problem is known explicitly. Based on our previous results in [13], we determine the worst possible initial data, i.e. the vectors r_0^w and e_0^w leading to the slowest possible convergence of the MR and CG method, respectively. Knowing the initial vector e_0^w explicitly, we identify source terms and boundary conditions in the one-dimensional reaction-diffusion equation that yield, after discretization, the slowest possible CG convergence. We also address the identification of such data for the MR method, which appears to be considerably more complicated than for CG. Moreover, we determine (or very tightly estimate) the worst-case convergence quantities in the next-to-last step, and compare these to the convergence quantities obtained from average (or unbiased) initial residuals as well as the classical convergence bound (1.5). The spectral structure of the considered matrices allows us to apply our worst-case results for the next-to-last step to derive worst-case bounds also for other iteration steps. Finally, we consider the case of one-dimensional Poisson equation, which is a popular model problem for the convergence analysis of Krylov subspace methods, in particular of CG; see, e.g., [1, 2, 15, 16].

We point out that the convergence of GMRES for nonsymmetric tridiagonal Toeplitz matrices is studied in [10]. The results in [10] hold explicitly for the highly nonnormal case, i.e. the case when a tridiagonal Toeplitz matrix can be considered a perturbed Jordan block. Hence the results presented in this paper are neither special cases nor generalizations of the results in [10].

The paper is organized as follows. Section 2 presents basic formulas for the next-to-last MR and CG iteration step. In Section 3 we focus on symmetric positive definite tridiagonal Toeplitz matrices that arise from the discretization of one-dimensional reaction-diffusion equations with Dirichlet boundary conditions, and study the MR and CG convergence quantities in the next-to-last step. Section 4 compares our results with known results for the Poisson equation model problem. Our conclusions are given in Section 5, and the Appendix lists all trigonometric formulas used in the proofs.

2. Formulas for the next-to-last MR and CG iteration step.

Let a symmetric positive definite matrix $A \in \mathbb{R}^{n\times n}$ be given and denote by $A = Q\Lambda Q^T$ its eigendecomposition, where $Q^TQ = I$ and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. To avoid unnecessary technical complications we assume that *all eigenvalues of A are distinct*. Next, we parameterize the initial residual r_0 and the initial error e_0 by

$$r_0 \;=\; Q\,[\varrho_1, \ldots, \varrho_n]^T\,, \qquad e_0 \;=\; Q\,[\xi_1, \ldots, \xi_n]^T\,. \tag{2.1}$$

Note that, since $r_0 = Ae_0$, we have $\varrho_k = \lambda_k \xi_k$ for all $k = 1, \ldots, n$. Without loss of generality we restrict our analysis to vectors r_0 with $\varrho_k \neq 0$ for all $k = 1, \ldots, n$. In case

$d \geq 1$ coordinates ϱ_j are zero, the corresponding eigencomponents do not play any role, and hence the formulas for $i = n - 1$ presented below will hold for $i = n - d - 1$.

2.1. General results. As shown in [13, Theorem 2.1], the MR residual norm in the $(n-1)$st (next-to-last) iteration step is given by

$$\|r_{n-1}^{MR}\| \;=\; \left(\sum_{j=1}^{n} \left| \frac{L_j}{\varrho_j} \right|^2 \right)^{-1/2} \;=\; \left(\sum_{j=1}^{n} \left| \frac{L_j}{\lambda_j \xi_j} \right|^2 \right)^{-1/2}, \tag{2.2}$$

where

$$L_k \;\equiv\; \prod_{\substack{j=1\\ j\neq k}}^{n} \frac{|\lambda_j|}{|\lambda_j - \lambda_k|}\,. \tag{2.3}$$

To obtain a similar result for the A-norm of the CG error, it suffices to realize that

$$\|p(A)e_0\|_A \;=\; \|p(A)A^{1/2}e_0\| \;\equiv\; \|p(A)\tilde{r}_0\|\,. \tag{2.4}$$

Hence the A-norm of the CG error can be seen as the MR residual norm, when MR is started with the initial residual $\tilde{r}_0 = A^{1/2}e_0$. Parameterizing $\tilde{r}_0$ by $\tilde{r}_0 = Q[\tilde{\varrho}_1, \ldots, \tilde{\varrho}_n]^T$, i.e. $\tilde{\varrho}_k = \lambda_k^{1/2}\xi_k = \lambda_k^{-1/2}\varrho_k$, we obtain

$$\|e_{n-1}^{CG}\|_A = \left(\sum_{j=1}^{n} \left| \frac{L_j}{\lambda_j^{1/2}\xi_j} \right|^2 \right)^{-1/2} \;=\; \left(\sum_{j=1}^{n} \left| \frac{\lambda_j^{1/2}L_j}{\varrho_j} \right|^2 \right)^{-1/2}. \tag{2.5}$$

The formulas (2.2) and (2.5) provide explicit a priori information about the next-to-last MR and CG convergence quantities in terms of the matrix eigenvalues and the coordinates of r_0 or e_0 in the matrix eigenvectors. To simplify the notation, we will write residuals and errors without superscript MR or CG. When we speak about residuals r_i, we always mean residuals r_i^{MR} of the MR method. Similarly, e_i always denotes the error e_i^{CG} of the CG method. The superscript can be now used to indicate the association of a residual or error with a particular initial residual or error.

2.2. Convergence quantities for different initial vectors. As described in the Introduction, we are interested in initial residuals and initial errors that lead to the maximal relative convergence quantities of the MR and CG method, respectively, in the next-to-last iteration step. We denote such a worst-case initial residual for the MR method by r_0^w, and the corresponding residual in the next-to-last step by r_{n-1}^w. In [13, Theorem 3.1] we show that

$$r_0^w = Q[\varrho_1^w, \ldots, \varrho_n^w]^T\,, \qquad |\varrho_k^w|^2 \;=\; \gamma\, L_k, \quad k = 1, \ldots, n\,, \tag{2.6}$$

where $\gamma > 0$ is any scaling factor, and that

$$\frac{\|r_{n-1}^w\|}{\|r_0^w\|} \;=\; \max_{r_0 \neq 0}\; \min_{p \in \pi_{n-1}} \frac{\|p(A)r_0\|}{\|r_0\|} \;=\; \left(\sum_{k=1}^{n} L_k \right)^{-1}. \tag{2.7}$$

Using the relation (2.4) and the definition of r_0^w it is not hard to see that the corresponding worst-case initial error e_0^w for CG is given by

$$e_0^w = Q[\xi_1^w, \ldots, \xi_n^w]^T\,, \qquad |\xi_k^w|^2 = \gamma\, \lambda_k^{-1} L_k \quad \text{for} \quad k = 1, \ldots, n\,, \tag{2.8}$$

where $\gamma > 0$ is any scaling factor, and that

$$\frac{\|e^w_{n-1}\|_A}{\|e^w_0\|_A} = \max_{e_0 \neq 0} \min_{p \in \pi_{n-1}} \frac{\|p(A)e_0\|_A}{\|e_0\|_A} = \left(\sum_{k=1}^{n} L_k \right)^{-1} . \tag{2.9}$$

We also consider the initial residual

$$r^u_0 = Q[\varrho^u_1, \ldots, \varrho^u_n]^T , \qquad \varrho^u_k = 1, \quad k = 1, \ldots, n. \tag{2.10}$$

The vector r^u_0 can be considered as a representative of the initial residuals which are uncorrelated with the matrix A, in the sense that their components in the eigenvectors of A are of (approximately) equal size. We call such vectors *unbiased* with respect to A. The MR method started with the initial residual (2.10) will produce, in the next-to-last iteration step, the residual vector r^u_{n-1}. Using (2.2), the relative MR residual norm is given by

$$\frac{\|r^u_{n-1}\|}{\|r^u_0\|} = \left(n \sum_{k=1}^{n} L^2_k \right)^{-1/2} . \tag{2.11}$$

The CG method started with the initial residual r^u_0, i.e. with the initial error

$$e^u_0 \;=\; A^{-1} r^u_0 \;=\; Q[\xi^u_1, \ldots, \xi^u_n]^T \;=\; Q[\lambda^{-1}_1, \ldots, \lambda^{-1}_n]^T, \tag{2.12}$$

generates in the next-to-last iteration step the error e^u_{n-1}. Based on (2.5), the relative A-norm of this error is given by

$$\frac{\|e^u_{n-1}\|_A}{\|e^u_0\|_A} = \left(\sum_{k=1}^{n} \lambda_k L^2_k \right)^{-1/2} \left(\sum_{k=1}^{n} \frac{1}{\lambda_k} \right)^{-1/2} . \tag{2.13}$$

The vector e^u_0 is by its definition correlated with the eigenvalue distribution of A and thus can be considered *biased*. We have deliberately made this choice to contrast the convergence quantities of MR and CG for the same initial residual.

3. Symmetric positive definite tridiagonal Toeplitz matrices. Consider the one-dimensional reaction-diffusion equation

$$-\,u''(z) + \sigma u(z) = f(z), \quad z \in (0,1), \tag{3.1}$$

for some parameter $\sigma \geq 0$, with Dirichlet boundary conditions

$$u(0) = u_0, \quad u(1) = u_1 \,. \tag{3.2}$$

Then for each positive integer n, the central finite difference approximation of (3.1)–(3.2) on the uniform grid kh, $k = 1, \ldots, n$, $h = (n+1)^{-1}$, leads to a linear system of the form

$$\underbrace{\begin{bmatrix} 2(1+\delta) & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2(1+\delta) \end{bmatrix}}_{A} x \;=\; \underbrace{h^2 \begin{bmatrix} f(h) \\ \vdots \\ \vdots \\ f(nh) \end{bmatrix} + \begin{bmatrix} u_0 \\ \\ \\ u_1 \end{bmatrix}}_{b} . \tag{3.3}$$

In the expression for A we have defined $\delta \equiv \sigma h^2/2$ for notational convenience.

The n distinct and positive eigenvalues λ_k, and the normalized eigenvectors q_k of A are given by

$$\lambda_k = 2(1+\delta) - 2\,\omega_k = 2\delta + 4\sin^2(k\pi h/2)\,, \quad \omega_k \equiv \cos(k\pi h)\,, \tag{3.4}$$

$$q_k = (2h)^{1/2}\left[\sin(k\pi h)\,, \sin(2k\pi h)\,, \ldots, \sin(nk\pi h)\right]^T\,, \quad k = 1, \ldots, n\,, \tag{3.5}$$

cf., e.g., [20, pp. 113–115]. We write the eigendecomposition of A as $A = Q\Lambda Q^T$, where $Q = [q_1, \ldots, q_n]$, and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$.

REMARK 3.1. We have chosen to derive our results for the tridiagonal Toeplitz matrix $A = \text{tridiag}(-1, 2(1+\delta), -1)$ in (3.3) because of its direct relation to the differential equation (3.1)–(3.2). However, our results hold equally well for any symmetric tridiagonal Toeplitz matrix of the form $B = \text{tridiag}(\beta, \alpha, \beta)$ with $\alpha = 2|\beta|(1+\delta) > 0$, for some $\delta > 0$. Obviously, $B = |\beta|\,\text{tridiag}(\beta/|\beta|, 2(1+\delta), \beta/|\beta|)$. If $\beta < 0$, then $B = |\beta|A$, and if $\beta > 0$, then $B = |\beta| I^{\pm} A I^{\pm}$, where $I^{\pm} = \text{diag}(1, -1, \ldots, (-1)^{n+1})$. In either case, A and B have the same set of orthogonal eigenvectors, and the eigenvalues B coincide with those of A up to a scaling by $|\beta|$. It is easy to check that all of our results are invariant under such scaling of the eigenvalues of A.

3.1. Connection with Chebyshev polynomials of the second kind.

The relation of the eigenvalues of A given in (3.4) to the roots of the nth Chebyshev polynomial of the *second* kind, denoted by $U_n(z)$, will prove useful in our context. The polynomial $U_n(z)$ has degree n, and its n distinct roots are the values $\omega_k = \cos(k\pi h)$, $k = 1, \ldots, n$. Hence all roots are contained in the open interval $(-1, 1)$. The leading coefficient of $U_n(z)$ is 2^n, which means that $U_n(z)$ can be written as

$$U_n(z) = 2^n \prod_{k=1}^{n} (z - \omega_k)\;.$$

This relation shows that the product of all eigenvalues of A can be expressed as

$$\prod_{k=1}^{n} \lambda_k = 2^n \prod_{k=1}^{n} (1 + \delta - \omega_k) = U_n(1+\delta)\,. \tag{3.6}$$

Below we study how much the MR and CG convergence quantities change with changing δ. For this we first need to understand the behavior of $U_n(1+\delta)$ as a function of $\delta \geq 0$. To get a feeling of the growth of $U_n(z)$ outside the interval $(-1, 1)$, we use the alternative representation

$$U_n(z) = \frac{1}{2}\,\frac{(z + \sqrt{z^2-1})^{n+1} - (z - \sqrt{z^2-1})^{n+1}}{\sqrt{z^2-1}}\,, \tag{3.7}$$

see, e.g., [14, p. 15]. Using this formula, elementary real analysis shows that

$$U_n(1) = |U_n(-1)| = n+1\,,$$

and that $U_n'(z) > 0$ for $z \geq 1$. In particular, $U_n(1+\delta)$ is positive and strictly increasing for $\delta \geq 0$. As shown by (3.7), $|U_n(z)|$ grows exponentially outside $(-1, 1)$. This is illustrated in Fig. 3.1, where we plot $U_n(z)/(n+1)$ for $n = 4, 6, 10$.

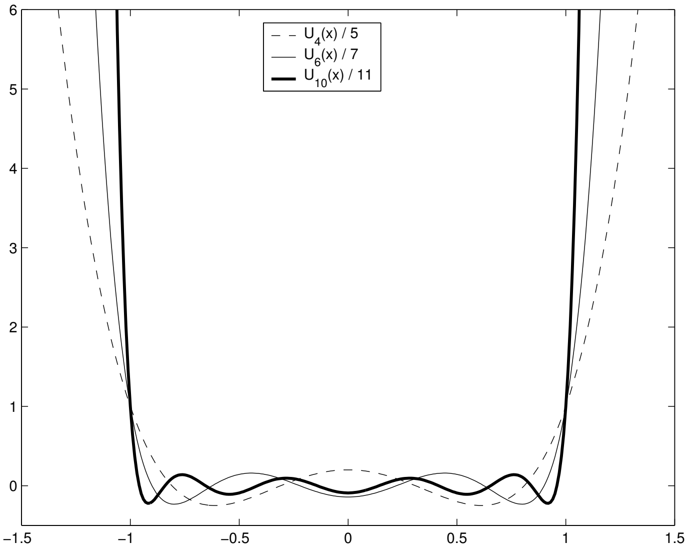


FIG. 3.1. $U_n(z)/(n+1)$ *for different* n.

3.2. Worst-case data. Our goal here is to characterize data (source term f and boundary conditions) in (3.1)–(3.2), that lead to the maximal relative convergence quantities in the next-to-last step when MR and CG with the initial guess $x_0 = 0$ are applied to the discretized system (3.3). Our main tools are the parameterizations (2.6) and (2.8) of the worst-case initial vectors r_0^w and e_0^w, which we evaluate explicitly using the known eigendecomposition of A, and then translate back into data for (3.1)–(3.2). The vectors r_0^w and e_0^w depend on the terms L_k, which are characterized by the following lemma.

LEMMA 3.2. *Suppose that* $\lambda_1, \ldots, \lambda_n$ *are given by* (3.4) *for some* $\delta \geq 0$. *Then* L_k *as defined in* (2.3) *satisfies*

$$L_k = h\, U_n\,(1+\delta)\, \frac{\sin^2(k\pi h)}{\delta + 2\sin^2\left(\frac{k\pi h}{2}\right)}\,. \tag{3.8}$$

In particular, for $\delta = 0$,

$$L_k \;=\; 2\cos^2\left(\frac{k\pi h}{2}\right). \tag{3.9}$$

Proof. The denominator of L_k can be written as

$$\prod_{\substack{j=1\\ j\neq k}}^{n} |\lambda_j - \lambda_k| \;=\; \prod_{\substack{j=1\\ j\neq k}}^{n} |2\,\omega_k - 2\,\omega_j| \;=\; 2^{2n-2} \prod_{\substack{j=1\\ j\neq k}}^{n} \left|\sin^2\left(\frac{jh\pi}{2}\right) - \sin^2\left(\frac{kh\pi}{2}\right)\right|$$

$$\;=\; \frac{n+1}{2\sin^2(k\pi h)}\,, \tag{3.10}$$

cf. identity (A.1). According to (2.3), (3.6) and (3.10),

$$L_k \;=\; \frac{U_n\,(1+\delta)}{\lambda_k} \cdot \frac{2\sin^2(k\pi h)}{n+1} \;=\; h\, U_n\,(1+\delta)\, \frac{\sin^2(k\pi h)}{\delta + 2\sin^2\left(\frac{k\pi h}{2}\right)}\,. \tag{3.11}$$

The relation (3.9) for $\delta = 0$ follows immediately from $U_n(1) = n+1 = h^{-1}$ and $\sin(k\pi h) = 2\sin(k\pi h/2)\cos(k\pi h/2)$. $\square$

Now consider the parameterization of e_0^w given in (2.8). Clearly, for any $\gamma > 0$, the set of coefficients

$$\xi_k^w \;\equiv\; \left(\gamma\lambda_k^{-1}L_k\right)^{1/2}, \quad k = 1, \dots, n, \tag{3.12}$$

leads to a worst-case initial error $e_0^w = Q[\xi_1^w, \dots, \xi_n^w]^T$ for CG. If CG is started with initial guess $x_0 = 0$, then e_0^w represents the solution, and Ae_0^w the right hand side of a linear system that leads to the maximal relative A-norm of the error in the next-to-last iteration step.

Using the coefficients (3.12), and the explicit form of L_k in (3.11),

$$\begin{aligned} \lambda_k\xi_k^w &= \lambda_k\left(\gamma\lambda_k^{-1}\frac{U_n(1+\delta)}{\lambda_k}\cdot\frac{2\sin^2(k\pi h)}{n+1}\right)^{1/2} \\ &= \left(\gamma\, 2h\, U_n(1+\delta)\right)^{1/2}\sin(k\pi h), \end{aligned}$$

and, therefore,

$$\begin{aligned} Ae_0^w &= (Q\Lambda Q^T)\,(Q[\xi_1^w, \dots, \xi_n^w]^T) \\ &= Q\,[\lambda_1\xi_1^w, \dots, \lambda_n\xi_n^w]^T \\ &= \left(\gamma\, 2h\, U_n(1+\delta)\right)^{1/2} Q\,[\sin(\pi h), \dots, \sin(n\pi h)]^T \\ &= \left(\gamma\, 2h\, U_n(1+\delta)\right)^{1/2} Q\,(2h)^{-1/2}\, q_1 \\ &= \left(\gamma U_n(1+\delta)\right)^{1/2}[1, 0, \dots, 0]^T. \end{aligned} \tag{3.13}$$

Since $\gamma > 0$ can be chosen arbitrarily, we conclude that any right hand side vector b that is a positive multiple of the first unit vector leads to the worst possible relative A-norm of the error in the next-to-last step of CG (with $x_0 = 0$) for the linear system $Ax = b$ given by (3.3). The convergence of CG (with $x_0 = 0$) for $Ax = b$ is obviously the same as for $Ax = -b$, and therefore any negative multiple of the first unit vector is a worst-case right hand side in the just described sense as well.

Instead of the coefficients (3.12) we may define

$$\xi_k^w \;\equiv\; (-1)^{k+1}\left(\gamma\lambda_k^{-1}L_k\right)^{1/2}, \quad k = 1, \dots, n. \tag{3.14}$$

Then, using $(-1)^{k+1}\sin(k\pi h) = \sin(nk\pi h)$, we obtain

$$\lambda_k\xi_k^w \;=\; \left(\gamma\, 2h\, U_n(1+\delta)\right)^{1/2}\sin(nk\pi h).$$

A computation analogous to the one leading to (3.13) shows that, for the initial error e_0^w defined by the coefficients (3.14),

$$Ae_0^w \;=\; \left(\gamma U_n(1+\delta)\right)^{1/2}[0, \dots, 0, 1]^T, \tag{3.15}$$

i.e., any nonzero multiple of the nth unit vector also is a worst-case right hand side for CG.

Both examples show that the right hand sides leading to the very unfavorable convergence behavior of CG may look rather unsuspicious at first sight. In terms of the differential equation (3.1)–(3.2), the worst possible relative A-norm of the next-to-last error in CG (for $x_0 = 0$) is obtained simply by

$$f = 0 \quad \text{and} \quad u_0 = c,\; u_1 = 0, \quad \text{or} \quad u_0 = 0,\; u_1 = c, \tag{3.16}$$

for any nonzero constant c.

As shown in (2.4), CG for the initial error e_0^w defined by (3.12) is equivalent to MR for the initial residual $A^{1/2}e_0^w$ that can be written in the form

$$\begin{aligned} A^{1/2}e_0^w &= A^{-1/2}Ae_0^w \\ &= (\gamma U_n\,(1+\delta))^{1/2}\;A^{-1/2}[1,0,\ldots,0]^T \\ &= (\gamma U_n\,(1+\delta))^{1/2}\;Q\Lambda^{-1/2}q_1\,. \end{aligned}$$

Therefore, any nonzero multiple of the vector $r_0^w \equiv Q\Lambda^{-1/2}q_1$ leads to the worst-case relative residual norm in the next-to-last MR step. Obviously, the coordinates of r_0^w in the eigenvectors of A are given by

$$\varrho_k^w = [2\delta + 4\sin^2(k\pi h/2)]^{-1/2}\sin(k\pi h), \quad k = 1,\ldots,n\,. \tag{3.17}$$

Because of the complicated form of the ϱ_k^w, no simple expression for the vector $r_0^w = Q[\varrho_1^w,\ldots,\varrho_n^w]^T$ exists in general. An exception for which r_0^w can be found in a relatively simple form is the case $\delta = 0$, where $\varrho_k^w = \cos(k\pi h/2)$, and the jth entry of r_0^w, denoted by $r_{0,j}^w$ for $j = 1,\ldots,n$, satisfies

$$r_{0,j}^w \;=\; (2h)^{1/2}\frac{\sin\,(j\pi h)}{\cos\left(\frac{\pi h}{2}\right) - \cos\,(j\pi h)}\,. \tag{3.18}$$

As (3.18) indicates, for MR it is not as straightforward as for CG to find data for (3.1)–(3.2) that leads to the worst case in the next-to-last step. For more details and a proof of (3.18) we refer to [11].

3.3. Worst-case and unbiased convergence quantities. After having characterized the worst-case initial vectors r_0^w and e_0^w for the system (3.3), we next evaluate the corresponding convergence quantity (2.7) and compare it to the quantities (2.11) and (2.13) resulting from the initial vectors r_0^u and e_0^u. We start with deriving bounds on (2.7) and (2.11).

THEOREM 3.3. *Suppose that MR is applied to a system of the form (3.3), and the initial residual is either r_0^w or r_0^u. Then*

$$3^{-1}\,\frac{2+\delta}{U_n\,(1+\delta)} \;<\; \frac{\|r_{n-1}^u\|}{\|r_0^u\|} \;<\; \frac{\|r_{n-1}^w\|}{\|r_0^w\|} \;\le\; 3\,\frac{2+\delta}{U_n\,(1+\delta)}\,. \tag{3.19}$$

In particular, for $\delta = 0$,

$$\frac{1}{n}\sqrt{\frac{2}{3}} \;<\; \sqrt{\frac{2}{3n^2-n}} \;=\; \frac{\|r_{n-1}^u\|}{\|r_0^u\|} \;<\; \frac{\|r_{n-1}^w\|}{\|r_0^w\|} \;=\; \frac{1}{n}\,. \tag{3.20}$$

Proof. We first prove (3.19). The middle inequality is trivial. To show the leftmost inequality it suffices to use the relation (2.11) and to find an upper bound on the sum of the L_k^2. Using (3.8) and (A.4),

$$\begin{aligned} \sum_{k=1}^{n} L_k^2 &\le \frac{U_n^2\,(1+\delta)}{(n+1)^2(\frac{\delta}{2}+1)^2}\sum_{k=1}^{n}\frac{\sin^4(k\pi h)}{4\sin^4\left(\frac{k\pi h}{2}\right)} \\ &= \frac{16\,U_n^2\,(1+\delta)}{(n+1)^2(\delta+2)^2}\sum_{k=1}^{n}\cos^4\left(\frac{k\pi h}{2}\right) \\ &= \frac{(6\,n-2)\,U_n^2\,(1+\delta)}{(n+1)^2(\delta+2)^2}\,. \end{aligned} \tag{3.21}$$

Then (2.11) implies

$$\left(n \sum_{k=1}^{n} L_k^2 \right)^{-1/2} \geq \frac{(n+1)(\delta+2)}{\sqrt{(6n-2)n}\, U_n\,(1+\delta)} > \frac{1}{3}\, \frac{\delta+2}{U_n\,(1+\delta)}\,.$$

Next note that, using (A.3),

$$\begin{aligned} \sum_{k=1}^{n} L_k &\geq \frac{U_n\,(1+\delta)}{\delta+2} \sum_{k=1}^{n} \frac{\sin^2(k\pi h)}{n+1} = \frac{1}{2}\,\frac{U_n\,(\delta+1)}{\delta+2}\,\frac{n}{n+1} \\ &\geq \frac{1}{3}\,\frac{U_n\,(\delta+1)}{\delta+2}\,, \end{aligned} \tag{3.22}$$

and thus the rightmost inequality in (3.19) follows from applying (3.22) to (2.7).

For $\delta = 0$ we have

$$\sum_{k=1}^{n} L_k = 2 \sum_{k=1}^{n} \cos^2\left(\frac{k\pi h}{2}\right) = n\,, \tag{3.23}$$

cf. (A.3), and

$$\begin{aligned} \sum_{k=1}^{n} L_k^2 &= \frac{U_n^2\,(1)}{(n+1)^2} \sum_{k=1}^{n} \frac{\sin^4(k\pi h)}{4\sin^4\left(\frac{k\pi h}{2}\right)} \\ &= 4 \sum_{k=1}^{n} \cos^4\left(\frac{k\pi h}{2}\right) = \frac{3n-1}{2}\,, \end{aligned} \tag{3.24}$$

cf. (A.4). Substituting (3.23) and (3.24) into (2.7) and (2.11), we obtain (3.20). □

Since $\|r_{n-1}^w\|/\|r_0^w\| = \|e_{n-1}^w\|_A/\|e_0^w\|_A$ (compare (2.7) and (2.9)) the theorem also characterizes $\|e_{n-1}^w\|_A/\|e_0^w\|_A$, the next-to-last worst-case relative A-norm of the error for CG.

The rightmost equation in (3.20) shows that, for $\delta = 0$, MR in the worst case decreases the relative residual norm in the first $n-1$ iteration steps only to n^{-1}. On the other hand, since $\|r_{n-1}^w\|/\|r_0^w\| \approx (1+\delta)/U_n(1+\delta)$ for all δ, the next-to-last worst-case MR residual norm decreases exponentially with increasing δ, and hence increasing diagonal dominance of A. Moreover, Theorem 3.3 shows that the progress MR has made in the next-to-last iteration step for the unbiased initial residual r_0^u is at most a *constant factor* (less than $1/9$) apart from the worst case. In general the two cases may differ by a factor of up to $n^{1/2}$; see [13, Section 5], [7, Section 5].

The spectral structure of A allows to use the worst-case convergence result for the next-to-last step in Theorem 3.3 to obtain a worst-case convergence bound also for other iteration steps.

COROLLARY 3.4. *Suppose that the positive integer m divides $n+1$. Then for all $i \equiv (n+1)/m - 2 > 1$,*

$$\max_{r_0 \neq 0}\, \min_{p \in \pi_i} \frac{\|p(A) r_0\|}{\|r_0\|} = \max_{e_0 \neq 0}\, \min_{p \in \pi_i} \frac{\|p(A) e_0\|_A}{\|e_0\|_A} > 3^{-1}\, \frac{2+\delta}{U_{i+1}\,(1+\delta)}\,. \tag{3.25}$$

Proof. Consider the subset $\{\mu_1, \ldots, \mu_{i+1}\} \subseteq \{\lambda_1, \ldots, \lambda_n\}$ of $i+1$ eigenvalues of A given by

$$\mu_j = 2\left(1 + \delta - \cos\left(\frac{j\,\pi}{i+2}\right)\right), \qquad j = 1, \ldots, i+1\,.$$

It is easy to see that the set $\{\mu_1, \ldots, \mu_{i+1}\}$ consists of the $i+1$ distinct eigenvalues of $A_{i+1} \equiv \text{tridiag}(-1, 2(1+\delta), -1) \in \mathbb{R}^{(i+1)\times(i+1)}$. Then

$$\begin{aligned} \max_{r_0 \neq 0} \min_{p \in \pi_i} \frac{\|p(A) r_0\|}{\|r_0\|} &= \min_{p \in \pi_i} \max_{k} |p(\lambda_k)| \\ &\geq \min_{p \in \pi_i} \max_{k} |p(\mu_k)| \\ &= \max_{r_0 \neq 0} \min_{p \in \pi_i} \frac{\|p(A_{i+1}) r_0\|}{\|r_0\|} \\ &> 3^{-1} \frac{2+\delta}{U_{i+1}\,(1+\delta)}\,, \end{aligned}$$

where the final lower bound results from applying Theorem 3.3 to the linear system with A_{i+1}. □

For example, in case $n = 99$, the lower bound (3.25) would apply in the steps $i = 2, 8, 18, 23, 48, 98$. Hence in addition to just the lower bound on $\|r^w_{n-1}\|/\|r^w_0\|$ in (3.19), which corresponds to (3.25) for $i = n-1$, we get additional lower bounds particularly for the earlier phase of the iteration.

Theorem 3.3 does not characterize (2.13), i.e. the case of CG for the initial error e^u_0. This is done in the following result.

THEOREM 3.5. *Suppose that CG is applied to a system of the form (3.3), and the initial error is e^u_0. Then*

$$3^{-1} \frac{\delta}{U_n\,(1+\delta)} < \frac{\|e^u_{n-1}\|_A}{\|e^u_0\|_A} < 3\,\frac{2+\delta}{U_n\,(1+\delta)}\,. \tag{3.26}$$

For $\delta < 1/4$,

$$3^{-1} \frac{\delta+2}{n^{1/2} U_n\,(1+\delta)} < \frac{\|e^u_{n-1}\|_A}{\|e^u_0\|_A}\,, \tag{3.27}$$

and for $\delta = 0$,

$$\frac{\|e^u_{n-1}\|_A}{\|e^u_0\|_A} = \frac{\sqrt{6}}{\sqrt{n(n+1)(n+2)}} > n^{-3/2}\,. \tag{3.28}$$

Proof. The second inequality in (3.26) follows easily from (3.19). We prove the first inequality. Using Cauchy's inequality we obtain, cf. (2.13),

$$\frac{\|e^u_0\|_A^2}{\|e^u_{n-1}\|_A^2} \leq \left(\sum_{k=1}^{n} L_k^4\right)^{1/2} \left(\sum_{k=1}^{n} \lambda_k^2\right)^{1/2} \left(\sum_{k=1}^{n} \frac{1}{\lambda_k}\right)\,. \tag{3.29}$$

Since λ_n is the largest eigenvalue,

$$\begin{aligned}\left(\sum_{k=1}^{n}\lambda_k^2\right)^{1/2}\left(\sum_{k=1}^{n}\frac{1}{\lambda_k}\right) &< n^{1/2}\lambda_n\left(\sum_{k=1}^{n}\frac{1}{\lambda_k}\right) \;<\; n^{3/2}\frac{\lambda_n}{\lambda_1}\\ &= n^{3/2}\frac{1+\delta+\omega_1}{1+\delta-\omega_1}\\ &< n^{3/2}\frac{2+\delta}{\delta}\,.\end{aligned}\tag{3.30}$$

It remains to find a bound on the sum of the L_k^4. Using (3.8) and (A.5),

$$\begin{aligned}\sum_{k=1}^{n}L_k^4 \;&\le\; \frac{U_n^4\,(1+\delta)}{(n+1)^4(\frac{\delta}{2}+1)^4}\sum_{k=1}^{n}\frac{\sin^8(k\pi h)}{2^4\sin^8\left(\frac{k\pi h}{2}\right)}\\ &=\; 2^8\frac{U_n^4\,(1+\delta)}{(n+1)^4(\delta+2)^4}\sum_{k=1}^{n}\cos^8\left(\frac{k\pi h}{2}\right)\\ &<\; 3^4\frac{n\,U_n^4\,(1+\delta)}{(n+1)^4(\delta+2)^4}\,.\end{aligned}\tag{3.31}$$

From (3.29)–(3.31) we now obtain (3.26).

Now consider the case $\delta < 1/4$. Then

$$\begin{aligned}\left(\sum_{k=1}^{n}\lambda_k^2\right)^{1/2}\sum_{k=1}^{n}\frac{1}{\lambda_k} \;&=\; \left(\sum_{k=1}^{n}(1+\delta-\omega_k)^2\right)^{1/2}\sum_{k=1}^{n}\frac{1}{1+\delta-\omega_k}\\ &<\; \left(\sum_{k=1}^{n}(5/4-\omega_k)^2\right)^{1/2}\sum_{k=1}^{n}\frac{1}{1-\omega_k}\\ &=\; \left(\frac{33}{16}n-\frac{1}{2}\right)^{1/2}\sum_{k=1}^{n}\frac{1}{2\sin^2\left(\frac{k\pi h}{2}\right)}\\ &<\; \left(\frac{36}{16}n\right)^{1/2}\frac{n(n+2)}{3}\\ &=\; \frac{n^{1/2}n(n+2)}{2}\\ &<\; \frac{n^{1/2}(n+1)^2}{2}\,,\end{aligned}\tag{3.32}$$

where we have used the identities (A.7) and (A.8). Then (3.27) follows from (3.29), (3.31) and (3.32).

For $\delta = 0$,

$$\begin{aligned}\frac{\|e_0^u\|_A^2}{\|e_{n-1}^u\|_A^2} \;&=\; \left(\sum_{k=1}^{n}4\sin^2\left(\frac{k\pi h}{2}\right)4\cos^4\left(\frac{k\pi h}{2}\right)\right)\left(\sum_{k=1}^{n}\frac{1}{4\sin^2\left(\frac{k\pi h}{2}\right)}\right)\\ &=\; (n+1)\left(\frac{n(n+2)}{6}\right)\,,\end{aligned}$$

where we have used (A.6) and (A.7). □

A comparison of Theorems 3.3 and 3.5 shows that, for small δ,

$$\text{(MR)} \qquad \frac{\|r_{n-1}^u\|}{\|r_0^u\|} \;\approx\; n^{1/2}\,\frac{\|e_{n-1}^u\|_A}{\|e_0^u\|_A} \qquad \text{(CG)}\,.$$

For larger δ, this difference is much less pronounced, and these MR and CG quantities are at most a small constant apart from each other.

3.4. Comparison of the worst-case bound and the classical bound. We next compare our worst-case convergence results in Theorem 3.3 with the classical convergence bound (1.5),

$$\min_{p\in\pi_i}\max_k |p(\lambda_k)| \;\leq\; 2\nu^i\,, \qquad i = 0,\ldots,n-1\,, \tag{3.33}$$

where $\nu \equiv (\sqrt{\kappa(A)}-1)/(\sqrt{\kappa(A)}+1) < 1$, for $i = n-1$.

For our comparison we express $U_n(1+\delta)$ in terms of the condition number of A, which is given by $\kappa(A) = \lambda_n/\lambda_1$. First note that, by (3.4),

$$1+\delta \;=\; \omega_1\,\frac{\lambda_n+\lambda_1}{\lambda_n-\lambda_1} \;=\; \omega_1\,\frac{\kappa(A)+1}{\kappa(A)-1} \;\equiv\; \omega_1\,\tau\,.$$

Next,

$$\tau - \sqrt{\tau^2-1} \;=\; \frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1} \;\equiv\; \nu\,, \qquad \tau + \sqrt{\tau^2-1} \;=\; \nu^{-1}\,, \tag{3.34}$$

which, inserted into (3.7), yields

$$U_n(\tau) \;=\; \frac{\nu^{n+1}-\nu^{-(n+1)}}{\nu-\nu^{-1}}\,. \tag{3.35}$$

Since $U_n(z)$ is strictly monotonically increasing for $z \geq 1$, and $\omega_1 \lesssim 1$,

$$U_n(1+\delta) \;\lesssim\; U_n(\tau) \;=\; \nu^{-n}+\nu^{-n+2}+\nu^{-n+4}+\ldots+\nu^n\,, \tag{3.36}$$

where "$\lesssim$" means that the inequality is close. In the notation established above,

$$\begin{aligned}
2\nu^{n-1} \;&\geq\; \frac{\|r_{n-1}^w\|}{\|r_0^w\|} \;=\; \frac{\|e_{n-1}^w\|_A}{\|e_0^w\|_A} && (3.37)\\
&\gtrsim\; \frac{\|r_{n-1}^u\|}{\|r_0^u\|} \;\approx\; \frac{4}{\omega_1}\,\frac{2+\delta}{U_n\,(1+\delta)} && (3.38)\\
&\gtrsim\; \frac{4\tau}{U_n(\tau)} && (3.39)\\
&\gtrsim\; \frac{2}{\nu\,U_n(\tau)} \;=\; \frac{2\,\nu^{n-1}}{1+\nu^2+\ldots+\nu^{2(n-1)}+\nu^{2n}}\,. && (3.40)
\end{aligned}$$

In (3.37) we use (3.33) for $i = n-1$, and in (3.38) we use (3.19), where the unimportant multiplicative factor (between $1/3$ and 3) was replaced by $4/\omega_1$ for convenience. Next, in (3.39) we use (3.36) as well as the relation $\tau = (1+\delta)/\omega_1$, from which we receive (3.40) using (3.36) and the inequality $2\,\tau \geq \nu^{-1} \geq \tau$.

The main point in this derivation is that the actual convergence quantities on the right hand side of the inequality in (3.37) are always quite close to (3.40), i.e.

$$\frac{\|r^w_{n-1}\|}{\|r^w_0\|} = \frac{\|e^w_{n-1}\|_A}{\|e^w_0\|_A} \approx \frac{2\,\nu^{n-1}}{1+\nu^2+\ldots+\nu^{2(n-1)}+\nu^{2n}}\,.$$

The tightness of the *upper* bound (3.37) to the actual convergence quantities therefore depends on the size of ν, and hence on $\kappa(A)$, which for a fixed matrix size n is a strictly decreasing function of the parameter $\delta \geq 0$.

For small $\kappa(A)$ (or δ bounded away from zero), the difference between (3.37) and (3.40) is small, i.e. the classical bound provides accurate information about the actual convergence quantities of CG and MR in (3.37) and (3.38). On the other hand, when $\kappa(A)$ is large (or δ is close to zero), then the lower bound (3.40), and with it the CG and MR convergence quantities will be smaller (up to the factor n^{-1}) than predicted by the classical upper bound (3.37). In the limiting case $\delta = 0$,

$$\min_{p\in\pi_{n-1}} \max_{1\leq k\leq n} |p(\lambda_k)| = \frac{1}{n} \ll 2\nu^{n-1} \overset{n\to\infty}{\longrightarrow} 2e^{-\pi}\,.$$

This clearly demonstrates that, for reasonably large n, the classical bound (3.33) cannot describe the worst-case convergence values of CG or MR in later iterations. Asymptotically (for $n \to \infty$) the weakness of the classical bound in this context has also been noticed before by Axelsson [1, Example 13.7] and others.

4. Poisson equation. Now we consider the case of one-dimensional Poisson equation with Dirichlet boundary conditions, i.e. the problem (3.1)–(3.2) with $\sigma = 0$. Then $\delta = 0$ and the corresponding system matrix in (3.3) is $A = \text{tridiag}(-1, 2, -1)$. In this case, simple explicit expressions for r^w_0 as well as e^w_0 are known (see Section 3.2). Moreover, we have determined the exact MR and CG convergence quantities in the next-to-last step for the worst-case as well as the unbiased initial vectors (see Theorems 3.3 and 3.5). In addition, it is possible, in this particular case and for special starting vectors including the ones considered in this paper, to determine the whole MR and CG convergence curve a priori. In the following we recall known results from [15] for the unbiased case, and state (without proof) a new convergence result for the worst case.

Assuming that $x_0 = 0$, and hence $e_0 = x$, the papers [15, 16] present exact analytic expressions for the relative A-norm of the CG errors for solutions of the form

$$x^{(s)} = Q[\xi^{(s)}_1, \ldots, \xi^{(s)}_n]^T\,, \qquad \xi^{(s)}_k = \sin^{-s}\left(\frac{k\pi h}{2}\right)\,, \tag{4.1}$$

for some parameter $s \in \mathbb{N}_0$. Two of these solutions are of particular interest in our context. A simple calculation shows that $x^{(2)} = 4e^u_0$ as defined in (2.12). Moreover, $A^{1/2}x^{(1)} = 2r^u_0$, where r^u_0 is defined in (2.10). Using these relations and the exact analytic convergence curves derived in [15] gives the following result.

PROPOSITION 4.1. *Suppose that CG and MR are applied to the system (3.3) with $\delta = 0$, and the respective initial error and residual are given by e^u_0 and r^u_0. Then the resulting CG errors e^u_i and MR residuals r^u_i, $i = 0, \ldots, n$, satisfy*

$$\frac{\|e^u_i\|_A}{\|e^u_0\|_A} = \left[\frac{(n-i)^3+3(n-i)^2+2(n-i)}{n(n+1)(n+2)}\right]^{1/2} \equiv \varphi_C(i)\,, \tag{4.2}$$

$$\frac{\|r^u_i\|}{\|r^u_0\|} = \left[\frac{(n-i)+(n-i)^2}{n(n+1)+2ni(n-i)}\right]^{1/2} \equiv \varphi_M(i)\,. \tag{4.3}$$

An elementary computation using (4.2) shows that

$$\frac{\varphi_C(i)}{\varphi_C(i-1)} = \left(\frac{n-i}{n-i+3}\right)^{1/2}, \qquad i=1,\ldots,n,$$

which represents a strictly decreasing function of the iteration step i. The “superlinear” behavior of $\varphi_C(i)$ can be related to the distribution of the eigenvector coordinates of the initial error e_0^u. As proved asymptotically by Beckermann and Kuijlaars [2], CG may for the model problem (3.3) with $\delta = 0$ converge superlinearly, when the initial error exhibits a certain distribution of eigencomponents that is far from an equilibrium distribution. This appears to be the case in our example, where e_0^u is *biased*, cf. (2.12).

Using the same techniques as in [15] based on Lagrange multipliers, it is also possible to determine the exact values of the relative A-norm of the error in every step of CG with the initial error e_0^w. This technique is quite involved, and the full proof would take us several pages to state. The final result is the following,

$$\text{(4.4)} \qquad \frac{\|e_i^w\|_A}{\|e_0^w\|_A} = \left[\frac{n-i}{n\,(i+1)}\right]^{1/2} \equiv \varphi_W(i)\,, \quad i=0,\ldots,n\,.$$

Because of the equivalence (2.4) between CG and MR, the relative MR residual norms for the initial residual r_0^w also satisfy $\|r_i^w\|/\|r_0^w\| = \varphi_W(i)$. Note that

$$\text{(4.5)} \qquad \varphi_M(i) \;<\; \varphi_W(i) \;<\; \sqrt{2}\,\varphi_M(i)\,, \quad i=1,\ldots,n-1\,.$$

Obviously, the worst-case convergence value (1.4) of CG and MR at each step i must be larger than (or equal) to any other attainable convergence value. Hence the maximum of the three convergence curves $\varphi_C(i)$, $\varphi_M(i)$ and $\varphi_W(i)$ forms a lower bound on the worst-case value,

$$\text{(4.6)} \qquad \min_{p\in\pi_i}\max_k |p(\lambda_k)| \geq \max\{\varphi_C(i),\varphi_M(i),\varphi_W(i)\}\,, \quad i=0,\ldots,n-1\,.$$

Figure 4.1 illustrates the above results for the model problem (3.3) with $n = 120$ and $\delta = 0$. The computations were performed in MATLAB [21], on an AMD Athlon XP 2100+ personal computer with machine precision $\varepsilon \sim 10^{-16}$.

As predicted by (4.5), the curves $\varphi_M(i)$ (dashed dotted) and $\varphi_W(i)$ (solid) are very close. The left hand side of (4.6) (bold) was computed by the function `cheby0` of the semidefinite programming package SDPT3 [22]. Except for the last few steps, the maximum on the right hand side of (4.6) is given by $\varphi_C(i)$ (dashed). Overall, the bound (4.6) is quite tight. The bound (3.33) is tight in step i, if there exist $i-1$ eigenvalues of A, that closely approximate extrema of the ith scaled and shifted Chebyshev polynomial of the first kind. In our example this is not the case for the later phase of the iteration, where the two sides of (3.33) differ significantly.

As mentioned above, MR with the right hand side r_0^w (we used $x_0 = 0$ for MR and CG) and CG with the right hand side Ae_0^w have the same convergence curve given by $\varphi_W(i)$ (solid). However, the curves of MR with the right hand side Ae_0^w (dotted) and CG with the right hand side r_0^w (dashed dotted; coincides with $\varphi_M(i)$) differ by orders of magnitude from each other. Hence a right hand side that leads to the worst-case convergence for one method does not lead (in general) to similar convergence for the other method.

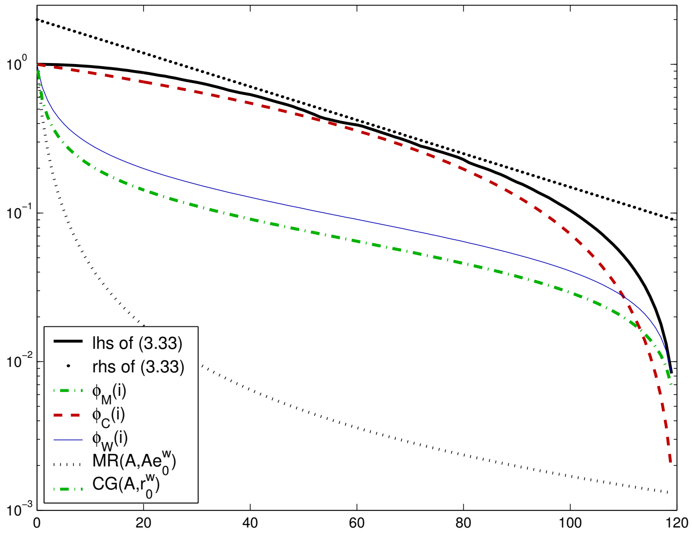


FIG. 4.1. *CG and MR convergence curves, and both sides of* (3.33).

5. Conclusions. In this paper we have applied our previous results in [13] to study the convergence of the CG and MR methods for linear systems with symmetric positive definite tridiagonal Toeplitz matrices. The structure of the matrix spectra allowed us to answer the questions how slow the convergence of the iterative solvers might possibly be for the considered model problems, which initial vectors lead to the maximal convergence quantity in the next-to-last iteration step, and how much the convergence quantity in this case differs from an “average” (or unbiased) case. We also were able to derive lower bounds on the worst-case convergence quantities in other iteration steps using the lower bound for the next-to-last step. The presented approach can be applied also to other classes of model problems in which the matrix eigenvalues are known, and the Lagrange factors L_k in (2.3) can be evaluated.

Acknowledgments. We thank Miro Rozložník, Daniel Szyld and anonymous referees for helpful comments and suggestions that improved the paper.

Appendix. Let $h = (n+1)^{-1}$, $n \in \mathbb{N}$. Then the following identities hold:

$$\frac{n+1}{2^{2n-1}}\,\frac{1}{\sin^2(k\pi h)} = \prod_{\substack{j=1\\ j\neq k}}^{n} \left| \sin^2\left(\frac{j\pi h}{2}\right) - \sin^2\left(\frac{k\pi h}{2}\right) \right| , \tag{A.1}$$

$$\frac{n+1}{2^n} = \prod_{j=1}^{n} \sin(j\pi h)\ , \tag{A.2}$$

$$\frac{n}{2} = \sum_{j=1}^{n} \cos^2(j\pi h) = \sum_{j=1}^{n} \sin^2(j\pi h)\,, \tag{A.3}$$

$$\frac{3\,n-1}{2^3} = \sum_{j=1}^{n} \cos^4\left(\frac{j\pi h}{2}\right), \tag{A.4}$$

$$\frac{35\,n-29}{2^7} = \sum_{j=1}^{n} \cos^8\left(\frac{j\pi h}{2}\right), \tag{A.5}$$

$$\frac{n+1}{16} = \sum_{j=1}^{n} \sin^2\left(\frac{j\pi h}{2}\right)\cos^4\left(\frac{j\pi h}{2}\right), \tag{A.6}$$

$$\frac{2\,n(n+2)}{3} = \sum_{j=1}^{n} \sin^{-2}\left(\frac{j\pi h}{2}\right), \tag{A.7}$$

$$\frac{33}{16}n - \frac{1}{2} = \sum_{j=1}^{n} \left(\frac{5}{4} - \cos(j\pi h)\right)^2. \tag{A.8}$$

Identity (A.2) can be found in [3, p. 40], and the sums (A.3)–(A.8) can be verified using MAPLE [23]. To prove the non-standard identity (A.1), we note that

$$\begin{aligned}&\prod_{\substack{j=1\\ j\neq k}}^{n}\left[\sin^2\left(\frac{j\pi h}{2}\right) - \sin^2\left(\frac{k\pi h}{2}\right)\right]\\ &= \prod_{\substack{j=1\\ j\neq k}}^{n} \sin\left(\frac{(j+k)\pi h}{2}\right) \prod_{\substack{j=1\\ j\neq n+1-k}}^{n} \cos\left(\frac{(j+k)\pi h}{2}\right).\end{aligned}$$

If $kh = \frac{1}{2}$ then, $n+1-k = k$, and the product in (A.1) takes the form

$$\begin{aligned}\prod_{\substack{j=1\\ j\neq k}}^{n}\left|\sin\left(\frac{(j+k)\pi h}{2}\right)\cos\left(\frac{(j+k)\pi h}{2}\right)\right| &= \frac{1}{2^{n-1}}\prod_{\substack{j=1\\ j\neq k}}^{n} |\sin((j+k)\pi h)|\\ &= \frac{1}{2^{n-1}}\prod_{j=1}^{n}\sin(j\pi h) = \frac{n+1}{2^{2n-1}},\end{aligned}$$

cf. (A.2). Clearly, (A.1) holds since $\sin^2(k\pi h) = 1$ for $kh = \frac{1}{2}$.

If $kh \neq \frac{1}{2}$, then the product in (A.1) can be written as

$$\begin{aligned}&|\cos(k\pi h)| \prod_{\substack{j=1\\ j\neq k\\ j\neq n+1-k}}^{n}\left|\sin\left(\frac{(j+k)\pi h}{2}\right)\cos\left(\frac{(j+k)\pi h}{2}\right)\right|\\ &= \frac{|\cos(k\pi h)|}{2^{n-2}} \prod_{\substack{j=1\\ j\neq k\\ j\neq n+1-k}}^{n} |\sin((j+k)\pi h)|\\ &= \frac{|\cos(k\pi h)|}{2^{n-2}|\sin(2k\pi h)|}\cdot \prod_{\substack{j=1\\ j\neq n+1-k}}^{n} |\sin((j+k)\pi h)|\\ &= \frac{2\sin(k\pi h)\cos(k\pi h)}{2^{n-1}\sin(k\pi h)\sin(2k\pi h)}\cdot\frac{1}{\sin(k\pi h)}\prod_{j=1}^{n}\sin(j\pi h)\\ &= \frac{n+1}{2^{2n-1}}\,\frac{1}{\sin^2(k\pi h)}.\end{aligned}$$

REFERENCES

[1] O. AXELSSON, *Iterative Solution Methods*, Cambridge University Press, 1994.

[2] B. Beckermann and A. B. J. Kuijlaars, *Superlinear CG convergence for special right-hand sides*, Electron. Trans. Numer. Anal., 14 (2002), pp. 1–19.

[3] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*, Academic Press Inc., San Diego, CA, sixth ed., 2000. Translated from the Russian, Translation edited and with a preface by Alan Jeffrey and Daniel Zwillinger.

[4] A. Greenbaum, *Comparison of splittings used with the conjugate gradient algorithm*, Numer. Math., 33 (1979), pp. 181–193.

[5] ———, *Iterative methods for solving linear systems*, vol. 17 of Frontiers in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

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

[7] A. Greenbaum and L. N. Trefethen, *GMRES/CR and Arnoldi/Lanczos as matrix approximation problems*, SIAM J. Sci. Comput., 15 (1994), pp. 359–368.

[8] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards , 49 (1952), pp. 409–435.

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

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

[11] J. Liesen and P. Tichý, *Behavior of CG and MINRES for symmetric tridiagonal Toeplitz matrices*, Preprint 34-2004, Institute of Mathematics, Technical University of Berlin, 2004. Available at http://www.math.tu-berlin.de/preprints.

[12] ———, *Convergence analysis of Krylov subspace methods*, GAMM Mitt. Ges. Angew. Math. Mech., 27 (2004), pp. 153–173 (2005).

[13] ———, *The worst-case GMRES for normal matrices*, BIT, 44 (2004), pp. 79–98.

[14] J. C. Mason and D. C. Handscomb, *Chebyshev polynomials*, Chapman & Hall/CRC, Boca Raton, FL, 2003.

[15] A. E. Naiman, I. M. Babuška, and H. C. Elman, *A note on conjugate gradient convergence*, Numer. Math., 76 (1997), pp. 209–230.

[16] A. E. Naiman and S. Engelberg, *A note on conjugate gradient convergence. II, III*, Numer. Math., 85 (2000), pp. 665–683, 685–696.

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

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

[19] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[20] G. D. Smith, *Numerical solution of partial differential equations*, The Clarendon Press Oxford University Press, New York, second ed., 1978. Finite difference methods, Oxford Applied Mathematics and Computing Science Series.

[21] The MathWorks, Inc., *MATLAB 6.5, Release 13.* Natick, Massachusetts, USA, 2002.

[22] K. Toh, M. Todd, and R. Tütüncü, *SDPT3 – a Matlab software package for semidefinite programming, version 2.1. Interior point methods.* June 2001.

[23] Waterloo Maple, Inc., *Maple 8.0.* Waterloo, Ontario, Canada, 2002.

SIAM J. MATRIX ANAL. APPL.
Vol. 31, No. 2, pp. 853–863

ON BEST APPROXIMATIONS OF POLYNOMIALS IN MATRICES IN THE MATRIX 2-NORM[*]

JÖRG LIESEN[†] AND PETR TICHÝ[‡]

Abstract. We show that certain matrix approximation problems in the matrix 2-norm have uniquely defined solutions, despite the lack of strict convexity of the matrix 2-norm. The problems we consider are generalizations of the ideal Arnoldi and ideal GMRES approximation problems introduced by Greenbaum and Trefethen [*SIAM J. Sci. Comput.*, 15 (1994), pp. 359–368]. We also discuss general characterizations of best approximation in the matrix 2-norm and provide an example showing that a known sufficient condition for uniqueness in these characterizations is not necessary.

Key words. matrix approximation problems, polynomials in matrices, matrix functions, matrix 2-norm, GMRES, Arnoldi's method

AMS subject classifications. 41A52, 15A60, 65F35

DOI. 10.1137/080728299

1. Introduction.

Much of the work in approximation theory concerns the approximation of a given function f on some (compact) set Ω in the complex plane by polynomials. Classical results in this area deal with the *best approximation problem*

$$\min_{p \in \mathcal{P}_m} \|f - p\|_\Omega \,, \tag{1.1}$$

where $\|g\|_\Omega \equiv \max_{z\in\Omega} |g(z)|$ and $\mathcal{P}_m$ denotes the set of polynomials of degree at most m. (Note that since in (1.1) we seek an approximation from a finite-dimensional subspace, the minimum is indeed attained by some polynomial $p_* \in \mathcal{P}_m$.)

Scalar approximation problems of the form (1.1) have been studied since the mid 1850s. Accordingly, numerous results on the existence and uniqueness of the solution as well as estimates for the value of (1.1) are known. Here we consider a problem that at first sight looks similar, but apparently is much less understood: Let f be a function that is analytic in an open neighborhood of the spectrum of a given matrix $A \in \mathbb{C}^{n\times n}$, so that $f(A)$ is well defined, and let $|\cdot|$ be a given matrix norm. Consider the *matrix* approximation problem

$$\min_{p \in \mathcal{P}_m} |f(A) - p(A)| \,. \tag{1.2}$$

Does this problem have a unique solution?

An answer to this question of course depends on the norm used in (1.2). A norm $|\cdot|$ on a vector space $\mathcal{V}$ is called *strictly convex* when for all vectors $v_1, v_2 \in \mathcal{V}$ the equation $|v_1| = |v_2| = \frac{1}{2}|v_1 + v_2|$ implies that $v_1 = v_2$. A geometric interpretation of strict convexity is that the unit sphere in $\mathcal{V}$ with respect to the norm $|\cdot|$ does not

[*] Received by the editors June 24, 2008; accepted for publication (in revised form) by M. Embree May 21, 2009; published electronically July 30, 2009.
http://www.siam.org/journals/simax/31-2/72829.html

[†] Institute of Mathematics, Technical University of Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany (liesen@math.tu-berlin.de). The work of this author was supported by the Heisenberg Program of the Deutsche Forschungsgemeinschaft.

[‡] Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 2, 18207 Prague, Czech Republic (tichy@cs.cas.cz). The work of this author was supported by GAAS grant IAA100300802.

contain any line segments. If $\mathcal{S} \subseteq \mathcal{V}$ is a finite-dimensional subspace, then for any given $v \in \mathcal{V}$ there exists a *unique* $s_* \in \mathcal{S}$ so that

$$|v - s_*| \;=\; \min_{s\in\mathcal{S}} |v - s| \,.$$

A proof of this classical result can be found in most books on approximation theory; see, e.g., [3, Chapter 1]. In particular, if the norm is strictly convex, then (1.2) is guaranteed to have a unique solution as long as the value of (1.2) is positive.

A useful matrix norm that is met in many applications is the matrix 2-norm (or spectral norm), which for a given matrix A is equal to the largest singular value of A. We denote the 2-norm of A by $\|A\|$. This norm is *not* strictly convex, as can be seen from the following simple example: Suppose that we have two matrices $A_1, A_2 \in \mathbb{C}^{n\times n}$ of the form

$$A_1 \;=\; \begin{bmatrix} B & 0 \\ 0 & C \end{bmatrix}, \qquad A_2 \;=\; \begin{bmatrix} B & 0 \\ 0 & D \end{bmatrix},$$

with $\|A_1\| = \|A_2\| = \|B\| \geq \frac{1}{2}\|C + D\|$. Then $\frac{1}{2}\|A_1 + A_2\| = \|B\|$, but whenever $C \neq D$, we have $A_1 \neq A_2$. Consequently, in the case of the matrix 2-norm, the classical uniqueness result mentioned above does not apply, and our question about the uniqueness of the solution of the matrix approximation problem (1.2) is nontrivial.

It is well known that when the function f is analytic in an open neighborhood of the spectrum of the matrix $A \in \mathbb{C}^{n\times n}$, then $f(A)$ is a well-defined complex $n \times n$ matrix. In fact, $f(A) = p_f(A)$, where p_f is a polynomial that depends on the values and possibly the derivatives of f on the spectrum of A. The recent book of Higham [5] gives an extensive overview of definitions, applications, and computational techniques for matrix functions. Our above question now naturally leads to the following mathematical problem: *Let a polynomial b and a nonnegative integer $m < \deg b$ be given. Determine conditions so that the best approximation problem*

$$\min_{p\in\mathcal{P}_m} \|b(A) - p(A)\| \tag{1.3}$$

has a unique solution, where $\|\cdot\|$ is the matrix 2-norm and $\mathcal{P}_m$ denotes the set of polynomials of degree at most m.

When searching the literature we found a number of results on general characterizations of best approximations in normed linear spaces of matrices, e.g., in [7, 9, 15, 16], but just a few papers related to our specific problem. In particular, Greenbaum and Trefethen consider in [4] the two approximation problems

$$\min_{p\in\mathcal{P}_m} \left\|A^{m+1} - p(A)\right\|, \tag{1.4}$$

$$\min_{p\in\mathcal{P}_m} \|I - Ap(A)\|. \tag{1.5}$$

They state that both (1.4) and (1.5) (for nonsingular A) have a unique minimizer.[1] The problem (1.4) is equal to (1.3) with $b(A) = A^{m+1}$. Because of its relation to the convergence of the Arnoldi method [1] for approximating eigenvalues of A, the uniquely defined monic polynomial $z^{m+1} - p_*$ that solves (1.4) is called the $(m+1)$*st*

[1] The statement of uniqueness is true, but the proof given in [4], which was later repeated in [14, Chapter 29], contains a small error at the very end. After the error was spotted by Michael Eiermann, it was fixed by Anne Greenbaum in 2005, but the correction has not been published.

ideal Arnoldi polynomial of A. In a paper that is mostly concerned with algorithmic and computational results, Toh and Trefethen [13] call this polynomial the $(m+1)$*st Chebyshev polynomial of A*. The reason for this terminology is the following: When the matrix A is *normal*, i.e., unitarily diagonalizable, problem (1.4) becomes a scalar approximation problem of the form (1.1) with $f(z) = z^{m+1}$ and Ω being the spectrum of A. The resulting monic polynomial is the $(m+1)$st Chebyshev polynomial on this (discrete) set Ω, i.e., the unique monic polynomial of degree $m+1$ with minimal maximum norm on Ω. In this sense, the matrix approximation problem (1.3) we study here can be considered a generalization of the classical scalar approximation problem (1.1). Some further results on Chebyshev polynomials of matrices are given in [11] and [14, Chapter 29].

The quantity (1.5) can be used for bounding the relative residual norm in the GMRES method [8]; for details see, e.g., [10, 12]. Therefore, the uniquely defined polynomial $1 - z\,p_*$ that solves (1.5) is called the $(m+1)$*st ideal GMRES polynomial of A*.

In this paper we show that, despite the lack of strict convexity of the matrix 2-norm, the approximation problem (1.3) as well as a certain related problem that generalizes (1.5) have a unique minimizer. Furthermore, we discuss some of the above-mentioned general characterizations of best approximations with respect to the 2-norm in linear spaces of matrices. On the example of a Jordan block, we show that a sufficient condition for the uniqueness of the best approximation obtained by Ziętak [15] does not hold. We are not aware that such an example for a nonnormal matrix has been given before.

2. Uniqueness results.

Let $\ell \geq 0$ and $m \geq 0$ be given integers, and consider a given polynomial b of the form

$$b = \sum_{j=0}^{\ell+m+1} \beta_j z^j \in \mathcal{P}_{\ell+m+1}\,.$$

Let us rewrite the approximation problem (1.3) in a more convenient equivalent form:

$$\begin{aligned}
\min_{p\in\mathcal{P}_m} \|b(A) - p(A)\| &= \min_{p\in\mathcal{P}_m} \left\| b(A) - \left(p(A) + \sum_{j=0}^{m} \beta_j A^j \right) \right\| \\
&= \min_{p\in\mathcal{P}_m} \left\| \sum_{j=m+1}^{\ell+m+1} \beta_j A^j - p(A) \right\| \\
&= \min_{p\in\mathcal{P}_m} \left\| A^{m+1} \sum_{j=0}^{\ell} \beta_{j+m+1} A^j - p(A) \right\| . \qquad (2.1)
\end{aligned}$$

The polynomials in (2.1) are of the form $z^{m+1}g + h$, where the polynomial $g \in \mathcal{P}_\ell$ is *given* and $h \in \mathcal{P}_m$ is *sought*. Hence (1.3) is equivalent to the problem

$$\min_{h\in\mathcal{P}_m} \left\| A^{m+1} g(A) + h(A) \right\|, \qquad (2.2)$$

where $g \in \mathcal{P}_\ell$ is a given polynomial or

$$\min_{p\in\mathcal{G}_{\ell,m}^{(g)}} \|p(A)\|, \quad \text{where } \mathcal{G}_{\ell,m}^{(g)} \equiv \left\{ z^{m+1}g + h \,:\, g \in \mathcal{P}_\ell \text{ is given}, h \in \mathcal{P}_m \right\}. \qquad (2.3)$$

With $\ell = 0$ and $g = 1$, (2.3) reduces to (1.4).

Similarly, we may consider the approximation problem

$$(2.4)\quad \min_{p\in\mathcal{H}^{(h)}_{\ell,m}} \|p(A)\|, \quad\text{where } \mathcal{H}^{(h)}_{\ell,m} \equiv \left\{ z^{m+1}g+h \,:\, h\in\mathcal{P}_m \text{ is given, } g\in\mathcal{P}_\ell \right\}.$$

Setting $m=0$ and $h=1$ in (2.4), we retrieve a problem of the form (1.5).

The problems (2.3) and (2.4) are trivial for $g=0$ and $h=0$, respectively. Both cases are unconstrained minimizations problems, and it is easily seen that the resulting minimum value is zero. In the following we will therefore exclude the cases $g=0$ in (2.3) and $h=0$ in (2.4). Under this assumption, both $\mathcal{G}^{(g)}_{\ell,m}$ and $\mathcal{H}^{(h)}_{\ell,m}$ are subsets of $\mathcal{P}_{\ell+m+1}$, where certain coefficients are *fixed*. In the case of $\mathcal{G}^{(g)}_{\ell,m}$, these are the coefficients at the $\ell+1$ largest powers of z, namely, $z^{m+1},\dots,z^{\ell+m+1}$. For $\mathcal{H}^{(h)}_{\ell,m}$ these are the coefficients at the $m+1$ smallest powers of z, namely, $1,\dots,z^m$.

We start with conditions so that the values of (2.3) and (2.4) are positive for all given nonzero polynomials $g\in\mathcal{P}_\ell$ and $h\in\mathcal{P}_m$, respectively.

LEMMA 2.1. *Consider the approximation problems* (2.3) *and* (2.4), *where* $\ell\ge 0$ *and* $m\ge 0$ *are given integers. Denote by* $d(A)$ *the degree of the minimal polynomial of the given matrix* $A\in\mathbb{C}^{n\times n}$. *Then the following two assertions are equivalent:*

(1) $\min_{p\in\mathcal{G}^{(g)}_{\ell,m}} \|p(A)\| > 0$ *for all nonzero polynomials* $g\in\mathcal{P}_\ell$.

(2) $m+\ell+1 < d(A)$.

If A *is nonsingular, the two assertions are equivalent with*

(3) $\min_{p\in\mathcal{H}^{(h)}_{\ell,m}} \|p(A)\| > 0$ *for all nonzero polynomials* $h\in\mathcal{P}_m$.

Proof. (1) ⇒ (2): We suppose that $m+\ell+1\ge d(A)$ and show that (1) fails to hold. Denote the minimal polynomial of A by Ψ_A. If $m+1\le d(A)\le \ell+m+1$, then there exist uniquely determined polynomials $\widehat{g}\in\mathcal{P}_\ell$, $\widehat{g}\ne 0$, and $\widehat{h}\in\mathcal{P}_m$, so that $z^{m+1}\cdot\widehat{g}+\widehat{h}=\Psi_A$. Hence $\min_{p\in\mathcal{G}^{(g)}_{\ell,m}}\|p(A)\|=0$ for $g=\widehat{g}$. If $0\le d(A)\le m$, let $\widehat{g}$ be any nonzero polynomial of degree at most ℓ. By the division theorem for polynomials,[2] there exist uniquely defined polynomials $q\in\mathcal{P}_{m+\ell+1-d(A)}$ and $h\in\mathcal{P}_{m-1}$, so that $z^{m+1}\cdot\widehat{g}=q\cdot\Psi_A-h$, or, equivalently, $z^{m+1}\cdot\widehat{g}+h=q\cdot\Psi_A$. Hence $A^{m+1}\widehat{g}(A)+h(A)=0$, which means that $\min_{p\in\mathcal{G}^{(g)}_{\ell,m}}\|p(A)\|=0$ for the nonzero polynomial $g=\widehat{g}\in\mathcal{P}_\ell$.

(2) ⇒ (1): If $m+\ell+1<d(A)$, then $\mathcal{G}^{(g)}_{\ell,m}\subset\mathcal{P}_{m+\ell+1}$ implies $\min_{p\in\mathcal{G}^{(g)}_{\ell,m}}\|p(A)\|>0$ for every nonzero polynomial $g\in\mathcal{P}_\ell$.

(2) ⇒ (3): If $m+\ell+1<d(A)$, then $\mathcal{H}^{(h)}_{\ell,m}\subset\mathcal{P}_{m+\ell+1}$ implies $\min_{p\in\mathcal{H}^{(h)}_{\ell,m}}\|p(A)\|>0$ for every nonzero polynomial $h\in\mathcal{P}_m$.

(3) ⇒ (2): For this implication we use that A is nonsingular. Suppose that (2) does not hold, i.e., that $0\le d(A)\le m+\ell+1$. Then there exist uniquely defined polynomials $\widehat{g}\in\mathcal{P}_\ell$ and $\widehat{h}\in\mathcal{P}_m$ such that $z^{m+1}\cdot\widehat{g}+\widehat{h}=\Psi_A$. Since A is assumed to be nonsingular, we must have $\widehat{h}\ne 0$. Consequently, $\min_{p\in\mathcal{H}^{(h)}_{\ell,m}}\|p(A)\|=0$ for the nonzero polynomial $h=\widehat{h}\in\mathcal{P}_m$. □

In the following Theorem 2.2, we show that the problem (2.3) has a uniquely defined minimizer when the value of this problem is positive (and not zero). In the previous lemma we have shown that $m+\ell+1<d(A)$ is necessary and sufficient so that the value of (2.3) is positive *for all* nonzero polynomials $g\in\mathcal{P}_\ell$. However, it is possible that *for some* nonzero polynomial $g\in\mathcal{P}_\ell$ the value of (2.3) is positive even when $m+1\le d(A)\le m+\ell+1$. It is possible to further analyze this special case, but

[2] If f and $g\ne 0$ are polynomials over a field $\mathbb{F}$, then there exist uniquely defined polynomials s and r over $\mathbb{F}$ such that (i) $f=g\cdot s+r$, and (ii) either $r=0$ or $\deg r<\deg g$. If $\deg f\ge\deg g$, then $\deg f=\deg g+\deg s$. For a proof of this standard result, see, e.g., [6, Chapter 4].

for the ease of the presentation we simply assume that the value of (2.3) is positive. The same assumption is made in Theorem 2.3 below, where we prove the uniqueness of the minimizer of (2.4) (under the additional assumption that A is nonsingular).

We point out that Lemma 2.1 implies that the approximation problems (1.4) and (1.5) for nonsingular A have positive values if and only if $m+1 < d(A)$. Of course, if $m+1 = d(A)$, then the value of both problems is zero. In this case, the $(m+1)$st ideal Arnoldi polynomial that solves (1.4) is equal to the minimal polynomial of A, and the $(m+1)$st ideal GMRES polynomial that solves (1.5) is a scalar multiple of that polynomial.

THEOREM 2.2. *Let $A \in \mathbb{C}^{n\times n}$ be a given matrix, $\ell \geq 0$ and $m \geq 0$ be given integers, and $g \in \mathcal{P}_\ell$ be a given nonzero polynomial. If the value of* (2.3) *is positive, then this problem has a uniquely defined minimizer.*

Proof. The general strategy in the following is similar to the construction in [4, section 5]. We suppose that $q_1 = z^{m+1}g + h_1 \in \mathcal{G}^{(g)}_{\ell,m}$ and $q_2 = z^{m+1}g + h_2 \in \mathcal{G}^{(g)}_{\ell,m}$ are two distinct solutions to (2.3) and derive a contradiction. Suppose that the minimal norm attained by the two polynomials is

$$C = \|q_1(A)\| = \|q_2(A)\|.$$

By assumption, $C > 0$. Define $q \equiv \frac{1}{2}(q_1 + q_2) \in \mathcal{G}^{(g)}_{\ell,m}$, then

$$\|q(A)\| \;\leq\; \frac{1}{2}(\|q_1(A)\| + \|q_2(A)\|) = C.$$

Since C is assumed to be the minimal value of (2.3), we must have $\|q(A)\| = C$. Denote the singular value decomposition of $q(A)$ by

$$q(A) \;=\; V\,\mathrm{diag}(\sigma_1, \ldots, \sigma_n)\,W^*\,. \tag{2.5}$$

Suppose that the maximal singular value $\sigma_1 = C$ of $q(A)$ is J-fold, with left and right singular vectors given by $v_1, \ldots, v_J$ and $w_1, \ldots, w_J$, respectively.

It is well known that the 2-norm for vectors $v \in \mathbb{C}^n$, $\|v\| \equiv (v^*v)^{1/2}$, is strictly convex. For each w_j, $1 \leq j \leq J$, we have

$$C = \|q(A)w_j\| \leq \frac{1}{2}\left(\|q_1(A)w_j\| + \|q_2(A)w_j\|\right) \leq C\,,$$

which implies

$$\|q_1(A)w_j\| = \|q_2(A)w_j\| = C, \qquad 1 \leq j \leq J\,.$$

By the strict convexity of the vector 2-norm,

$$q_1(A)w_j = q_2(A)w_j\,, \qquad 1 \leq j \leq J\,.$$

Similarly, one can show that

$$q_1(A)^*v_j = q_2(A)^*v_j\,, \qquad 1 \leq j \leq J.$$

Thus,

$$(q_2(A) - q_1(A))w_j = 0, \qquad (q_2(A) - q_1(A))^*v_j = 0, \qquad 1 \leq j \leq J. \tag{2.6}$$

By assumption, $q_2 - q_1 \in \mathcal{P}_m$ is a nonzero polynomial. By the division theorem for polynomials (see footnote 2), there exist uniquely defined polynomials s and r, with $\deg s \leq \ell + m + 1$ and $\deg r < \deg(q_2 - q_1) \leq m$ (or $r = 0$), so that

$$z^{m+1} g \; = \; (q_2 - q_1) \cdot s + r \, .$$

Hence we have shown that for the given polynomials $q_2 - q_1$ and g there exist polynomials s and r such that

$$\widetilde{q} \;\equiv\; (q_2 - q_1) \cdot s \;=\; z^{m+1} g - r \;\in\; \mathcal{G}^{(g)}_{\ell,m}.$$

Since $g \neq 0$, we must have $\widetilde{q} \neq 0$. For a fixed $\epsilon \in (0, 1)$, consider the polynomial

$$q_\epsilon \;=\; (1 - \epsilon) q + \epsilon \widetilde{q} \;\in\; \mathcal{G}^{(g)}_{\ell,m} \, .$$

By (2.6),

$$\widetilde{q}(A) w_j = 0, \qquad \widetilde{q}(A)^* v_j = 0, \qquad 1 \leq j \leq J,$$

and thus

$$\begin{aligned} q_\epsilon(A)^* q_\epsilon(A) w_j &= (1-\epsilon) q_\epsilon(A)^* q(A) w_j = (1-\epsilon) C q_\epsilon(A)^* v_j \\ &= (1-\epsilon)^2 C q(A)^* v_j = (1-\epsilon)^2 C^2 w_j \, , \end{aligned}$$

which shows that $w_1, \dots, w_J$ are right singular vectors of $q_\epsilon(A)$ corresponding to the singular value $(1 - \epsilon)C$. Note that $(1 - \epsilon)C < C$ since $C > 0$.

Now there are two cases: Either $\|q_\epsilon(A)\| = (1 - \epsilon)C$, or $(1 - \epsilon)C$ is not the largest singular value of $q_\epsilon(A)$. In the first case we have a contradiction to the fact that C is the minimal value of (2.3). Therefore, the second case must hold. In that case, none of the vectors $w_1, \dots, w_J$ correspond to the largest singular value of $q_\epsilon(A)$. Using this fact and the singular value decomposition (2.5), we get

$$\begin{aligned} \|q_\epsilon(A)\| &= \|q_\epsilon(A) W\| \\ &= \|q_\epsilon(A)[w_{J+1}, \dots, w_n]\| \\ &= \|(1-\epsilon) q(A)[w_{J+1}, \dots, w_n] \, + \, \epsilon \widetilde{q}(A)[w_{J+1}, \dots, w_n]\| \\ &\leq (1-\epsilon) \, \|[v_{J+1}, \dots, v_n] \, \mathrm{diag}(\sigma_{J+1}, \dots, \sigma_n)\| \, + \, \epsilon \|\widetilde{q}(A)[w_{J+1}, \dots, w_n]\| \\ &\leq (1-\epsilon) \sigma_{J+1} \, + \, \epsilon \|\widetilde{q}(A)[w_{J+1}, \dots, w_n]\| \, . \end{aligned} \tag{2.7}$$

Note that the norm $\|\widetilde{q}(A)[w_{J+1}, \dots, w_n]\|$ in (2.7) does not depend on the choice of ϵ and that (2.7) goes to σ_{J+1} as ϵ goes to zero. Since $\sigma_J > \sigma_{J+1}$, one can find a positive $\epsilon_* \in (0, 1)$ such that (2.7) is less than σ_J for all $\epsilon \in (0, \epsilon_*)$. Any of the corresponding polynomials q_ϵ gives a matrix $q_\epsilon(A)$ whose norm is less than σ_J. This contradiction finishes the proof. □

In the following theorem we prove that the problem (2.4), and hence in particular the problem (1.5), has a uniquely defined minimizer.

THEOREM 2.3. *Let $A \in \mathbb{C}^{n \times n}$ be a given nonsingular matrix, $\ell \geq 0$ and $m \geq 0$ be given integers, and $h \in \mathcal{P}_m$ be a given nonzero polynomial. If the value of* (2.4) *is positive, then this problem has a uniquely defined minimizer.*

Proof. Most parts of the following proof are analogous to the proof of Theorem 2.2 and are stated only briefly. However, the construction of the polynomial q_ϵ used to derive the contradiction is different.

We suppose that $q_1 = z^{m+1}g_1+h \in \mathcal{H}^{(h)}_{\ell,m}$ and $q_2 = z^{m+1}g_2+h \in \mathcal{H}^{(h)}_{\ell,m}$ are two distinct solutions to (2.4) and that the minimal norm attained by them is $C = \|q_1(A)\| = \|q_2(A)\|$. By assumption, $C > 0$. Define $q \equiv \frac{1}{2}(q_1 + q_2) \in \mathcal{H}^{(h)}_{\ell,m}$; then $\|q(A)\| = C$. Denote the singular value decomposition of $q(A)$ by $q(A) = V\text{diag}(\sigma_1, \ldots, \sigma_n)W^*$, and suppose that the maximal singular value $\sigma_1 = C$ of $q(A)$ is J-fold, with left and right singular vectors given by $v_1, \ldots, v_J$ and $w_1, \ldots, w_J$, respectively. As previously, we can show that

$$(q_2(A) - q_1(A))w_j = 0, \qquad (q_2(A) - q_1(A))^*v_j = 0, \qquad 1 \le j \le J.$$

Since A is nonsingular and $q_2 - q_1 = z^{m+1}(g_2 - g_1)$, these relations imply that

$$(g_2(A) - g_1(A))w_j = 0, \qquad (g_2(A) - g_1(A))^*v_j = 0, \qquad 1 \le j \le J. \tag{2.8}$$

By assumption, $0 \neq g_2 - g_1 \in \mathcal{P}_\ell$. Hence there exists an integer d, $0 \le d \le \ell$, so that

$$g_2 - g_1 \;=\; \sum_{i=d}^{\ell} \gamma_i z^i\,, \;\text{ with }\; \gamma_d \neq 0\,.$$

Now define

$$\widetilde{g} \;\equiv\; z^{-d}(g_2 - g_1) \;\in\; \mathcal{P}_{\ell-d}\,.$$

By construction, $\widetilde{g}$ is a polynomial with a nonzero constant term. Furthermore, define

$$\widehat{h} \;\equiv\; z^{-m-1-\ell+d}\,h \qquad\text{and}\qquad \widehat{g} \;\equiv\; z^{-\ell+d}\,\widetilde{g}\,.$$

After a formal change of variables $z^{-1} \mapsto y$, we obtain

$$\widehat{h}(y) \;\in\; \mathcal{P}_{m+1+\ell-d} \qquad\text{and}\qquad \widehat{g}(y) \;\in\; \mathcal{P}_{\ell-d} \setminus \mathcal{P}_{\ell-d-1}\,.$$

(Here $\mathcal{P}_{-1} \equiv \emptyset$ in case $d = \ell$.) By the division theorem for polynomials (see footnote 2), there exist uniquely defined polynomials $s(y)$ and $r(y)$ with $\deg s \le m+1$ (since $\widehat{g} \neq 0$ is of exact degree $\ell - d$) and $\deg r < \ell - d$ (or $r = 0$) such that

$$\widehat{h}(y) \;=\; \widehat{g}(y) \cdot s(y) - r(y)\,.$$

We now multiply the preceding equation by $y^{-m-1-\ell+d}$, which gives

$$y^{-m-1-\ell+d}\,\widehat{h}(y) \;=\; \left(y^{-\ell+d}\widehat{g}(y)\right) \cdot \left(y^{-m-1}s(y)\right) - y^{-m-1}\,\left(y^{-\ell+d}\,r(y)\right)\,.$$

Since $y^{-1} = z$, this equation is equivalent to

$$h \;=\; \widetilde{g} \cdot \widetilde{s} - z^{m+1}\,\widetilde{r},$$

where $\widetilde{s} \in \mathcal{P}_{m+1}$ and $\widetilde{r} \in \mathcal{P}_{\ell-d-1}$. Hence we have shown that for the given polynomials h and $\widetilde{g}$ there exist polynomials $\widetilde{s} \in \mathcal{P}_{m+1}$ and $\widetilde{r} \in \mathcal{P}_{\ell-d-1}$ such that

$$\widetilde{q} \;\equiv\; \widetilde{g} \cdot \widetilde{s} \;=\; z^{m+1}\,\widetilde{r} + h \;\;\in\;\; \mathcal{H}^{(h)}_{\ell,m}.$$

For a fixed $\epsilon \in (0, 1)$, consider

$$q_\epsilon \;=\; (1 - \epsilon)q + \epsilon\widetilde{q} \;\in\; \mathcal{H}^{(h)}_{\ell,m}\,.$$

Since $\widetilde{q} = \widetilde{s}z^{-d}(g_2 - g_1)$, (2.8) implies that

$$\widetilde{q}(A)w_j = 0, \qquad \widetilde{q}(A)^* v_j = 0, \qquad 1 \le j \le J,$$

which can be used to show that

$$q_\epsilon(A)^* q_\epsilon(A) w_j = (1-\epsilon)^2 C^2 w_j, \qquad 1 \le j \le J.$$

Now the same argument as in the proof of Theorem 2.2 gives a contradiction to the original assumption that $q_2 \ne q_1$. □

Remark 2.4. Similarly as in Lemma 2.1, the assumption of nonsingularity in the previous theorem is in general necessary. In other words, when A is singular the approximation problem (2.4) might have more than one solution even when the value of (2.4) is positive. The following example demonstrating this fact was pointed out to us by Ziętak: Consider a normal matrix $A = U\Lambda U^*$, where $U^*U = I$ and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Suppose that A is singular with n distinct eigenvalues, and $\lambda_1 = 0$. Furthermore, suppose that $h \in \mathcal{P}_m$ is any given polynomial that satisfies $h(0) \ne 0$ and $|h(0)| > |h(\lambda_j)|$ for $j = 2, \ldots, n$. Then for *any* integer $\ell \ge 0$,

$$\min_{p \in \mathcal{H}_{\ell,m}^{(h)}} \|p(A)\| = \min_{g \in \mathcal{P}_\ell} \max_j \left|\lambda_j^{m+1} g(\lambda_j) + h(\lambda_j)\right| \; = \; |h(0)| \; > \; 0\,.$$

One solution of this problem is given by the polynomial $g = 0$. Moreover, the minimum value is attained for any polynomial $g \in \mathcal{P}_\ell$ that satisfies

$$\min_{g \in \mathcal{P}_\ell} \max_{2 \le j \le n} \left|\lambda_j^{m+1} g(\lambda_j) + h(\lambda_j)\right| \; \le |h(0)|\,,$$

i.e., for any polynomial $g \in \mathcal{P}_\ell$ that is close enough to the zero polynomial.

3. Characterization of best approximation with respect to the matrix 2-norm.

In this section we discuss general characterizations of best approximation in linear spaces of matrices with respect to the matrix 2-norm obtained by Ziętak [15, 16], and we give an example from our specific problem. To state Ziętak's results, we need some notation. Suppose that we are given m matrices $A_1, \ldots, A_m \in \mathbb{C}^{n\times n}$ that are linearly independent in $\mathbb{C}^{n\times n}$. We assume that $1 \le m < n^2$ to avoid trivialities. Denote $\mathbb{A} \equiv \operatorname{span}\{A_1, \ldots, A_m\}$, which is an m-dimensional subspace of $\mathbb{C}^{n\times n}$. As above, let $\|\cdot\|$ denote the matrix 2-norm. For a given matrix $B \in \mathbb{C}^{n\times n}\backslash\mathbb{A}$, we consider the best approximation (or matrix nearness) problem

$$\min_{M \in \mathbb{A}} \|B - M\|\,. \tag{3.1}$$

A matrix $A_* \in \mathbb{A}$ for which this minimum is achieved (such a matrix exists, since $\mathbb{A}$ is finite dimensional) is called a *spectral approximation of B from the subspace* $\mathbb{A}$. The corresponding matrix $R(A_*) = B - A_*$ is called a *residual matrix.*

The approximation problems (2.3) and (2.4) studied in the previous section are both special cases of (3.1). In the case of (2.3),

$$B = A^{m+1}g(A), \;\text{ where } g \in \mathcal{P}_\ell \text{ is given and } \quad \mathbb{A} = \{I, A, \ldots, A^m\}\,,$$

while in case of (2.4),

$$B = h(A), \;\text{ where } h \in \mathcal{P}_m \text{ is given and } \quad \mathbb{A} = \left\{A^{m+1}, \ldots, A^{\ell+m+1}\right\}.$$

We have shown that when the values of these approximation problems are positive (which is true if $\ell + m + 1 < d(A)$), for both these problems there exists a uniquely defined spectral approximation A_* of B from the subspace $\mathbb{A}$ (in the case of (2.4), we have assumed that A is nonsingular). Another approximation problem that fits into the template (3.1) arises in the convergence theory for Arnoldi eigenvalue iterations in [2], where the authors study the problem of minimizing $\|I - h(A)p(A)\|$ over polynomials $p \in \mathcal{P}_{\ell-2m}$, $\ell \geq 2m \geq 2$, and $h \in \mathcal{P}_m$ is a given polynomial.

In general, the spectral approximation of a matrix $B \in \mathbb{C}^{n\times n}$ from a subspace $\mathbb{A} \subset \mathbb{C}^{n\times n}$ is not unique. Ziętak [15] studies the problem (3.1) and gives a general characterization of spectral approximations based on the singular value decomposition of the residual matrices. In particular, combining results of [16] with [15, Theorem 4.3] yields the following sufficient condition for uniqueness of the spectral approximation.

Lemma 3.1. *In the notation established above, let A_* be a spectral approximation of B from the subspace $\mathbb{A}$. If the residual matrix $R(A_*) = B - A_*$ has an n-fold singular value, then the spectral approximation A_* of B from the subspace $\mathbb{A}$ is unique.*

It is quite obvious that the sufficient condition in Lemma 3.1 is, in general, not necessary. To construct a nontrivial counterexample, we recall that the dual norm to the matrix 2-norm is the trace norm (also called energy norm or c_1-norm)

$$|||\, M \,||| \equiv \sum_{j=1}^{r} \sigma_j(M)\,, \tag{3.2}$$

where $\sigma_1(M), \ldots, \sigma_r(M)$ denote the singular values of the matrix $M \in \mathbb{C}^{n\times n}$ with $\operatorname{rank}(M) = r$. For $X \in \mathbb{C}^{n\times n}$ and $Y \in \mathbb{C}^{n\times n}$ we define the inner product $\langle X, Y\rangle \equiv \operatorname{trace}(Y^*X)$. Using this notation, we can state the following result, which is given in [16, p. 173].

Lemma 3.2. *The matrix $A_* \in \mathbb{A}$ is a spectral approximation of B from the subspace $\mathbb{A}$ if and only if there exists a matrix $Z \in \mathbb{C}^{n\times n}$, with $|||\, Z \,||| = 1$ such that*

$$\langle Z, X\rangle = 0 \ \textit{ for all } \ X \in \mathbb{A} \qquad \textit{and} \qquad \operatorname{Re}\, \langle Z, B - A_*\rangle = \|B - A_*\|\,. \tag{3.3}$$

Remark 3.3. Lemmas 3.1 and 3.2 are both stated here for square complex matrices. Originally, Lemma 3.1 is formulated in [15] for real rectangular matrices and Lemma 3.2 given in [16] for square complex matrices. A further generalization to rectangular complex matrices seems possible, but it is out of our focus here.

Based on Lemma 3.2 we can prove the following result.

Theorem 3.4. *For $\lambda \in \mathbb{C}$, consider the $n \times n$ Jordan block*

$$J_\lambda \ \equiv \ \begin{pmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{pmatrix}.$$

Then for any nonnegative integer m with $m+1 \leq n$, the solution to the approximation problem (1.4) with $A = J_\lambda$, i.e., the $(m+1)$st ideal Arnoldi (or Chebyshev) polynomial of J_λ, is uniquely defined and given by $(z-\lambda)^{m+1}$.

Proof. With $A = J_\lambda$, the approximation problem (1.4) reads

$$\min_{p\in\mathcal{P}_m} \left\| J_\lambda^{m+1} - p(J_\lambda)\right\|\,. \tag{3.4}$$

In the notation established in this section, we seek a spectral approximation A_* of $B = J_\lambda^{m+1}$ from the subspace $\mathbb{A} = \operatorname{span}\{I, J_\lambda, \dots, J_\lambda^m\}$. We claim that the uniquely defined solution is given by the matrix $A_* = J_\lambda^{m+1} - (J_\lambda - \lambda I)^{m+1}$. For this matrix A_* we get

$$B - A_* = J_\lambda^{m+1} - A_* \;=\; (J_\lambda - \lambda I)^{m+1} = J_0^{m+1}\,.$$

For $m+1 = n$, $A_* = J_\lambda^n - (J_\lambda - \lambda I)^n = J_\lambda^n$ yields $B - A_* = J_0^n = 0$. The corresponding ideal Arnoldi polynomial of J_λ is uniquely defined and equal to $(z-\lambda)^n$, the minimal polynomial of J_λ.

For $m+1 < n$, the value of (3.4) is positive, and hence Theorem 2.2 ensures that the spectral approximation of J_λ^{m+1} from the subspace $\mathbb{A}$ is uniquely defined. We prove our claim using Lemma 3.2. Define $Z \equiv e_1 e_{m+2}^T$ then $|||Z||| = 1$,

$$\left\langle Z, J_\lambda^j \right\rangle \;=\; 0 \quad \text{for } j = 0, \dots, m\,,$$

and $\|B - A_*\| = \|J_0^{m+1}\| = 1$, so that

$$\langle Z, B - A_* \rangle \;=\; \left\langle Z, J_0^{m+1} \right\rangle \;=\; 1 \;=\; \|B - A_*\|,$$

which shows (3.3) and completes the proof. □

The proof of this theorem shows that the residual matrix of the spectral approximation A_* of $B = J_\lambda^{m+1}$ from the subspace $\mathbb{A} = \operatorname{span}\{I, J_\lambda, \dots, J_\lambda^m\}$ is given by $R(A_*) = J_0^{m+1}$. This matrix $R(A_*)$ has $m+1$ singular values equal to zero and $n-m-1$ singular values equal to one. Hence, for $m+1 < n$, the maximal singular value of the residual matrix is not n-fold, and the sufficient condition of Lemma 3.1 does not hold. Nevertheless, the spectral approximation of B from the subspace $\mathbb{A}$ is unique whenever $m+1 < n$.

As shown above, for $m = 0, 1, \dots, n-1$, the polynomial $(z-\lambda)^{m+1}$ solves the ideal Arnoldi approximation problem (1.4) for $A = J_\lambda$. For $\lambda \neq 0$, we can write

$$(z-\lambda)^{m+1} \;=\; (-\lambda)^{m+1} \cdot \left(1 - \lambda^{-1} z\right)^{m+1}\,.$$

Note that the rightmost factor is a polynomial that has value one at the origin. Hence it is a candidate for the solution of the ideal GMRES approximation problem (1.5) for $A = J_\lambda$. More generally, it is tempting to assume that the $(m+1)$st ideal GMRES polynomial for a given matrix A is equal to a scaled version of its $(m+1)$st ideal Arnoldi (or Chebyshev) polynomial. However, this assumption is false, as we can already see in case $A = J_\lambda$. As shown in [10], the determination of the ideal GMRES polynomials for a Jordan block is an intriguing problem, since these polynomials can become quite complicated. They are of the simple form $(1 - \lambda^{-1} z)^{m+1}$ if and only if $0 \le m+1 < n/2$ and $|\lambda| \ge \varrho_{m+1,n-m-1}^{-1}$; cf. [10, Theorem 3.2]. Here $\varrho_{k,n}$ denotes the radius of the polynomial numerical hull of degree k of an $n \times n$ Jordan block (this radius is independent of the eigenvalue λ).

Now let n be even, and consider $m+1 = n/2$. If $|\lambda| \le 2^{-2/n}$, the ideal GMRES polynomial of degree $n/2$ of J_λ is equal to the constant polynomial 1. If $|\lambda| \ge 2^{-2/n}$, the ideal GMRES polynomial of degree $n/2$ of J_λ is equal to

$$\frac{2}{4\lambda^n + 1} + \frac{4\lambda^n - 1}{4\lambda^n + 1}\left(1 - \lambda^{-1} z\right)^{n/2}; \tag{3.5}$$

cf. [10, p. 465]. Obviously, neither the polynomial 1 nor the polynomial (3.5) are scalar multiples of $(z-\lambda)^{n/2}$, the ideal Arnoldi polynomial of degree $n/2$ of J_λ.

Acknowledgments. We thank Krystyna Ziętak for many discussions and suggestions that helped to improve the content of this paper. We also thank Mark Embree, Shmuel Friedland, Anne Greenbaum, Nick Trefethen, and two anonymous referees for their helpful comments.

REFERENCES

[1] W. E. Arnoldi, *The principle of minimized iteration in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 17–29.

[2] C. A. Beattie, M. Embree, and D. C. Sorensen, *Convergence of polynomial restart Krylov methods for eigenvalue computations*, SIAM Rev., 47 (2005), pp. 492–515.

[3] E. W. Cheney, *Introduction to Approximation Theory*, McGraw-Hill, New York, 1966.

[4] A. Greenbaum and L. N. Trefethen, *GMRES/CR and Arnoldi/Lanczos as matrix approximation problems*, SIAM J. Sci. Comput., 15 (1994), pp. 359–368.

[5] N. J. Higham, *Functions of Matrices: Theory and Computation*, SIAM, Philadelphia, 2008.

[6] K. Hoffman and R. Kunze, *Linear Algebra*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1971.

[7] K. K. Lau and W. O. J. Riha, *Characterization of best approximations in normed linear spaces of matrices by elements of finite-dimensional linear subspaces*, Linear Algebra Appl., 35 (1981), pp. 109–120.

[8] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[9] I. Singer, *Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces*, Springer-Verlag, Berlin, 1970.

[10] P. Tichý, J. Liesen, and V. Faber, *On worst-case GMRES, ideal GMRES, and the polynomial numerical hull of a Jordan block*, Electron. Trans. Numer. Anal., 26 (2007), pp. 453–473 (electronic).

[11] K.-C. Toh, *Matrix Approximation Problems and Nonsymmetric Iterative Methods*, Ph.D. thesis, Cornell University, Ithaca, NY, 1996.

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

[13] K.-C. Toh and L. N. Trefethen, *The Chebyshev polynomials of a matrix*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 400–419.

[14] L. N. Trefethen and M. Embree, *Spectra and Pseudospectra*, Princeton University Press, Princeton, NJ, 2005.

[15] K. Ziętak, *Properties of linear approximations of matrices in the spectral norm*, Linear Algebra Appl., 183 (1993), pp. 41–60.

[16] K. Ziętak, *On approximation problems with zero-trace matrices*, Linear Algebra Appl., 247 (1996), pp. 169–183.

Electronic Transactions on Numerical Analysis.
Volume 41, pp. 159-166, 2014.
Copyright © 2014, Kent State University.
ISSN 1068-9613.

MAX-MIN AND MIN-MAX APPROXIMATION PROBLEMS FOR NORMAL MATRICES REVISITED[*]

JÖRG LIESEN[†] AND PETR TICHÝ[‡]

In memory of Bernd Fischer

Abstract. We give a new proof of an equality of certain max-min and min-max approximation problems involving normal matrices. The previously published proofs of this equality apply tools from matrix theory, (analytic) optimization theory, and constrained convex optimization. Our proof uses a classical characterization theorem from approximation theory and thus exploits the link between the two approximation problems with normal matrices on the one hand and approximation problems on compact sets in the complex plane on the other.

Key words. matrix approximation problems, min-max and max-min approximation problems, best approximation, normal matrices

AMS subject classifications. 41A10, 30E10, 49K35, 65F10

1. Introduction.

Let A be a real or complex square matrix, i.e., $A \in \mathbb{F}^{n\times n}$ with $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$. Suppose that f and $\varphi_1, \dots, \varphi_k$ are given (scalar) functions so that $f(A) \in \mathbb{F}^{n\times n}$ and $\varphi_1(A), \dots, \varphi_k(A) \in \mathbb{F}^{n\times n}$ are well defined matrix functions in the sense of [9, Definition 1.2]. (In the case $\mathbb{F} = \mathbb{R}$, this requires a subtle assumption which is explicitly stated in (2.4) below.) Let $\mathcal{P}_k(\mathbb{F})$ denote the linear span of the functions $\varphi_1, \dots, \varphi_k$ with coefficients in $\mathbb{F}$ so that in particular $p(A) \in \mathbb{F}^{n\times n}$ for each linear combination $p = \alpha_1\varphi_1 + \cdots + \alpha_k\varphi_k \in \mathcal{P}_k(\mathbb{F})$.

With this notation, the optimality property of many useful methods of numerical linear algebra can be formulated as an approximation problem of the form

$$\min_{p\in\mathcal{P}_k(\mathbb{F})} \|f(A)v - p(A)v\|, \tag{1.1}$$

where $v \in \mathbb{F}^n$ is a given vector and $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{F}^n$. In (1.1) we seek a best approximation (with respect to the given norm) of the vector $f(A)v \in \mathbb{F}^n$ from the subspace of $\mathbb{F}^n$ spanned by the vectors $\varphi_1(A)v, \dots, \varphi_k(A)v$. An example of such a method is the GMRES method [15] for solving the linear algebraic problem $Ax = b$ with $A \in \mathbb{F}^{n\times n}$, $b \in \mathbb{F}^n$, and the initial guess $x_0 \in \mathbb{F}^n$. Its optimality property is of the form (1.1) with $f(z) = 1$, $\varphi_i(z) = z^i$, for $i = 1, \dots, k$, and $v = b - Ax_0$.

If the given vector v has unit norm, which usually can be assumed without loss of generality, then an upper bound on (1.1) is given by

$$\min_{p\in\mathcal{P}_k(\mathbb{F})} \|f(A) - p(A)\|, \tag{1.2}$$

where $\|\cdot\|$ denotes the matrix norm associated with the Euclidean vector norm, i.e., the matrix 2-norm or spectral norm on $\mathbb{F}^{n\times n}$. In (1.2) we seek a best approximation (with respect to the given norm) of the matrix $f(A) \in \mathbb{F}^{n\times n}$ from the subspace of $\mathbb{F}^{n\times n}$ spanned by the

[*] Received October 22, 2013. Accepted April 14, 2014. Published online on July 4, 2014. Recommended by R. Freund. The work of the second author was supported by the Grant Agency of the Czech Republic under grant No. P201/13-06684 S and by the project M100301201 of the institutional support of the Academy of Sciences of the Czech Republic.

[†] Institute of Mathematics, Technical University of Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany (liesen@math.tu-berlin.de).

[‡] Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 18207 Prague, Czech Republic (tichy@cs.cas.cz).

matrices $\varphi_1(A), \dots, \varphi_k(A)$. An example of this type is the Chebyshev matrix approximation problem with $A \in \mathbb{F}^{n\times n}$, $f(z) = z^k$, and $\varphi_i(z) = z^{i-1}$, $i = 1, \dots, k$. This problem was introduced in [8] and later studied, for example, in [3] and [17].

In order to analyse how close the upper bound (1.2) can possibly be to the quantity (1.1), one can maximize (1.1) over all unit norm vectors $v \in \mathbb{F}^n$ and investigate the sharpness of the inequality

$$\begin{aligned} \max_{\substack{v\in\mathbb{F}^n\\ \|v\|=1}} \min_{p\in\mathcal{P}_k(\mathbb{F})} \|f(A)v - p(A)v\| &\le \min_{p\in\mathcal{P}_k(\mathbb{F})} \|f(A) - p(A)\| \\ &= \min_{p\in\mathcal{P}_k(\mathbb{F})} \max_{\substack{v\in\mathbb{F}^n\\ \|v\|=1}} \|f(A)v - p(A)v\|. \end{aligned} \tag{1.3}$$

From analyses of the GMRES method it is known that the inequality (1.3) can be strict. For example, certain nonnormal matrices $A \in \mathbb{R}^{4\times 4}$ were constructed in [2, 16] for which (1.3) is strict with $k = 3$, $f(z) = 1$, and $\varphi_i(z) = z^i$, $i = 1, 2, 3$. More recently, nonnormal matrices $A \in \mathbb{R}^{2n\times 2n}$, $n \ge 2$, were derived in [4] for which the inequality (1.3) is strict for all $k = 3, \dots, 2n-1$, $f(z) = 1$, and $\varphi_i(z) = z^i$, $i = 1, \dots, k$.

On the other hand, the following result is well known.

THEOREM 1.1. *Under the assumptions made in the first paragraph of the introduction, if $A \in \mathbb{F}^{n\times n}$ is normal, then equality holds in (1.3).*

At least three different proofs of this theorem or variants of it can be found in the literature. Greenbaum and Gurvits proved it for $\mathbb{F} = \mathbb{R}$ using mostly methods from matrix theory; see [7, Section 2] as well as Section 3 below for their formulation of the result. Using (analytic) methods of optimization theory, Joubert proved the equality for the case of the GMRES method with $f(z) = 1$, $\varphi_i(z) = z^i$, $i = 1, \dots, k$, and he distinguished the cases $\mathbb{F} = \mathbb{R}$ and $\mathbb{F} = \mathbb{C}$; see [11, Theorem 4]. Finally, Bellalij, Saad, and Sadok also considered the GMRES case with $\mathbb{F} = \mathbb{C}$, and they applied methods from constrained convex optimization; see [1, Theorem 2.1].

In this paper we present yet another proof of Theorem 1.1, which is rather simple because it fully exploits the link between matrix approximation problems for normal matrices and scalar approximation problems in the complex plane. We observe that when formulating the matrix approximation problems in (1.3) in terms of scalar approximation problems, the proof of Theorem 1.1 reduces to a straightforward application of a well-known characterization theorem of polynomials of best approximation in the complex plane. While the proof of the theorem for $\mathbb{F} = \mathbb{C}$ can be accomplished in just a few lines, the case $\mathbb{F} = \mathbb{R}$ contains some technical details that require additional attention.

The characterization theorem from approximation theory we use in this paper and some of its variants have been stated and applied also in other publications in this context, in particular in [1, Theorem 5.1]. To our knowledge the theorem has, however, not been used to give a simple and direct proof of Theorem 1.1.

Personal note. We have written this paper in memory of our colleague Bernd Fischer, who passed away on July 15, 2013. Bernd's achievements in the analysis of iterative methods for linear algebraic systems using results of approximation theory, including his nowadays classical monograph [5], continue to inspire us in our own work. One of Bernd's last publications in this area (before following other scientific interests), written jointly with Franz Peherstorfer (1950–2009) and published in 2001 in ETNA [6], is also based on a variant of the characterization theorem that we apply in this paper.

2. Characterization theorem and proof of Theorem 1.1.

In order to formulate the characterization theorem of best approximation in the complex plane, we follow the treatment of Rivlin and Shapiro [14] that has been summarized in Lorentz' book [13, Chapter 2].

Let Γ be a compact subset of $\mathbb{F}$, where either $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$, and let $C(\Gamma)$ denote the set of continuous functions on Γ. If Γ consists of finitely many single points (which is the case of interest in this paper), then $g \in C(\Gamma)$ means that the function g has a well defined (finite) value at each point of Γ. For $g \in C(\Gamma)$ we denote the maximum norm on Γ by

$$\|g\|_\Gamma \equiv \max_{z\in\Gamma} |g(z)|.$$

Now let $f \in C(\Gamma)$ and $\varphi_1, \ldots, \varphi_k \in C(\Gamma)$ be given functions with values in $\mathbb{F}$. As above, let $\mathcal{P}_k(\mathbb{F})$ denote the linear span of the functions $\varphi_1, \ldots, \varphi_k$ with coefficients in $\mathbb{F}$. For $p \in \mathcal{P}_k(\mathbb{F})$, define

$$\Gamma(p) \equiv \{z \in \Gamma : \ |f(z) - p(z)| = \|f - p\|_\Gamma\}.$$

A function $p_* = \alpha_1\varphi_1 + \cdots + \alpha_k\varphi_k \in \mathcal{P}_k(\mathbb{F})$ is called a *polynomial of best approximation* for f on Γ when

$$\|f - p_*\|_\Gamma \ = \min_{p\in\mathcal{P}_k(\mathbb{F})} \|f - p\|_\Gamma. \tag{2.1}$$

Under the given assumptions, such a polynomial of best approximation exists; see, e.g., [13, Theorem 1, p. 17]. The following well known result (see, e.g., [13, Theorem 3, p. 22] or [14, pp. 672-674]) characterizes the polynomials of best approximation.

THEOREM 2.1. *In the notation established above, the following two statements are equivalent:*

1. *The function $p_* \in \mathcal{P}_k(\mathbb{F})$ is a polynomial of best approximation for f on Γ.*
2. *For the function $p_* \in \mathcal{P}_k(\mathbb{F})$ there exist ℓ pairwise distinct points $\mu_i \in \Gamma(p_*)$, $i = 1, \ldots, \ell$, where $1 \le \ell \le k + 1$ for $\mathbb{F} = \mathbb{R}$ and $1 \le \ell \le 2k + 1$ for $\mathbb{F} = \mathbb{C}$, and ℓ real numbers $\omega_1, \ldots, \omega_\ell > 0$ with $\omega_1 + \cdots + \omega_\ell = 1$, such that*

$$\sum_{j=1}^{\ell} \omega_j \,[f(\mu_j) - p_*(\mu_j)]\overline{p(\mu_j)} = 0, \quad \textit{for all } p \in \mathcal{P}_k(\mathbb{F}). \tag{2.2}$$

A well known geometric interpretation of the condition (2.2) is that the origin is contained in the convex hull of the points

$$\left\{ \Big([f(\mu) - p_*(\mu)]\overline{\varphi_1(\mu)}, \ldots, [f(\mu) - p_*(\mu)]\overline{\varphi_k(\mu)}\Big) \in \mathbb{F}^k \ : \ \mu \in \Gamma(p_*) \right\};$$

see, e.g., [13, Equation (5), p. 21]. Here we will not use this interpretation but rewrite (2.2) in terms of an algebraic orthogonality condition involving vectors and matrices. Using that condition we will be able to prove Theorem 1.1 in a straightforward way. We will distinguish the cases of complex and real normal matrices because the real case contains some subtleties.

2.1. Proof of Theorem 1.1 for $\mathbb{F} = \mathbb{C}$. Let $A \in \mathbb{C}^{n\times n}$ be normal. Then A is unitarily diagonalizable, $A = Q\Lambda Q^H$ with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and $QQ^H = Q^HQ = I_n$. In the notation established above, let $\Gamma = \{\lambda_1, \ldots, \lambda_n\}$ and suppose that $p_* \in \mathcal{P}_k(\mathbb{C})$ is a polynomial of best approximation for f on Γ so that statement 2 from Theorem 2.1 applies to p_*. With this setting, the matrix approximation problem (1.2) can be seen as the scalar best approximation problem (2.1), i.e.,

$$\min_{p\in\mathcal{P}_k(\mathbb{C})} \|f(A) - p(A)\| = \min_{p\in\mathcal{P}_k(\mathbb{C})} \|f(\Lambda) - p(\Lambda)\| = \min_{p\in\mathcal{P}_k(\mathbb{C})} \|f - p\|_\Gamma\,.$$

Without loss of generality, we may assume that the eigenvalues of A are ordered so that $\lambda_j = \mu_j$ for $j = 1, \ldots, \ell$. We denote

$$\delta \equiv \|f - p_*\|_\Gamma = |f(\lambda_j) - p_*(\lambda_j)|, \quad j = 1, \ldots, \ell.$$

Next, we define the vector

$$v_* \equiv Q\xi, \quad \text{where} \quad \xi \equiv [\xi_1, \ldots, \xi_\ell, 0, \ldots, 0]^T \in \mathbb{C}^n, \quad \xi_j \equiv \sqrt{\omega_j},\ j = 1, \ldots, \ell. \tag{2.3}$$

Since Q is unitary and $\omega_1 + \cdots + \omega_\ell = 1$, we have $\|v_*\| = 1$.

The condition (2.2) can be written as

$$\begin{aligned} 0 &= \sum_{j=1}^{\ell} |\xi_j|^2 \overline{p(\lambda_j)}\,[f(\lambda_j) - p_*(\lambda_j)] = \xi^H p(\Lambda)^H\,[f(\Lambda) - p_*(\Lambda)]\,\xi \\ &= v_*^H p(A)^H\,[f(A) - p_*(A)]\,v_*\,, \quad \text{for all } p \in \mathcal{P}_k(\mathbb{C}), \end{aligned}$$

or, equivalently,

$$f(A)v_* - p_*(A)v_* \perp p(A)v_*\,, \quad \text{for all } p \in \mathcal{P}_k(\mathbb{C}).$$

It is well known that this algebraic orthogonality condition with respect to the Euclidean inner product is equivalent to the optimality condition

$$\|f(A)v_* - p_*(A)v_*\| = \min_{p \in \mathcal{P}_k(\mathbb{C})} \|f(A)v_* - p(A)v_*\|;$$

see, e.g., [12, Theorem 2.3.2].

Using the previous relations we now obtain

$$\begin{aligned} \min_{p \in \mathcal{P}_k(\mathbb{C})} \|f(A) - p(A)\| = \delta &= \left(\sum_{j=1}^{\ell} |\xi_j|^2 \delta^2 \right)^{1/2} \\ &= \left(\sum_{j=1}^{\ell} |\xi_j|^2\,|f(\lambda_j) - p_*(\lambda_j)|^2 \right)^{1/2} \\ &= \|\,[f(\Lambda) - p_*(\Lambda)]\,\xi\| \\ &= \|Q\,[f(\Lambda) - p_*(\Lambda)]\,Q^H Q\xi\| \\ &= \|f(A)v_* - p_*(A)v_*\| \\ &= \min_{p \in \mathcal{P}_k(\mathbb{C})} \|f(A)v_* - p(A)v_*\| \\ &\le \max_{\substack{v \in \mathbb{C}^n \\ \|v\|=1}} \min_{p \in \mathcal{P}_k(\mathbb{C})} \|f(A)v - p(A)v\|. \end{aligned}$$

This is just the reverse of the inequality (1.3) for $\mathbb{F} = \mathbb{C}$, and hence the proof of Theorem 1.1 for $\mathbb{F} = \mathbb{C}$ is complete.

2.2. Proof of Theorem 1.1 for $\mathbb{F} = \mathbb{R}$. If $A \in \mathbb{R}^{n\times n}$ is symmetric, then we can write $A = Q\Lambda Q^T$ with a real diagonal matrix Λ and a real orthogonal matrix Q. The proof presented in the previous section also works in this case. In particular, for a real matrix Q, the vector $v_* = Q\xi$ constructed in (2.3) is real, and for a real matrix A, the maximization in (1.3) is performed over $v \in \mathbb{R}^n$.

From now on we consider a general normal matrix $A \in \mathbb{R}^{n\times n}$. In the spectral decomposition $A = Q\Lambda Q^H$, the diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and the unitary matrix Q are in general complex. Since this would lead to a complex vector $v_* = Q\xi$ in (2.3), the previous proof requires some modifications.

As above, let $\Gamma = \{\lambda_1, \ldots, \lambda_n\}$. Since A is real, the set Γ may contain non-real points (appearing in complex conjugate pairs), and thus we must allow complex-valued functions $f \in C(\Gamma)$ and $\varphi_1, \ldots, \varphi_k \in C(\Gamma)$. This means that we must work with Theorem 2.1 for $\mathbb{F} = \mathbb{C}$, although A is real. However, we will assume that for each eigenvalue λ_j of A the given functions f and $\varphi_1, \ldots, \varphi_k$ satisfy

$$\overline{f(\lambda_j)} = f(\overline{\lambda}_j) \quad \text{and} \quad \overline{\varphi_i(\lambda_j)} = \varphi_i(\overline{\lambda}_j), \quad i = 1, \ldots, k. \tag{2.4}$$

This is a natural assumption for real matrices A since it guarantees that the matrices $f(A)$ and $\varphi_1(A), \ldots, \varphi_k(A)$ are real as well; see [9, Remark 1.9] (for analytic functions it is actually a necessary and sufficient condition; see [9, Theorem 1.18]).

Now let $q_* = \sum_{i=1}^{k} \alpha_i \varphi_i \in \mathcal{P}_k(\mathbb{C})$ be a polynomial of best approximation for f on Γ. Then, for any eigenvalue λ_j of A,

$$\Big|f(\lambda_j) - \sum_{i=1}^{k} \alpha_i \varphi_i(\lambda_j)\Big| = \Big|\overline{f(\lambda_j)} - \sum_{i=1}^{k} \overline{\alpha}_i \overline{\varphi_i(\lambda_j)}\Big| = \Big|f(\overline{\lambda}_j) - \sum_{i=1}^{k} \overline{\alpha}_i \varphi_i(\overline{\lambda}_j)\Big|.$$

Since both λ_j and $\overline{\lambda}_j$ are elements of Γ, we see that also $\overline{q}_* \equiv \sum_{i=1}^{k} \overline{\alpha}_i \varphi_i$ is a polynomial of best approximation for f on Γ. Denote

$$\delta \equiv \|f - q_*\|_\Gamma = \|f - \overline{q}_*\|_\Gamma \,,$$

then for any $0 \le \alpha \le 1$ we obtain

$$\begin{aligned} \delta &\le \|f - \alpha q_* - (1-\alpha)\overline{q}_*\|_\Gamma = \|\alpha(f - q_*) + (1-\alpha)(f - \overline{q}_*)\|_\Gamma \\ &\le \alpha \,\|f - q_*\|_\Gamma + (1-\alpha)\,\|f - \overline{q}_*\|_\Gamma = \delta, \end{aligned}$$

which shows that any polynomial of the form $\alpha q_* + (1-\alpha)\overline{q}_*$, $0 \le \alpha \le 1$, is also a polynomial of best approximation for f on Γ. In particular, for $\alpha = \frac{1}{2}$ we obtain the *real* polynomial of best approximation

$$p_* \equiv \frac{1}{2}(q_* + \overline{q}_*) \;\in\; \mathcal{P}_k(\mathbb{R}).$$

Using $p_* \in \mathcal{P}_k(\mathbb{R})$ and (2.4) we get

$$|f(z) - p_*(z)| = |\overline{f(z) - p_*(z)}| = |f(\overline{z}) - p_*(\overline{z})|, \quad \text{for all } z \in \Gamma.$$

Therefore, the set $\Gamma(p_*)$ of all points z which satisfy $|f(z) - p_*(z)| = \|f - p_*\|_\Gamma$ is symmetric with respect to the real axis, i.e., $z \in \Gamma(p_*)$ if and only if $\overline{z} \in \Gamma(p_*)$.

For simplicity of notation we denote

$$\zeta_p(z) \equiv [f(z) - p_*(z)]\overline{p(z)}.$$

In the definition of $\zeta_p(z)$ we indicate only its dependence on p and z since f is a given function and p_* is fixed. If $p \in \mathcal{P}_k(\mathbb{R})$, then the corresponding function $\zeta_p(z)$ satisfies $\overline{\zeta_p(z)} = \zeta_p(\overline{z})$ for all $z \in \Gamma$.

Now, Theorem 2.1 (with $\mathbb{F} = \mathbb{C}$) implies the existence of a set

$$G_* \equiv \{\mu_1, \dots, \mu_\ell\} \subseteq \Gamma(p_*) \subseteq \Gamma,$$

and the existence of positive real numbers $\omega_1, \dots, \omega_\ell$ with $\sum_{j=1}^{\ell} \omega_j = 1$ such that

$$\sum_{j=1}^{\ell} \omega_j \, \zeta_p(\mu_j) = 0, \quad \text{for all } p \in \mathcal{P}_k(\mathbb{R}), \tag{2.5}$$

where we have used that $\mathcal{P}_k(\mathbb{R}) \subset \mathcal{P}_k(\mathbb{C})$. To define a convenient real vector v_* similar to the construction leading to (2.3), we will "symmetrize" the condition (2.5) with respect to the real axis.

Taking complex conjugates in (2.5) and using that $\zeta_p(z) = \zeta_p(\overline{z})$ for any $z \in \Gamma$, we obtain another relation of the form

$$\sum_{j=1}^{\ell} \omega_j \, \zeta_p(\overline{\mu}_j) = 0, \quad \text{for all } p \in \mathcal{P}_k(\mathbb{R}),$$

and therefore

$$\frac{1}{2} \sum_{j=1}^{\ell} \omega_j \, \zeta_p(\mu_j) + \frac{1}{2} \sum_{j=1}^{\ell} \omega_j \, \zeta_p(\overline{\mu}_j) = 0, \quad \text{for all } p \in \mathcal{P}_k(\mathbb{R}). \tag{2.6}$$

Here (2.6) is the desired "symmetrized" condition. We now define the set

$$G_*^{\mathrm{sym}} \equiv \{\theta_1, \dots, \theta_m\} \equiv G_* \cup \overline{G}_*,$$

where each $\theta_i \in G_*^{\mathrm{sym}}$ corresponds to some μ_j or $\overline{\mu}_j$, and clearly $\ell \leq m \leq 2\ell$. (The exact value of m is unimportant for our construction.) Writing the condition (2.6) as a single sum over all points from G_*^{sym}, we get

$$\sum_{i=1}^{m} \widetilde{\omega}_i \, \zeta_p(\theta_i) = 0, \quad \text{for all } p \in \mathcal{P}_k(\mathbb{R}), \tag{2.7}$$

where the coefficients $\widetilde{\omega}_i$ are defined as follows.

If $\mu_j \in \mathbb{R}$, then $\zeta_p(\mu_j)$ appears in both sums in (2.6) with the same coefficient $\omega_j/2$. Since $\theta_i = \mu_j \in \mathbb{R}$, the term $\zeta_p(\theta_i)$ appears in (2.7) with the coefficient $\widetilde{\omega}_i = \omega_j$.

If $\mu_j \notin \mathbb{R}$ and $\overline{\mu}_j \notin G_*$, then $\zeta_p(\mu_j)$ appears only in the left sum in (2.6) with the coefficient $\omega_j/2$. Therefore, the term $\zeta_p(\mu_j)$ corresponds to a single term $\zeta_p(\theta_i)$ in (2.7) with the coefficient $\widetilde{\omega}_i = \omega_j/2$. Similarly, $\zeta_p(\overline{\mu}_j)$ appears only in the right sum in (2.6) with the coefficient $\omega_j/2$, and it corresponds to a single term, say $\zeta_p(\theta_s)$, in (2.7) with the coefficient $\widetilde{\omega}_s = \omega_j/2$.

If $\mu_j \notin \mathbb{R}$ and $\overline{\mu}_j \in G_*$, then $\overline{\mu}_j = \mu_s$ for some index $s \neq j$, $1 \leq s \leq \ell$. Therefore, the term $\zeta_p(\mu_j)$ appears in both sums in (2.6), in the left sum with the coefficient $\omega_j/2$ and in the right sum with the coefficient $\omega_s/2$. Hence, $\zeta_p(\mu_j)$ corresponds to a single term $\zeta_p(\theta_i)$ in (2.7) with the coefficient $\widetilde{\omega}_i = \omega_j/2 + \omega_s/2$. Similarly, $\zeta_p(\overline{\mu}_j)$ corresponds to the term $\zeta_p(\overline{\theta}_i)$ in (2.7) with the coefficient equal to $\omega_j/2 + \omega_s/2$.

One can easily check that $\widetilde{\omega}_i > 0$, for $i = 1, \dots, m$, and that

$$\sum_{i=1}^{m} \widetilde{\omega}_i = 1.$$

Moreover, if $\theta_j = \overline{\theta}_i$ for $j \neq i$, then $\widetilde{\omega}_j = \widetilde{\omega}_i$.

Based on the relation (2.7) we set

$$v_* \equiv Q\xi, \quad \xi \equiv [\xi_1, \dots, \xi_n]^T \in \mathbb{R}^n,$$

where the ξ_j, $j = 1, \dots, n$, are defined as follows: if $\lambda_j \in G_*^{\rm sym}$, then there exits an index i such that $\lambda_j = \theta_i$, and we define $\xi_j \equiv \sqrt{\widetilde{\omega}_i}$. If $\lambda_j \notin G_*^{\rm sym}$, we set $\xi_j = 0$.

It remains to justify that the resulting vector v_* is real. If $\lambda_j \in \mathbb{R}$, then the corresponding eigenvector q_j (i.e., the jth column of the matrix Q) is real, and $\xi_j q_j$ is real. If $\lambda_j \notin \mathbb{R}$ and $\lambda_j \in G_*^{\rm sym}$, then also $\overline{\lambda}_j \in G_*^{\rm sym}$, and $\overline{\lambda}_j = \lambda_i$ for some $i \neq j$. The corresponding eigenvector is $q_i = \overline{q}_j$, and since $\xi_i = \xi_j$, the linear combination $\xi_j q_j + \xi_i q_i = \xi_j(q_j + \overline{q}_j)$ is a real vector. Therefore, the resulting vector $v_* = Q\xi$ is real.

Using (2.7), analogously to the previous section, we get

$$0 = v_*^T p(A)^T \left[f(A) - p_*(A)\right] v_*\,, \qquad \text{for all } p \in \mathcal{P}_k(\mathbb{R}),$$

or, equivalently,

$$\|f(A)v_* - p_*(A)v_*\| = \min_{p\in\mathcal{P}_k(\mathbb{R})} \|f(A)v_* - p(A)v_*\|$$

so that

$$\begin{aligned}\min_{p\in\mathcal{P}_k(\mathbb{R})} \|f(A) - p(A)\| = \delta &= \|f(A)v_* - p_*(A)v_*\| \\ &= \min_{p\in\mathcal{P}_k(\mathbb{R})} \|f(A)v_* - p(A)v_*\| \\ &\leq \max_{\substack{v\in\mathbb{R}^n \\ \|v\|=1}} \min_{p\in\mathcal{P}_k(\mathbb{R})} \|f(A)v - p(A)v\|.\end{aligned}$$

This is just the reverse of the inequality (1.3) for $\mathbb{F} = \mathbb{R}$, and hence the proof of Theorem 1.1 for $\mathbb{F} = \mathbb{R}$ is complete.

3. A different formulation.

Theorem 1.1 can be easily rewritten as a statement about pairwise commuting normal matrices. In the following we only discuss the complex case. The real case requires an analogous treatment as in Section 2.2.

Let $A_0, A_1, \dots, A_k \in \mathbb{C}^{n\times n}$ be pairwise commuting normal matrices. Then these matrices can be simultaneously unitarily diagonalized, i.e., there exists a unitary matrix $U \in \mathbb{C}^{n\times n}$ so that

$$U^H A_i U = \Lambda_i = \mathrm{diag}(\lambda_1^{(i)}, \dots, \lambda_n^{(i)}), \quad i = 0, 1, \dots, k;$$

see, e.g., [10, Theorem 2.5.5]. Let $\Gamma \equiv \{\lambda_1, \dots, \lambda_n\}$ be an arbitrary set containing n pairwise distinct complex numbers, and let $A \equiv U\mathrm{diag}(\lambda_1, \dots, \lambda_n)U^H \in \mathbb{C}^{n\times n}$. We now *define* the functions $f \in C(\Gamma)$ and $\varphi_1, \dots, \varphi_k \in C(\Gamma)$ to be any functions satisfying

$$f(\lambda_j) \equiv \lambda_j^{(0)}, \quad \varphi_i(\lambda_j) \equiv \lambda_j^{(i)}, \quad j = 1, \dots, n, \;\; i = 1, \dots, k.$$

Then $f(A) = A_0$ and $\varphi_i(A) = A_i$ for $i = 1, \dots, k$, so that Theorem 1.1 implies

$$\begin{aligned}\max_{\substack{v\in\mathbb{C}^n \\ \|v\|=1}} \min_{\alpha_1,\dots,\alpha_k\in\mathbb{C}} \|A_0 v - \sum_{i=1}^k \alpha_i A_i v\| &= \max_{\substack{v\in\mathbb{C}^n \\ \|v\|=1}} \min_{\alpha_1,\dots,\alpha_k\in\mathbb{C}} \|f(A)v - \sum_{i=1}^k \alpha_i \varphi_i(A) v\| \\ &= \min_{\alpha_1,\dots,\alpha_k\in\mathbb{C}} \|f(A) - \sum_{i=1}^k \alpha_i \varphi_i(A)\| \\ &= \min_{\alpha_1,\dots,\alpha_k\in\mathbb{C}} \|A_0 - \sum_{i=1}^k \alpha_i A_i\|.\end{aligned}$$

This equality is in fact the version of Theorem 1.1 proven by Greenbaum and Gurvits in [7, Theorem 2.3] for the case $\mathbb{F} = \mathbb{R}$.

Acknowledgments. We thank the anonymous referees for comments that helped to improve the presentation.

REFERENCES

[1] M. Bellalij, Y. Saad, and H. Sadok, *Analysis of some Krylov subspace methods for normal matrices via approximation theory and convex optimization*, Electron. Trans. Numer. Anal., 33 (2008/09), pp. 17–30. http://etna.mcs.kent.edu/vol.33.2008-2009/pp17-30.dir

[2] V. Faber, W. Joubert, E. Knill, and T. Manteuffel, *Minimal residual method stronger than polynomial preconditioning*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 707–729.

[3] V. Faber, J. Liesen, and P. Tichý, *On Chebyshev polynomials of matrices*, SIAM J. Matrix Anal. Appl., 31 (2009/10), pp. 2205–2221.

[4] ———, *Properties of worst-case GMRES*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 1500–1519.

[5] B. Fischer, *Polynomial Based Iteration Methods for Symmetric Linear Systems*, Wiley, Chichester, 1996.

[6] B. Fischer and F. Peherstorfer, *Chebyshev approximation via polynomial mappings and the convergence behaviour of Krylov subspace methods*, Electron. Trans. Numer. Anal., 12 (2001), pp. 205–215. http://etna.mcs.kent.edu/vol.12.2001/pp205-215.dir

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

[8] A. Greenbaum and L. N. Trefethen, *GMRES/CR and Arnoldi/Lanczos as matrix approximation problems*, SIAM J. Sci. Comput., 15 (1994), pp. 359–368.

[9] N. J. Higham, *Functions of Matrices. Theory and Computation*, SIAM, Philadelphia, 2008.

[10] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1990.

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

[12] J. Liesen and Z. Strakoš, *Krylov Subspace Methods. Principles and Analysis*, Oxford University Press, Oxford, 2013.

[13] G. G. Lorentz, *Approximation of Functions*, 2nd ed., Chelsea, New York, 1986.

[14] T. J. Rivlin and H. S. Shapiro, *A unified approach to certain problems of approximation and minimization*, J. Soc. Indust. Appl. Math., 9 (1961), pp. 670–699.

[15] Y. Saad and M. H. Schultz, *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

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

[17] K.-C. Toh and L. N. Trefethen, *The Chebyshev polynomials of a matrix*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 400–419.

Numer Algor (2013) 62:163–191
DOI 10.1007/s11075-012-9591-9

ORIGINAL PAPER

On computing quadrature-based bounds for the A-norm of the error in conjugate gradients

Gérard Meurant · Petr Tichý

Received: 12 December 2011 / Accepted: 9 May 2012 /
Published online: 25 May 2012

Abstract In their original paper, Golub and Meurant (BIT 37:687–705, 1997) suggest to compute bounds for the A-norm of the error in the conjugate gradient (CG) method using Gauss, Gauss-Radau and Gauss-Lobatto quadratures. The quadratures are computed using the $(1, 1)$-entry of the inverse of the corresponding Jacobi matrix (or its rank-one or rank-two modifications). The resulting algorithm called CGQL computes explicitly the entries of the Jacobi matrix and its modifications from the CG coefficients. In this paper, we use the fact that CG computes the Cholesky decomposition of the Jacobi matrix which is given implicitly. For Gauss-Radau and Gauss-Lobatto quadratures, instead of computing the entries of the modified Jacobi matrices, we directly compute the entries of the Cholesky decompositions of the (modified) Jacobi matrices. This leads to simpler formulas in comparison to those used in CGQL.

Keywords Conjugate Gradients · Norm of the error ·
Bounds for the error norm

Mathematics Subject Classification (2010) 15A06 · 65F10

This work was supported by the project No. IAA100300802 of the Grant Agency of the Academy of Sciences of the Czech Republic. It was finalized in 2011 during a visit of G. Meurant to the Institute of Computer Science of the Academy of Sciences of the Czech Republic.

G. Meurant
30 rue du sergent Bauchat, 75012 Paris, France
e-mail: gerard.meurant@gmail.com

P. Tichý ( )
Institute of Computer Science, Academy of Sciences of the Czech Republic,
Pod Vodárenskou věží 2, 18207 Prague, Czech Republic
e-mail: tichy@cs.cas.cz

1 Introduction

Today the Conjugate Gradient (CG) algorithm is the iterative method of choice for solving linear systems with a real positive definite symmetric matrix. It is almost always used with a preconditioner to speed up convergence. CG was introduced in the beginning of the 1950s by Magnus Hestenes and Eduard Stiefel [17]. It can be derived from several different perspectives, as an orthogonalization algorithm or as a minimization process. It can also be obtained from the Lanczos algorithm [19] that was published almost at the same time.

When using CG for solving a linear system $Ax = b$ an important question is when to stop the iterations. Ideally, one would like to stop the iterations when the norm of the error $\varepsilon_k = x - x_k$, where x_k are the CG iterates, is small enough. However, the error is unknown and most CG implementations rely on stopping criteria like $\|r_k\| \leq \epsilon \|b\|$ where $r_k = b - Ax_k$ is the residual vector, which is computed in CG or even $\|r_k\| \leq \epsilon \|r_0\|$. These types of stopping criteria can be misleading depending on the norm of A or the choice of the initial approximation. This was already noticed in the Hestenes and Stiefel paper [17, p. 410]. It can stop the iterations too early when the norm of the error is still too large, or too late in which case too many floating point operations have been done for obtaining the required accuracy. This motivated some researchers to look for ways to compute estimates of some norms of the error during CG iterations. The norm of the error which is particularly interesting for CG is the A-norm (also called the energy norm) which is minimized at each iteration. The A-norm of the error has an important meaning in physics and mechanics, and plays a fundamental role in evaluating convergence [1, 18]. It is defined as

$$\|\varepsilon_k\|_A \equiv \left(\varepsilon_k^T A \varepsilon_k\right)^{1/2}. \tag{1.1}$$

Inspired by the connection of CG with Riemann-Stieltjes integrals (already noticed in [17]), a way of research on this topic was started by Gene Golub in the 1970s and continued throughout the years with several collaborators [6–8, 11, 12, 14]. In particular, it was known that the A-norm of the error can be written as a Riemann-Stieltjes integral for an unknown stepwise constant measure depending on the eigenvalues of A. The main idea of Golub and his collaborators was to obtain bounds for this integral by using Gauss quadrature rules. It turns out that these bounds can be computed without the knowledge of the stepwise constant measure and at almost no cost during the CG iterations as we will see in the next sections.

In [11], these techniques were used for providing lower and upper bounds for quadratic forms $u^T f(A)u$ where f is a smooth function, A is a symmetric matrix and u is a given vector. The algorithm GQL (Gauss Quadrature and Lanczos) was based on the Lanczos algorithm and on computing functions of Jacobi matrices (and their rank-one or rank-two modifications). Later [12, 21], these techniques were adapted to the CG algorithm to compute lower and upper bounds on the A-norm of the error for which the function is $f(x) = 1/x$. The idea was to use CG instead of the Lanczos algorithm, to compute explicitly the entries of the corresponding Jacobi matrices and their modifications from the CG coefficients, and then to use the same formulas as in GQL. The formulas were summarized in the CGQL algorithm (QL standing again for Quadrature and Lanczos). Extensions to preconditioned CG were given in [22, 29]. This research is summarized in the books [13, 23]. The formula for the Gauss rule was analyzed for finite precision arithmetic in [28] where it is shown that it is still valid in finite precision up to small terms proportional to the unit roundoff.

The CGQL algorithm, whose most recent version is described in [13], may seem complicated, particularly for computing bounds with the Gauss-Radau or Gauss-Lobatto quadrature rules. It uses the tridiagonal Jacobi matrix obtained by translating the coefficients computed in CG into the Lanczos coefficients. Therefore the analysis of the formulas is difficult. Our aim in this paper is to show that these formulas can be considerably simplified by working with the LDL^T factorizations of the Jacobi matrices and their modifications instead of computing the Lanczos coefficients explicitly. In other words, one can obtain the bounds from the CG coefficients without computing the Lanczos coefficients. Therefore we hope that with the simpler new formulas the computation of upper bounds for the A-norm of the error can be incorporated more easily into existing CG codes.

It is fair to note that there exist other ways to compute estimates of the norms of the error; see [2, 3]. The paper [3] uses extrapolation techniques. However, this only gives estimates of the norm of the error and not bounds.

The outline of the paper is as follows. Section 2 recalls some basic facts about the Lanczos and CG algorithms, their connection to the approximation of the Riemann-Stieltjes integral using various quadrature rules, about computing quadratures using a convenient modification of the corresponding Jacobi matrix, and finally about estimating the A-norm of the error in CG. Section 3 describes the algebraic background for the new formulas; it is shown how to compute efficiently the entries of the LDL^T factorizations of modified Jacobi matrices. These results are then used in Section 4 in the formulation of the new algorithm called CGQ. Section 5 shows how to modify these new formulas when using preconditioning and finally, Section 6 presents numerical experiments which show that the new formulas are not only simpler but also slightly more accurate than the previous ones.

Throughout the paper e_k denotes the kth column of the identity matrix of appropriate order.

2 Conjugate Gradient, Lanczos and quadratures

In this section we briefly recall the Lanczos and Conjugate Gradient algorithms as well as their relationships; see, for instance, [15, 23].

Algorithm 1 Lanczos algorithm

```
**input** $A, v$
$\beta_0 = 0, v_0 = 0$
$v_1 = v/\|v\|$
**for** $k = 1, \ldots$ **do**
    $w = Av_k - \beta_{k-1}v_{k-1}$
    $\alpha_k = v_k^T w$
    $w = w - \alpha_k v_k$
    $\beta_k = \|w\|$
    $v_{k+1} = w/\beta_k$
**end for**
```

2.1 The Lanczos and CG algorithms

Given a starting vector v and a symmetric matrix $A \in \mathbb{R}^{N\times N}$, one can consider a sequence of nested subspaces

$$\mathcal{K}_k(A, v) \equiv \text{span}\left\{v, Av, \ldots, A^{k-1}v\right\}, \qquad k = 1, 2, \ldots,$$

called Krylov subspaces. The dimension of these subspaces is increasing up to an index n called *the grade of v with respect to A*, at which the maximal dimension is attained, and $\mathcal{K}_n(A, v)$ is invariant under multiplication with A. Assuming that $k < n$, the Lanczos algorithm (Algorithm 1) computes an orthonormal basis $v_1, \ldots, v_{k+1}$ of the Krylov subspace $\mathcal{K}_{k+1}(A, v)$. In Algorithm 1 we have used the modified Gram-Schmidt form of the algorithm. The basis vectors v_j satisfy the matrix relation

$$AV_k = V_k T_k + \beta_k v_{k+1} e_k^T$$

where $V_k = [v_1 \cdots v_k]$ and T_k is the $k \times k$ symmetric tridiagonal matrix of the recurrence coefficients computed in Algorithm 1:

$$T_k = \begin{bmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \ddots & \ddots & \\ & \ddots & \ddots & \beta_{k-1} \\ & & \beta_{k-1} & \alpha_k \end{bmatrix}.$$

The coefficients β_j being positive, T_k is a Jacobi matrix. The Lanczos algorithm works for any symmetric matrix, but if A is positive definite, then T_k is positive definite as well.

When solving a system of linear algebraic equations $Ax = b$ with symmetric and positive definite matrix A, the CG method (Algorithm 2) can be used. CG computes iterates x_k that are optimal since the A-norm of the error defined in (1.1) is minimized over the manifold $x_0 + \mathcal{K}_k(A, r_0)$,

$$\|x - x_k\|_A = \min_{y \in x_0 + \mathcal{K}_k(A, r_0)} \|x - y\|_A.$$

Algorithm 2 Conjugate gradient algorithm

```
input $A, b, x_0$
$r_0 = b - Ax_0$
$p_0 = r_0$
for $k = 1, \ldots, n$ until convergence do
    $\gamma_{k-1} = \frac{r_{k-1}^T r_{k-1}}{p_{k-1}^T A p_{k-1}}$
    $x_k = x_{k-1} + \gamma_{k-1} p_{k-1}$
    $r_k = r_{k-1} - \gamma_{k-1} A p_{k-1}$
    $\delta_k = \frac{r_k^T r_k}{r_{k-1}^T r_{k-1}}$
    $p_k = r_k + \delta_k p_{k-1}$
end for
```

The residual vectors $r_k = b - Ax_k$ are proportional to the Lanczos basis vectors v_j and hence mutually orthogonal,

$$v_{j+1} = (-1)^j \frac{r_j}{\|r_j\|}, \qquad j = 0, \ldots, k.$$

Therefore, the residual vectors r_j yield an orthogonal basis of the Krylov subspaces $\mathcal{K}_{k+1}(A, r_0)$. In this sense, CG can be seen as an algorithm for computing an orthogonal basis of the Krylov subspace $\mathcal{K}_{k+1}(A, r_0)$ and there is a close relationship between the CG and Lanczos algorithms. It is well-known (see, for instance [23]) that the recurrence coefficients computed in both algorithms are connected via

$$\beta_k = \frac{\sqrt{\delta_k}}{\gamma_{k-1}}, \quad \alpha_k = \frac{1}{\gamma_{k-1}} + \frac{\delta_{k-1}}{\gamma_{k-2}}. \quad \delta_0 = 0, \quad \gamma_{-1} = 1.$$

Writing these formulas in a matrix form, we get

$$T_k = L_k D_k L_k^T \tag{2.1}$$

where T_k is the Jacobi matrix resulting from the Lanczos algorithm and

$$L_k \equiv \begin{bmatrix} 1 & & & \\ \sqrt{\delta_1} & \ddots & & \\ & \ddots & \ddots & \\ & & \sqrt{\delta_{k-1}} & 1 \end{bmatrix}, \qquad D_k \equiv \begin{bmatrix} \gamma_0^{-1} & & & \\ & \ddots & & \\ & & \ddots & \\ & & & \gamma_{k-1}^{-1} \end{bmatrix}. \tag{2.2}$$

In other words, CG implicitly computes an LDL^T factorization of the Jacobi matrix T_k generated by the Lanczos algorithm. In this paper we are interested in computing bounds for the A-norm of the error. Noticing that the error ε_k and the residual r_k are related through $A\varepsilon_k = r_k$, we have

$$\|\varepsilon_k\|_A^2 = \varepsilon_k^T A \varepsilon_k = r_k^T A^{-1} r_k.$$

The quantity on the right-hand side is a quadratic form. In the next subsection we briefly recall how quadratic forms are related to Riemann-Stieltjes integrals. This will allow us to compute bounds for the norm of the error.

2.2 Connection with Riemann-Stieltjes integrals

Let

$$A = U\Lambda U^T, \qquad UU^T = U^TU = I, \tag{2.3}$$

be the eigendecomposition of the symmetric matrix A where $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_N)$ and $U = [u_1, \dots, u_N]$. For simplicity of notation we assume that all the eigenvalues of A are distinct and ordered as $\lambda_1 < \lambda_2 < \cdots < \lambda_N$ (the generalization of the below defined function $\omega(\lambda)$ to the case of multiple eigenvalues is straightforward). Let v_1 be a given unit vector. Define the weights ω_i by

$$\omega_i \equiv (v_1, u_i)^2 \qquad \text{so that} \qquad \sum_{i=1}^{N} \omega_i = 1\,, \tag{2.4}$$

and the (nondecreasing) distribution function $\omega(\lambda)$ with a finite number of points of increase $\lambda_1, \lambda_2, \dots, \lambda_N$,

$$\omega(\lambda) \equiv \begin{cases} 0 \text{ for } \lambda < \lambda_1\,, \\ \sum_{j=1}^{i} \omega_j \text{ for } \lambda_i \le \lambda < \lambda_{i+1}\,, \quad 1 \le i \le N-1\,, \\ 1 \text{ for } \lambda_N \le \lambda\,. \end{cases} \tag{2.5}$$

Having the distribution function $\omega(\lambda)$ and an interval $\langle \zeta, \xi \rangle$ such that $\zeta < \lambda_1 < \lambda_2 < \cdots < \lambda_N < \xi$, for any continuous function f, one can define the Riemann-Stieltjes integral (see, for instance [13])

$$\int_{\zeta}^{\xi} f(\lambda)\, d\omega(\lambda). \tag{2.6}$$

Since $\omega(\lambda)$ is a stepwise constant function and all points of increase lie in the open interval (ζ, ξ), the integral (2.6) is a finite sum and it holds that

$$\int_{\zeta}^{\xi} f(\lambda)\, d\omega(\lambda) = \sum_{i=1}^{N} \omega_i f(\lambda_i) \;=\; v_1^T f(A) v_1. \tag{2.7}$$

The quantity $v_1^T f(A) v_1$ can be expressed using the tridiagonal matrix T_n stemming from the Lanczos algorithm (note that n is the grade of v_1 with respect to A). In the nth step of the Lanczos algorithm we get the full orthonormal basis of $\mathcal{K}_n(A, v_1)$ and we have

$$AV_n = V_nT_n \qquad \Rightarrow \qquad f(A)V_n = V_n f(T_n)$$

and, therefore,

$$v_1^T f(A) v_1 = v_1^T f(A) V_n e_1 = v_1^T V_n f(T_n) e_1 = e_1^T f(T_n) e_1.$$

From this it is clear that the quadratic form we are interested in, $r_k^T A^{-1} r_k$, can be written as a Riemann-Stieltjes integral for the function $f(\lambda) = 1/\lambda$.

2.3 Quadrature formulas

The integral (2.7), i.e. the quantity $v_1^T f(A) v_1$, can be approximated by quadrature formulas, for example the Gauss, Gauss-Radau and Gauss-Lobatto rules; see, for instance, [9, 13]. The general quadrature formula we use has the form

$$\int_\zeta^\xi f\, d\omega(\lambda) = \sum_{i=1}^{k} w_i f(\nu_i) + \sum_{j=1}^{m} \widetilde{w}_j f(\widetilde{\nu}_j) + \mathcal{R}_k[f],$$

where the weights $[w_i]_{i=1}^k$, $[\widetilde{w}_j]_{j=1}^m$ and the nodes $[\nu_i]_{i=1}^k$ are *unknowns* and the nodes $[\widetilde{\nu}_j]_{j=1}^m$ are *prescribed* outside the open integration interval. In our case it is sufficient when the prescribed nodes are strictly smaller than λ_1 or strictly larger than λ_N. The unknown nodes and weights are chosen to maximize the degree of exactness of the quadrature rule. If $m = 0$, there are no prescribed nodes, and we obtain the Gauss rule. If $m = 1$ we have the Gauss-Radau rule and if $m = 2$, this is the Gauss-Lobatto rule. It is known (see, for instance, [27]) that if $f \in C^{2k+m}$, then the remainder is

$$\mathcal{R}_k[f] = \frac{f^{(2k+m)}(\upsilon)}{(2k+m)!} \int_\zeta^\xi \prod_{j=1}^{m}(\lambda - \widetilde{\nu}_j) \left[\prod_{i=1}^{k}(\lambda - \nu_i)\right]^2 d\omega(\lambda)\,, \qquad \upsilon \in (\zeta, \xi).$$

For some functions f of interest the sign of the remainder term is known.

Consider first the Gauss rule, i.e. $m = 0$. The nodes ν_i and the weights w_i of the kth Gauss quadrature approximation are implicitly determined by the Lanczos algorithm; the nodes are the eigenvalues of T_k generated by the Lanczos algorithm started from v_1 and the weights are the squares of the first components of the normalized eigenvectors of T_k; see [16, 30].

To obtain the Gauss-Radau and Gauss-Lobatto rules, we must extend the matrix T_k in such a way that it has the prescribed nodes as eigenvalues; see [10]. Suppose that μ is a prescribed node. For the Gauss-Radau quadrature rule, we have to determine the coefficient $\tilde{\alpha}_{k+1}^{(\mu)}$ so that μ is an eigenvalue of the extended matrix

$$\tilde{T}_{k+1}^{(\mu)} = \begin{bmatrix} \alpha_1 & \beta_1 & & & \\ \beta_1 & \ddots & \ddots & & \\ & \ddots & \ddots & \beta_{k-1} & \\ & & \beta_{k-1} & \alpha_k & \beta_k \\ & & & \beta_k & \tilde{\alpha}_{k+1}^{(\mu)} \end{bmatrix}. \tag{2.8}$$

Given two prescribed nodes μ and η, for the Gauss-Lobatto quadrature rule we have to find the coefficients $\tilde{\alpha}_{k+1}^{(\mu,\eta)}$ and $\tilde{\beta}_{k}^{(\mu,\eta)}$ such that μ and η are eigenvalues of the extended matrix

$$\tilde{T}_{k+1}^{(\mu,\eta)} = \begin{bmatrix} \alpha_1 & \beta_1 & & & \\ \beta_1 & \ddots & \ddots & & \\ & \ddots & \ddots & \beta_{k-1} & \\ & & \beta_{k-1} & \alpha_k & \tilde{\beta}_k^{(\mu,\eta)} \\ & & & \tilde{\beta}_k^{(\mu,\eta)} & \tilde{\alpha}_{k+1}^{(\mu,\eta)} \end{bmatrix}. \tag{2.9}$$

Having the matrices T_k, $\tilde{T}_{k+1}^{(\mu)}$ and $\tilde{T}_{k+1}^{(\mu,\eta)}$, the Gauss, Gauss-Radau and Gauss-Lobatto quadrature rules can be respectively written in the form (see [13])

$$\begin{aligned} e_1^T f(T_n) e_1 &= e_1^T f(T_k) e_1 + \mathcal{R}_k^{(G)}[f]\,, \\ e_1^T f(T_n) e_1 &= e_1^T f\left(\tilde{T}_{k+1}^{(\mu)}\right) e_1 + \mathcal{R}_k^{(R)}[f]\,, \\ e_1^T f(T_n) e_1 &= e_1^T f\left(\tilde{T}_{k+1}^{(\mu,\eta)}\right) e_1 + \mathcal{R}_k^{(L)}[f]\,. \end{aligned}$$

These rules can provide lower and upper bounds on the integral (2.7), based on the following implications (see, e.g., [13, Theorem 6.4 and 6.5]):

$$\text{If } f^{(2k)}(\lambda) > 0 \text{ for all } \lambda \in \langle \zeta, \xi \rangle, \text{ then } \mathcal{R}_k^{(G)}[f] > 0. \tag{2.10}$$

$$\text{If } f^{(2k+1)}(\lambda) < 0 \text{ for all } \lambda \in \langle \zeta, \xi \rangle, \text{ and } \mu \le \lambda_1, \text{ then } \mathcal{R}_k^{(R)}[f] < 0. \tag{2.11}$$

$$\text{If } f^{(2k+2)}(\lambda) > 0 \text{ for all } \lambda \in \langle \zeta, \xi \rangle, \text{ and } \mu \le \lambda_1 \text{ and } \lambda_N \le \eta, \text{ then } \mathcal{R}_k^{(L)}[f] < 0. \tag{2.12}$$

If the derivatives of the function f satisfy the assumptions in (2.10)–(2.12), then the Gauss rule gives a lower bound and the Gauss-Radau and Gauss-Lobatto rules give upper bounds for the integral (2.7). Note that the assumptions on the sign of the derivatives of f in (2.10)–(2.12) are satisfied for the function $f(\lambda) = 1/\lambda$.

2.4 CG and Gauss quadrature

For the quadratic form involved in CG we are interested in the function $f(\lambda) = 1/\lambda$. Previous results imply that we can express the Gauss quadrature rule using the Lanczos-related quantities as

$$\left(T_n^{-1}\right)_{1,1} = \left(T_k^{-1}\right)_{1,1} + \mathcal{R}_k^{(G)}[\lambda^{-1}].$$

In [28] the authors show that the same equation multiplied by $\|r_0\|^2$ can be written using the CG-related quantities

$$\|x - x_0\|_A^2 = \sum_{j=0}^{k-1} \gamma_j \|r_j\|^2 + \|x - x_k\|_A^2 \,.$$

In other words, CG can be see as a procedure that implicitly determines weights and nodes of the Gauss quadrature rule applied to the Riemann-Stieltjes integral

$$\int_\zeta^\xi \lambda^{-1}\, d\omega(\lambda) = \frac{\|x - x_0\|_A^2}{\|r_0\|^2}$$

for which the Gauss quadrature approximation is given by

$$\left(T_k^{-1}\right)_{1,1} = \frac{1}{\|r_0\|^2} \sum_{j=0}^{k-1} \gamma_j \|r_j\|^2. \tag{2.13}$$

The remainder is nothing but the scaled and squared A-norm of the kth error,

$$\mathcal{R}_k^{(G)}\left[\lambda^{-1}\right] = \frac{\|x - x_k\|_A^2}{\|r_0\|^2}.$$

For more information on this topic see, e.g., [14, 28, Section 3] or [24, Subsection 3.3].

2.5 Estimating the A-norm of the error in CG

Of course, at CG iteration k we do not know $(T_n^{-1})_{1,1}$ or $\|x - x_0\|_A$. For estimating the A-norm of the error in CG we consider the Gauss quadrature rule at step k,

$$\|x - x_0\|_A^2 = \|r_0\|^2 \left(T_k^{-1}\right)_{1,1} + \|x - x_k\|_A^2, \tag{2.14}$$

and a (eventually modified) quadrature rule at step $k + d$, $d > 0$,

$$\|x - x_0\|_A^2 = \|r_0\|^2 \left(\hat{T}_{k+d}^{-1}\right)_{1,1} + \hat{\mathcal{R}}_{k+d}\left[\lambda^{-1}\right], \tag{2.15}$$

where $\hat{T}_{k+d}$ stands for the matrix T_{k+d} (in the case of using Gauss rule) or a suitable modification of T_{k+d} (in the case of using Gauss-Radau or Gauss-Lobatto rules). From (2.14) and (2.15) we get

$$\|x - x_k\|_A^2 = \hat{Q}_{k,d} + \hat{\mathcal{R}}_{k+d}\left[\lambda^{-1}\right], \; \hat{Q}_{k,d} \equiv \|r_0\|^2 \left(\left(\hat{T}_{k+d}^{-1}\right)_{1,1} - \left(T_k^{-1}\right)_{1,1}\right). \tag{2.16}$$

$\hat{Q}_{k,d}$ represents either a lower bound on $\|x - x_k\|_A^2$ if $\hat{\mathcal{R}}_{k+d}\left[\lambda^{-1}\right] > 0$, or an upper bound in the case $\hat{\mathcal{R}}_{k+d}\left[\lambda^{-1}\right] < 0$. It means that at CG iteration $k + d$, by computing $\hat{Q}_{k,d}$, we can obtain a bound for the A-norm of the error at iteration k.

From the computational point of view, as noted in [14] and [21], it is not convenient to compute $\hat{Q}_{k,d}$ by first computing $(T_k^{-1})_{1,1}$, $\left(\hat{T}_{k+d}^{-1}\right)_{1,1}$, and then taking the difference. By subtracting both quantities we loose accuracy and, as a result, the use of the estimate is limited by the square root of machine precision. Instead of subtracting, it is better to use the following identity

$$\left(\hat{T}_{k+d}^{-1}\right)_{1,1} - (T_k^{-1})_{1,1} = \left(\hat{T}_{k+d}^{-1}\right)_{1,1} - (T_{k+d-1}^{-1})_{1,1} + \sum_{j=k}^{k+d-2} \left[\left(T_{j+1}^{-1}\right)_{1,1} - \left(T_j^{-1}\right)_{1,1}\right].$$

From (2.13) we have

$$\|r_0\|^2 \left[\left(T_{j+1}^{-1}\right)_{1,1} - \left(T_j^{-1}\right)_{1,1}\right] = \gamma_j \|r_j\|^2 \tag{2.17}$$

so that $\hat{Q}_{k,d}$ takes the form

$$\hat{Q}_{k,d} = \|r_0\|^2 \left[\left(\hat{T}_{k+d}^{-1}\right)_{1,1} - (T_{k+d-1}^{-1})_{1,1}\right] + \sum_{j=k}^{k+d-2} \gamma_j \|r_j\|^2.$$

Therefore, the problem of computing $\hat{Q}_{k,d}$ reduces to the problem of computing efficiently the difference

$$\|r_0\|^2 \left[\left(\hat{T}_{j+1}^{-1}\right)_{1,1} - \left(T_j^{-1}\right)_{1,1}\right], \qquad j = k + d - 1, \tag{2.18}$$

using the CG-related quantities that are available during the CG iterations. This is easy for the Gauss rule since it is given by (2.17) but, for the Gauss-Radau and Gauss-Lobatto rules we need to use results about factorizations of tridiagonal matrices. They are recalled in the next section.

3 Factorizations of tridiagonal matrices

In this section our aim is to show how to compute the quantities we need for the quadrature rules by relying only on the LDL^T factorizations of the tridiagonal matrices. For doing this we will use variants of the qd algorithm; see [25]. Although the matrices in our problem are positive definite in theory, it can happen during finite precision computations that the computed tridiagonal matrices are indefinite. Therefore, we will assume that our symmetric tridiagonal matrices are indefinite. However, we also assume that their LDL^T factorizations exist.

3.1 LDL^T factorization of T_k and of its extension $\hat{T}_{k+1}$

Consider a symmetric tridiagonal matrix T_k with diagonal entries α_j and subdiagonal entries $\beta_j \neq 0$,

$$T_k = \begin{bmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \ddots & \ddots & \\ & \ddots & \ddots & \beta_{k-1} \\ & & \beta_{k-1} & \alpha_k \end{bmatrix} \tag{3.1}$$

and its LDL^T factorization, $T_k = L_k D_k L_k^T$, denoted as

$$T_k = \begin{bmatrix} 1 & & & \\ \ell_1 & \ddots & & \\ & \ddots & \ddots & \\ & & \ell_{k-1} & 1 \end{bmatrix} \begin{bmatrix} d_1 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & d_k \end{bmatrix} \begin{bmatrix} 1 & \ell_1 & & \\ & \ddots & \ddots & \\ & & \ddots & \ell_{k-1} \\ & & & 1 \end{bmatrix}. \tag{3.2}$$

To compute this factorization, one can use the following recurrence relations, see, e.g., [13, p.25],

$$d_1 = \alpha_1, \quad \ell_j = \frac{\beta_j}{d_j}, \quad d_{j+1} = \alpha_{j+1} - \beta_j \ell_j, \qquad j = 1, \ldots, k-1. \tag{3.3}$$

If T_k is extended by one row and one column to the matrix $\hat{T}_{k+1}$,

$$\hat{T}_{k+1} = \begin{bmatrix} T_k & \hat{\beta}_k e_k \\ \hat{\beta}_k e_k^T & \hat{\alpha}_{k+1} \end{bmatrix}, \tag{3.4}$$

the LDL^T factorization of $\hat{T}_{k+1}$ is just a straightforward extension of the LDL^T factorization of T_k,

$$\hat{T}_{k+1} = \begin{bmatrix} 1 & & & & \\ \ell_1 & \ddots & & & \\ & \ddots & \ddots & & \\ & & \ell_{k-1} & 1 & \\ & & & \hat{\ell}_k & 1 \end{bmatrix} \begin{bmatrix} d_1 & & & & \\ & \ddots & & & \\ & & \ddots & & \\ & & & d_k & \\ & & & & \hat{d}_{k+1} \end{bmatrix} \begin{bmatrix} 1 & \ell_1 & & & \\ & \ddots & & & \\ & & \ddots & \ell_{k-1} & \\ & & & 1 & \hat{\ell}_k \\ & & & & 1 \end{bmatrix}$$

where the additional entries are given by

$$\hat{\ell}_k = \frac{\hat{\beta}_k}{d_k}, \qquad \hat{d}_{k+1} = \hat{\alpha}_{k+1} - \hat{\beta}_k \hat{\ell}_k = \hat{\alpha}_{k+1} - \frac{\hat{\beta}_k^2}{d_k}.$$

3.2 The difference between (1, 1) entries of $\hat{T}_{k+1}^{-1}$ and T_k^{-1}

To compute various types of quadratures, we need to compute efficiently the difference between (1, 1) entries of inverses of some tridiagonal matrices; see (2.18). This can be done using the following formula, see Theorem 3.9 in [13, p. 31],

$$\left(\hat{T}_{k+1}^{-1}\right)_{1,1} - \left(T_k^{-1}\right)_{1,1} = \hat{d}_{k+1}^{-1} \frac{\left(\beta_1 \dots \beta_{k-1} \hat{\beta}_k\right)^2}{(d_1 \dots d_k)^2} = \frac{\hat{\ell}_k^2}{\hat{d}_{k+1}} \left(\ell_1 \dots \ell_{k-1}\right)^2, \tag{3.5}$$

where we have used that $\beta_j = \ell_j d_j$, $j = 1, \dots, k-1$ and $\hat{\beta}_k = \hat{\ell}_k \hat{d}_k$. In other words, having the LDL^T factorizations of T_k and $\hat{T}_{k+1}$, one can compute the required difference without subtraction.

3.3 LDL^T factorization of a shifted tridiagonal matrix

We will see that for prescribing some eigenvalues we have to deal with shifted tridiagonal matrices. Let the shift μ be given such that it is different from any eigenvalue of T_k so that $T_k - \mu I$ is nonsingular. In the application μ will be smaller (resp. larger) than the smallest (resp. largest) eigenvalue of A. We denote the LDL^T factorization of $T_k - \mu I$ (when it exists) as,

$$T_k - \mu I = \bar{L}_k^{(\mu)} \bar{D}_k^{(\mu)} \left(\bar{L}_k^{(\mu)}\right)^T. \tag{3.6}$$

The entries of the LDL^T factorization of $T_k - \mu I$ are denoted with a bar, the dependence on the parameter μ is denoted by the superscript within parentheses. This factorization can be computed from scratch using (3.3) since $T_k - \mu I$ differs from T_k only in diagonal entries by

$$\bar{d}_1^{(\mu)} = \alpha_1 - \mu, \quad \bar{\ell}_j^{(\mu)} = \frac{\beta_j}{\bar{d}_j^{(\mu)}}, \quad \bar{d}_{j+1}^{(\mu)} = \alpha_{j+1} - \mu - \beta_j \bar{\ell}_j^{(\mu)}, \quad j = 1, \dots, k-1.$$

However, suppose now that T_k is given in the form of its LDL^T factorization, i.e. we know the entries $d_1, \dots, d_k$ and $\ell_1, \dots, \ell_{k-1}$ and want to compute the factorization (3.6) directly from these entries. This can be done using the

Algorithm 3 `stqds`

```
input $\mu, d_1, \dots, d_k, \ell_1, \dots, \ell_{k-1}$
$\bar{d}_1^{(\mu)} = d_1 - \mu$
for $j = 1, \dots, k-1$ do
    $\bar{\ell}_j^{(\mu)} = \frac{d_j \ell_j}{\bar{d}_j^{(\mu)}}$
    $\bar{d}_{j+1}^{(\mu)} = (d_{j+1} - \mu) + d_j \ell_j^2 - d_j \ell_j \bar{\ell}_j^{(\mu)}$
end for
output $\bar{d}_1^{(\mu)}, \dots, \bar{d}_k^{(\mu)}, \bar{\ell}_1^{(\mu)}, \dots, \bar{\ell}_{k-1}^{(\mu)}$
```

Algorithm 4 dstqds

```
input $\mu, d_1, \ldots, d_k, \ell_1, \ldots, \ell_{k-1}$
$s_1^{(\mu)} = \mu$
for $j = 1, \ldots, k-1$ do
    $\bar{d}_j^{(\mu)} = d_j - s_j^{(\mu)}$
    $\bar{\ell}_j^{(\mu)} = \frac{d_j \ell_j}{\bar{d}_j^{(\mu)}}$
    $s_{j+1}^{(\mu)} = \mu + \ell_j \bar{\ell}_j^{(\mu)} s_j^{(\mu)}$
end for
$\bar{d}_k^{(\mu)} = d_k - s_k^{(\mu)}$
output $\bar{d}_1^{(\mu)}, \ldots, \bar{d}_k^{(\mu)}, \bar{\ell}_1^{(\mu)}, \ldots, \bar{\ell}_{k-1}^{(\mu)}$
```

stqds algorithm (Algorithm 3) which is a variant of the Rutishauser qd algorithm, see [25, 26] or [13, p. 35]. This can be further improved by introducing the difference $s_j^{(\mu)} \equiv d_j - \bar{d}_j^{(\mu)}$. It yields another version of the stqds algorithm called dstqds (Algorithm 4) which avoids some subtractions.

3.4 Rank-one modification of T_{k+1} with a prescribed eigenvalue

Given a real number μ different from any eigenvalue of T_k, our aim in this subsection is to modify the $(k+1, k+1)$st entry of T_{k+1} so that the resulting matrix $\tilde{T}_{k+1}^{(\mu)}$ defined in (2.8) has μ as a prescribed eigenvalue. In [10] it has been shown that

$$\tilde{\alpha}_{k+1}^{(\mu)} = \mu + \xi_k^{(\mu)}$$

where $\xi_k^{(\mu)}$ is the last component of the solution of the tridiagonal system

$$\begin{bmatrix} \alpha_1 - \mu & \beta_1 & & \\ \beta_1 & \ddots & \ddots & \\ & \ddots & \ddots & \beta_{k-1} \\ & & \beta_{k-1} & \alpha_k - \mu \end{bmatrix} \begin{bmatrix} \xi_1^{(\mu)} \\ \vdots \\ \xi_{k-1}^{(\mu)} \\ \xi_k^{(\mu)} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \beta_k^2 \end{bmatrix}, \tag{3.7}$$

see also [13, pp. 88–89]. Considering the LDL^T factorization (3.6) of $T_k - \mu I$, it is easy to show that

$$\xi_k^{(\mu)} = \frac{\beta_k^2}{\bar{d}_k^{(\mu)}}, \qquad \text{i.e.} \qquad \tilde{\alpha}_{k+1}^{(\mu)} = \mu + \frac{\beta_k^2}{\bar{d}_k^{(\mu)}}.$$

Suppose now that the matrix T_{k+1} is given in the form of the LDL^T factorization (T_{k+1} is not given explicitly). We would like to modify this factorization in such a way that we obtain the LDL^T factorization of $\tilde{T}_{k+1}^{(\mu)}$.

First we observe that if $T_{k+1} = L_{k+1} D_{k+1} L_{k+1}^T$, then the LDL^T factorization of $\tilde{T}_{k+1}^{(\mu)}$ is given by

$$\tilde{T}_{k+1}^{(\mu)} = L_{k+1} \begin{bmatrix} d_1 & & & & \\ & \ddots & & & \\ & & \ddots & & \\ & & & d_k & \\ & & & & \tilde{d}_{k+1}^{(\mu)} \end{bmatrix} L_{k+1}^T, \tag{3.8}$$

(see (3.5)) where

$$\tilde{d}_{k+1}^{(\mu)} = \tilde{\alpha}_{k+1}^{(\mu)} - \beta_k \ell_k = \mu + \xi_k^{(\mu)} - d_k \ell_k^2 = \mu + \frac{\beta_k^2}{\bar{d}_k^{(\mu)}} - d_k \ell_k^2. \tag{3.9}$$

In the following lemma we show that $\tilde{d}_{k+1}^{(\mu)} = d_{k+1} - \bar{d}_{k+1}^{(\mu)}$ where $\bar{d}_{k+1}^{(\mu)}$ is the last diagonal entry of the factorization of $T_{k+1} - \mu I$.

Lemma 3.1 *Given μ different from any eigenvalue of T_k, consider the LDL^T factorizations of T_{k+1} and $T_{k+1} - \mu I$,*

$$T_{k+1} = L_{k+1} D_{k+1} L_{k+1}^T, \qquad T_{k+1} - \mu I = \bar{L}_{k+1}^{(\mu)} \bar{D}_{k+1}^{(\mu)} \left(\bar{L}_{k+1}^{(\mu)}\right)^T.$$

Let $\tilde{T}_{k+1}^{(\mu)}$ be the rank-one modification (2.8) of T_{k+1} such that μ is an eigenvalue of $\tilde{T}_{k+1}^{(\mu)}$ and consider its LDL^T factorization (3.8). Then it holds that

$$\tilde{d}_{k+1}^{(\mu)} = d_{k+1} - \bar{d}_{k+1}^{(\mu)}.$$

Proof By a simple algebraic manipulation we obtain

$$d_{k+1} - \bar{d}_{k+1}^{(\mu)} = d_{k+1} - \left((d_{k+1} - \mu) + d_k \ell_k^2 - d_k \ell_k \bar{\ell}_k^{(\mu)}\right) = \left(\mu - d_k \ell_k^2\right) + d_k \ell_k \bar{\ell}_k^{(\mu)}.$$

From (3.9) it follows that $\mu - d_k \ell_k^2 = \tilde{d}_{k+1}^{(\mu)} - \beta_k \frac{\beta_k}{\bar{d}_k^{(\mu)}}$, therefore

$$d_{k+1} - \bar{d}_{k+1}^{(\mu)} = \tilde{d}_{k+1}^{(\mu)} - \beta_k \frac{\beta_k}{\bar{d}_k^{(\mu)}} + d_k \ell_k \bar{\ell}_k^{(\mu)} = \tilde{d}_{k+1}^{(\mu)} - \ell_k d_k \bar{\ell}_k^{(\mu)} + d_k \ell_k \bar{\ell}_k^{(\mu)} \;=\; \tilde{d}_{k+1}^{(\mu)}$$

which proves the result. □

The formula for $\tilde{d}_{k+1}^{(\mu)}$ requires not only the (known) entry d_{k+1}, but also the (so far unknown) entry $\bar{d}_{k+1}^{(\mu)}$ from the LDL^T factorization of $T_{k+1} - \mu I$. In the following we show how to recursively compute the entry $\tilde{d}_{k+1}^{(\mu)}$ so that the LDL^T factorization of $T_{k+1} - \mu I$ need not to be computed.

Lemma 3.2 *With the notation above,*

$$\tilde{d}_1^{(\mu)} \equiv \mu, \qquad \tilde{d}_{k+1}^{(\mu)} = \mu + \ell_k^2 \frac{d_k \tilde{d}_k^{(\mu)}}{d_k - \tilde{d}_k^{(\mu)}} \quad \textit{for} \quad k \geq 1. \tag{3.10}$$

Proof From Lemma 3.1 it follows that $\tilde{d}_{k+1}^{(\mu)}$ is nothing but the difference $s_{k+1}^{(\mu)} = d_{k+1} - \bar{d}_{k+1}^{(\mu)}$ introduced in Algorithm 4. Using

$$s_{k+1}^{(\mu)} - \mu = \ell_k \bar{\ell}_k^{(\mu)} s_k^{(\mu)} = \ell_k \frac{d_k \ell_k}{\bar{d}_k^{(\mu)}} s_k^{(\mu)} = \ell_k \frac{d_k \ell_k}{d_k - s_k^{(\mu)}} s_k^{(\mu)}$$

and $\tilde{d}_k^{(\mu)} = s_k^{(\mu)}$ we obtain the formula (3.10). □

Formula (3.10) will be used to compute the inverse of $\tilde{d}_{k+1}^{(\mu)}$.

3.5 Rank-two modification of T_{k+1} with two prescribed eigenvalues

Given two numbers μ and η different from any eigenvalue of T_k, we would like to find a rank-two modification of the matrix T_{k+1} so that the matrix $\tilde{T}_{k+1}^{(\mu,\eta)}$ defined in (2.9) has μ and η as prescribed eigenvalues. In [10] it has been shown that $\tilde{\alpha}_{k+1}^{(\mu,\eta)}$ and $\tilde{\beta}_k^{(\mu,\eta)}$ satisfy

$$\tilde{\alpha}_{k+1}^{(\mu,\eta)} = \mu + \left(\tilde{\beta}_k^{(\mu,\eta)}\right)^2 \zeta_k^{(\mu)}, \qquad \tilde{\alpha}_{k+1}^{(\mu,\eta)} = \eta + \left(\tilde{\beta}_k^{(\mu,\eta)}\right)^2 \zeta_k^{(\eta)},$$

where $\zeta_k^{(\mu)}$, respectively $\zeta_k^{(\eta)}$, is the last component of the tridiagonal system

$$(T_k - \mu I)\zeta^{(\mu)} = e_k, \qquad \zeta^{(\mu)} \equiv \left[\zeta_1^{(\mu)}, \ldots, \zeta_k^{(\mu)}\right]^T,$$

respectively,

$$(T_k - \eta I)\zeta^{(\eta)} = e_k, \qquad \zeta^{(\eta)} \equiv \left[\zeta_1^{(\eta)}, \ldots, \zeta_k^{(\eta)}\right]^T.$$

Summarizing, we first solve systems $(T_k - \mu I)\zeta^{(\mu)} = e_k$, $(T_k - \eta I)\zeta^{(\eta)} = e_k$ and then we obtain $\tilde{\alpha}_{k+1}^{(\mu,\eta)}$ and $\left(\tilde{\beta}_k^{(\mu,\eta)}\right)^2$ as the solution of the 2×2 linear system

$$\begin{bmatrix} 1 & -\xi_k^{(\mu)} \\ 1 & -\xi_k^{(\eta)} \end{bmatrix} \begin{bmatrix} \tilde{\alpha}_{k+1} \\ \left(\tilde{\beta}_k^{(\mu,\eta)}\right)^2 \end{bmatrix} = \begin{bmatrix} \mu \\ \eta \end{bmatrix}.$$

We are now interested in the LDL^T factorization of the matrix $\tilde{T}_{k+1}^{(\mu,\eta)}$, see (2.9),

$$\begin{bmatrix} 1 & & & & \\ \ell_1 & \ddots & & & \\ & \ddots & \ddots & & \\ & & \ell_{k-1} & 1 & \\ & & & \tilde{\ell}_k^{(\mu,\eta)} & 1 \end{bmatrix} \begin{bmatrix} d_1 & & & \\ & \ddots & & \\ & & d_k & \\ & & & \tilde{d}_{k+1}^{(\mu,\eta)} \end{bmatrix} \begin{bmatrix} 1 & \ell_1 & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ell_{k-1} & \\ & & & 1 & \tilde{\ell}_k^{(\mu,\eta)} \\ & & & & 1 \end{bmatrix}. \tag{3.11}$$

In the following lemma we express $\tilde{\ell}_k^{(\mu,\eta)}$ and $\tilde{d}_{k+1}^{(\mu,\eta)}$ using the entries of the LDL^T factorizations of $T_k - \mu I$ and $T_k - \eta I$.

Lemma 3.3 *The entries $\tilde{\ell}_k^{(\mu,\eta)}$ and $\tilde{d}_{k+1}^{(\mu,\eta)}$ from (3.11) can be computed using the following formulas,*

$$\left(\tilde{\ell}_k^{(\mu,\eta)}\right)^2 = \frac{\bar{d}_k^{(\mu)}\bar{d}_k^{(\eta)}}{d_k^2}\frac{\eta-\mu}{\bar{d}_k^{(\eta)}-\bar{d}_k^{(\mu)}}, \quad \tilde{d}_{k+1}^{(\mu,\eta)} = \frac{\eta\bar{d}_k^{(\eta)}-\mu\bar{d}_k^{(\mu)}}{\bar{d}_k^{(\eta)}-\bar{d}_k^{(\mu)}} - d_k\left(\tilde{\ell}_k^{(\mu,\eta)}\right)^2. \tag{3.12}$$

Proof For simplicity of notation in this proof, we will omit the (μ, η) upper indices. From (3.11), it follows that $\tilde{\ell}_k$ and $\tilde{d}_{k+1}$ satisfy

$$\tilde{\ell}_k^2 = \frac{\tilde{\beta}_k^2}{d_k^2}, \qquad \tilde{d}_{k+1} = \tilde{\alpha}_{k+1} - \tilde{\beta}_k\tilde{\ell}_k = \tilde{\alpha}_{k+1} - \frac{\tilde{\beta}_k^2}{d_k}.$$

Therefore,

$$\begin{bmatrix}\tilde{d}_{k+1}\\ \tilde{\ell}_k^2\end{bmatrix} = \begin{bmatrix}1 & -d_k^{-1}\\ 0 & d_k^{-2}\end{bmatrix}\begin{bmatrix}\tilde{\alpha}_{k+1}\\ \tilde{\beta}_k^2\end{bmatrix}, \qquad \begin{bmatrix}\tilde{\alpha}_{k+1}\\ \tilde{\beta}_k^2\end{bmatrix} = \begin{bmatrix}1 & d_k\\ 0 & d_k^2\end{bmatrix}\begin{bmatrix}\tilde{d}_{k+1}\\ \tilde{\ell}_k^2\end{bmatrix}$$

and $\tilde{\ell}_k^2$ and $\tilde{d}_{k+1}$ solve the system

$$\begin{bmatrix}\mu\\ \eta\end{bmatrix} = \begin{bmatrix}1 & -\zeta_k^{(\mu)}\\ 1 & -\zeta_k^{(\eta)}\end{bmatrix}\begin{bmatrix}1 & d_k\\ 0 & d_k^2\end{bmatrix}\begin{bmatrix}\tilde{d}_{k+1}\\ \tilde{\ell}_k^2\end{bmatrix} = \begin{bmatrix}1 & d_k - d_k^2\zeta_k^{(\mu)}\\ 1 & d_k - d_k^2\zeta_k^{(\eta)}\end{bmatrix}\begin{bmatrix}\tilde{d}_{k+1}\\ \tilde{\ell}_k^2\end{bmatrix}.$$

Using Cramer's rule we obtain

$$\det\begin{bmatrix}1 & d_k - d_k^2\zeta_k^{(\mu)}\\ 1 & d_k - d_k^2\zeta_k^{(\eta)}\end{bmatrix} = d_k^2\left(\zeta_k^{(\mu)} - \zeta_k^{(\eta)}\right),$$

$$\det\begin{bmatrix}\mu & d_k - d_k^2\zeta_k^{(\mu)}\\ \eta & d_k - d_k^2\zeta_k^{(\eta)}\end{bmatrix} = d_k(\mu-\eta) - d_k^2\left(\mu\zeta_k^{(\eta)} - \eta\zeta_k^{(\mu)}\right), \quad \det\begin{bmatrix}1 & \mu\\ 1 & \eta\end{bmatrix} = \eta - \mu,$$

and, therefore

$$\tilde{\ell}_k^2 = \frac{\eta-\mu}{d_k^2\left(\zeta_k^{(\mu)} - \zeta_k^{(\eta)}\right)},$$

$$\tilde{d}_{k+1} = \frac{\mu - \eta - d_k\left(\mu\zeta_k^{(\eta)} - \eta\zeta_k^{(\mu)}\right)}{d_k\left(\zeta_k^{(\mu)} - \zeta_k^{(\eta)}\right)} = \frac{\eta\zeta_k^{(\mu)} - \mu\zeta_k^{(\eta)}}{\zeta_k^{(\mu)} - \zeta_k^{(\eta)}} - d_k\tilde{\ell}_k^2.$$

Using

$$\zeta_k^{(\eta)} = \frac{1}{\bar{d}_k^{(\eta)}}, \qquad \zeta_k^{(\mu)} = \frac{1}{\bar{d}_k^{(\mu)}}$$

we obtain

$$\tilde{\ell}_k^2 = \frac{\eta - \mu}{d_k^2\left(\frac{1}{\bar{d}_k^{(\mu)}} - \frac{1}{\bar{d}_k^{(\eta)}}\right)} = \frac{\bar{d}_k^{(\mu)}\bar{d}_k^{(\eta)}}{d_k^2}\frac{\eta - \mu}{\bar{d}_k^{(\eta)} - \bar{d}_k^{(\mu)}},$$

$$\tilde{d}_{k+1} = \frac{\frac{\eta}{\bar{d}_k^{(\mu)}} - \frac{\mu}{\bar{d}_k^{(\eta)}}}{\frac{1}{\bar{d}_k^{(\mu)}} - \frac{1}{\bar{d}_k^{(\eta)}}} - d_k\tilde{\ell}_k^2 = \frac{\eta\bar{d}_k^{(\eta)} - \mu\bar{d}_k^{(\mu)}}{\bar{d}_k^{(\eta)} - \bar{d}_k^{(\mu)}} - d_k\tilde{\ell}_k^2,$$

which completes the proof. □

Using the formula (3.10) we can update the entries $\tilde{d}_k^{(\mu)}$ and $\tilde{d}_k^{(\eta)}$. Then, we can use the relations

$$\bar{d}_k^{(\mu)} = d_k - \tilde{d}_k^{(\mu)}, \qquad \bar{d}_k^{(\eta)} = d_k - \tilde{d}_k^{(\eta)}$$

and compute $\tilde{\ell}_k^{(\mu,\eta)}$ and $\tilde{d}_{k+1}^{(\mu,\eta)}$ using $\tilde{d}_k^{(\mu)}$ and $\tilde{d}_k^{(\eta)}$, as it is shown the following lemma.

Lemma 3.4 *The entries* $\tilde{\ell}_k^{(\mu,\eta)}$ *and* $\tilde{d}_{k+1}^{(\mu,\eta)}$ *from* (3.11) *can be computed using the following formulas,*

$$\left(\tilde{\ell}_k^{(\mu,\eta)}\right)^2 = \frac{\left(d_k - \tilde{d}_k^{(\mu)}\right)\left(d_k - \tilde{d}_k^{(\eta)}\right)(\eta - \mu)}{\left(\tilde{d}_k^{(\mu)} - \tilde{d}_k^{(\eta)}\right)d_k^2}, \tag{3.13}$$

$$\tilde{d}_{k+1}^{(\mu,\eta)} = \frac{d_k(\eta - \mu) + \mu\tilde{d}_k^{(\mu)} - \eta\tilde{d}_k^{(\eta)}}{\tilde{d}_k^{(\mu)} - \tilde{d}_k^{(\eta)}} - d_k\left(\tilde{\ell}_k^{(\mu,\eta)}\right)^2. \tag{3.14}$$

Proof Using $\bar{d}_k^{(\mu)} = d_k - \tilde{d}_k^{(\mu)}$, $\bar{d}_k^{(\eta)} = d_k - \tilde{d}_k^{(\eta)}$ we get

$$\bar{d}_k^{(\eta)} - \bar{d}_k^{(\mu)} = d_k - \tilde{d}_k^{(\eta)} - (d_k - \tilde{d}_k^{(\mu)}) = \tilde{d}_k^{(\mu)} - \tilde{d}_k^{(\eta)},$$

$$\eta\bar{d}_k^{(\eta)} - \mu\bar{d}_k^{(\mu)} = \eta(d_k - \tilde{d}_k^{(\eta)}) - \mu(d_k - \tilde{d}_k^{(\mu)}) = d_k(\eta - \mu) + \mu\tilde{d}_k^{(\mu)} - \eta\tilde{d}_k^{(\eta)}$$

and by substituting into the formulas (3.12) we obtain

$$\left(\tilde{\ell}_k^{(\mu,\eta)}\right)^2 = \frac{\bar{d}_k^{(\mu)}\bar{d}_k^{(\eta)}}{d_k^2}\frac{\eta - \mu}{\bar{d}_k^{(\eta)} - \bar{d}_k^{(\mu)}} = \frac{\left(d_k - \tilde{d}_k^{(\mu)}\right)\left(d_k - \tilde{d}_k^{(\eta)}\right)(\eta - \mu)}{\left(\tilde{d}_k^{(\mu)} - \tilde{d}_k^{(\eta)}\right)d_k^2},$$

$$\tilde{d}_{k+1}^{(\mu,\eta)} = \frac{\eta\bar{d}_k^{(\eta)} - \mu\bar{d}_k^{(\mu)}}{\bar{d}_k^{(\eta)} - \bar{d}_k^{(\mu)}} - d_k\left(\tilde{\ell}_k^{(\mu,\eta)}\right)^2 = \frac{d_k(\eta - \mu) + \mu\tilde{d}_k^{(\mu)} - \eta\tilde{d}_k^{(\eta)}}{\tilde{d}_k^{(\mu)} - \tilde{d}_k^{(\eta)}} - d_k\left(\tilde{\ell}_k^{(\mu,\eta)}\right)^2,$$

which completes the proof. □

In the formulas that we will use later, we need the ratio

$$\frac{\left(\tilde{\ell}_k^{(\mu,\eta)}\right)^2}{\tilde{d}_{k+1}^{(\mu,\eta)}}$$

rather than the values $\tilde{\ell}_k^{(\mu,\eta)}$ and $\tilde{d}_{k+1}^{(\mu,\eta)}$. The following lemma shows the formula for computing this ratio.

Lemma 3.5 *It holds that*

$$\frac{\left(\tilde{\ell}_k^{(\mu,\eta)}\right)^2}{\tilde{d}_{k+1}^{(\mu,\eta)}} = \frac{\left(\left(\tilde{d}_k^{(\eta)}\right)^{-1} - d_k^{-1}\right)\left(\left(\tilde{d}_k^{(\mu)}\right)^{-1} - d_k^{-1}\right)(\eta-\mu)}{\eta\left(\left(\tilde{d}_k^{(\mu)}\right)^{-1} - d_k^{-1}\right) - \mu\left(\left(\tilde{d}_k^{(\eta)}\right)^{-1} - d_k^{-1}\right)}. \tag{3.15}$$

Proof Using formulas (3.13) and (3.14) and simple algebraic manipulations we get

$$\begin{aligned}
\left(\frac{\left(\tilde{\ell}_k^{(\mu,\eta)}\right)^2}{\tilde{d}_{k+1}^{(\mu,\eta)}}\right)^{-1} &= \frac{d_k(\eta-\mu)+\mu\tilde{d}_k^{(\mu)}-\eta\tilde{d}_k^{(\eta)}}{\tilde{d}_k^{(\mu)}-\tilde{d}_k^{(\eta)}}\left(\tilde{\ell}_k^{(\mu,\eta)}\right)^{-2} - d_k \\
&= \frac{d_k(\eta-\mu)+\mu\tilde{d}_k^{(\mu)}-\eta\tilde{d}_k^{(\eta)}}{\tilde{d}_k^{(\mu)}-\tilde{d}_k^{(\eta)}}\,\frac{\left(\tilde{d}_k^{(\mu)}-\tilde{d}_k^{(\eta)}\right)d_k^2}{\left(d_k-\tilde{d}_k^{(\mu)}\right)\left(d_k-\tilde{d}_k^{(\eta)}\right)(\eta-\mu)} - d_k \\
&= d_k\left[d_k\frac{d_k(\eta-\mu)+\mu\tilde{d}_k^{(\mu)}-\eta\tilde{d}_k^{(\eta)}}{\left(d_k-\tilde{d}_k^{(\mu)}\right)\left(d_k-\tilde{d}_k^{(\eta)}\right)(\eta-\mu)} - 1\right] \\
&= \frac{d_k^2\left(\eta\tilde{d}_k^{(\mu)}-\mu\tilde{d}_k^{(\eta)}\right) - d_k\left(\tilde{d}_k^{(\mu)}\tilde{d}_k^{(\eta)}\right)(\eta-\mu)}{\left(d_k-\tilde{d}_k^{(\mu)}\right)\left(d_k-\tilde{d}_k^{(\eta)}\right)(\eta-\mu)} \\
&= \frac{\eta\left(\tilde{d}_k^{(\eta)}\right)^{-1} - \mu\left(\tilde{d}_k^{(\mu)}\right)^{-1} - d_k^{-1}(\eta-\mu)}{\left(\left(\tilde{d}_k^{(\mu)}\right)^{-1} - d_k^{-1}\right)\left(\left(\tilde{d}_k^{(\eta)}\right)^{-1} - d_k^{-1}\right)(\eta-\mu)} \\
&= \frac{\eta\left(\left(\tilde{d}_k^{(\eta)}\right)^{-1} - d_k^{-1}\right) - \mu\left(\left(\tilde{d}_k^{(\mu)}\right)^{-1} - d_k^{-1}\right)}{\left(\left(\tilde{d}_k^{(\mu)}\right)^{-1} - d_k^{-1}\right)\left(\left(\tilde{d}_k^{(\eta)}\right)^{-1} - d_k^{-1}\right)(\eta-\mu)}
\end{aligned}$$

which completes the proof. □

3.6 Another rank-two modification of T_{k+1}

The last modification of T_{k+1} that we consider, is to replace β_k in T_{k+1} by $c\,\beta_k$ where c is a given constant. The corresponding Jacobi matrix $\hat{T}^{(c)}_{k+1}$ has the same LDL^T factorization as T_{k+1} up to

$$\left(\hat{\ell}^{(c)}_k\right)^2 = \frac{c^2\beta_k^2}{d_k^2} = c^2\,\ell_k^2, \qquad \hat{d}^{(c)}_{k+1} = \alpha_{k+1} - c^2\,d_k\ell_k^2 = d_{k+1} + (1-c^2)\,d_k\ell_k^2\,,$$

and, therefore,

$$\frac{\left(\hat{\ell}^{(c)}_k\right)^2}{\hat{d}^{(c)}_{k+1}} = \frac{c^2}{d_{k+1} + (1-c^2)\,d_k\ell_k^2}\,\ell_k^2\,. \tag{3.16}$$

4 Algorithms

In this section we use the results from the previous sections for the tridiagonal matrices resulting from the Lanczos and CG algorithms. This will allow us to obtain simple formulas for computing lower and upper bounds for the A-norm of the error. Matching the LDL^T of T_k in (3.2) and (2.1), we obtain

$$\ell_j^2 = \delta_j, \qquad d_j = \gamma_{j-1}^{-1}.$$

Given two prescribed nodes μ and η, let us now consider various modifications of the matrix T_{k+1} and define

$$\tilde{\gamma}^{(\mu)}_k \equiv \left(\tilde{d}^{(\mu)}_{k+1}\right)^{-1}, \qquad \tilde{\gamma}^{(\eta)}_k \equiv \left(\tilde{d}^{(\eta)}_{k+1}\right)^{-1}, \qquad \tilde{\gamma}^{(\mu,\eta)}_k \equiv \left(\tilde{\ell}^{(\mu,\eta)}_k\right)^2\left(\tilde{d}^{(\mu,\eta)}_{k+1}\right)^{-1}.$$

Using (3.10) and (3.15) we get the updating formulas

$$\begin{aligned}
\tilde{\gamma}^{(\mu)}_0 &= \frac{1}{\mu}, & \tilde{\gamma}^{(\mu)}_k &= \frac{\tilde{\gamma}^{(\mu)}_{k-1} - \gamma_{k-1}}{\mu\left(\tilde{\gamma}^{(\mu)}_{k-1} - \gamma_{k-1}\right) + \delta_k}\,,\\
\tilde{\gamma}^{(\eta)}_0 &= \frac{1}{\eta}, & \tilde{\gamma}^{(\eta)}_k &= \frac{\tilde{\gamma}^{(\eta)}_{k-1} - \gamma_{k-1}}{\eta\left(\tilde{\gamma}^{(\eta)}_{k-1} - \gamma_{k-1}\right) + \delta_k}\,,\\
& & \tilde{\gamma}^{(\mu,\eta)}_k &= \frac{\left(\tilde{\gamma}^{(\mu)}_{k-1} - \gamma_{k-1}\right)\left(\tilde{\gamma}^{(\eta)}_{k-1} - \gamma_{k-1}\right)(\eta-\mu)}{\eta\left(\tilde{\gamma}^{(\eta)}_{k-1} - \gamma_{k-1}\right) - \mu\left(\tilde{\gamma}^{(\mu)}_{k-1} - \gamma_{k-1}\right)}.
\end{aligned}$$

Using (3.5) and

$$\|r_0\|^2\,(\ell_1\ldots\ell_{k-1})^2 = \|r_0\|^2\delta_1\ldots\delta_{k-1} = \|r_0\|^2\frac{\|r_1\|^2}{\|r_0\|^2}\cdots\frac{\|r_{k-1}\|^2}{\|r_{k-1}\|^2} = \|r_{k-1}\|^2,$$

one obtains

$$\|r_0\|^2 \left(\left[\left(\tilde{T}_{k+1}^{(\mu)} \right)^{-1} \right]_{1,1} - \left(T_k^{-1} \right)_{1,1} \right) = \frac{\ell_k^2}{\tilde{d}_{k+1}^{(\mu)}} \|r_{k-1}\|^2 = \frac{\|r_k\|^2}{\tilde{d}_{k+1}^{(\mu)}}.$$

We can now compute the quantities of the form (2.18),

$$g_k \equiv \|r_0\|^2 \left(\left(T_{k+1}^{-1} \right)_{1,1} - \left(T_k^{-1} \right)_{1,1} \right) = \gamma_k \|r_k\|^2,$$

$$g_k^{(\mu)} \equiv \|r_0\|^2 \left(\left[\left(\tilde{T}_{k+1}^{(\mu)} \right)^{-1} \right]_{1,1} - \left(T_k^{-1} \right)_{1,1} \right) = \tilde{\gamma}_k^{(\mu)} \|r_k\|^2,$$

$$g_k^{(\eta)} \equiv \|r_0\|^2 \left(\left[\left(\tilde{T}_{k+1}^{(\eta)} \right)^{-1} \right]_{1,1} - \left(T_k^{-1} \right)_{1,1} \right) = \tilde{\gamma}_k^{(\eta)} \|r_k\|^2,$$

$$g_k^{(\mu,\eta)} \equiv \|r_0\|^2 \left(\left[\left(\tilde{T}_{k+1}^{(\mu,\eta)} \right)^{-1} \right]_{1,1} - \left(T_k^{-1} \right)_{1,1} \right) = \tilde{\gamma}_k^{(\mu,\eta)} \|r_{k-1}\|^2.$$

Hence, the quantities for the three different rules are given almost in the same form. The index of the residual differs for the Gauss-Lobatto rule because the term $(\tilde{\ell}_k^{(\mu,\eta)})^2$ is incorporated in $\tilde{\gamma}_k^{(\mu,\eta)}$. Using the updating formulas for the coefficients $\tilde{\gamma}_k^{(\mu)}$, $\tilde{\gamma}_k^{(\eta)}$ and $\tilde{\gamma}_k^{(\mu,\eta)}$ we can derive updating formulas for $g_k^{(\mu)}$, $g_k^{(\eta)}$ and $g_k^{(\mu,\eta)}$ such that the coefficients $\tilde{\gamma}_k^{(\mu)}$, $\tilde{\gamma}_k^{(\eta)}$ and $\tilde{\gamma}_k^{(\mu,\eta)}$ need not to be computed. Since

$$\begin{aligned}\tilde{\gamma}_k^{(\mu)} &= \frac{\tilde{\gamma}_{k-1}^{(\mu)} - \gamma_{k-1}}{\mu \left(\tilde{\gamma}_{k-1}^{(\mu)} - \gamma_{k-1} \right) + \delta_k} = \frac{\|r_{k-1}\|^2 \tilde{\gamma}_{k-1}^{(\mu)} - \|r_{k-1}\|^2 \gamma_{k-1}}{\mu \left(\|r_{k-1}\|^2 \tilde{\gamma}_{k-1}^{(\mu)} - \|r_{k-1}\|^2 \gamma_{k-1} \right) + \|r_k\|^2} \\ &= \frac{g_{k-1}^{(\mu)} - g_{k-1}}{\mu \left(g_{k-1}^{(\mu)} - g_{k-1} \right) + \|r_k\|^2},\end{aligned}$$

the formulas for $g_k^{(\mu)}$ and $g_k^{(\eta)}$ can be written as

$$g_0^{(\mu)} = \frac{\|r_0\|^2}{\mu}, \qquad g_0^{(\eta)} = \frac{\|r_0\|^2}{\eta}, \tag{4.1}$$

$$g_k^{(\mu)} = \|r_k\|^2 \frac{g_{k-1}^{(\mu)} - g_{k-1}}{\mu \left(g_{k-1}^{(\mu)} - g_{k-1} \right) + \|r_k\|^2}, \qquad g_k^{(\eta)} = \|r_k\|^2 \frac{g_{k-1}^{(\eta)} - g_{k-1}}{\mu \left(g_{k-1}^{(\eta)} - g_{k-1} \right) + \|r_k\|^2}.$$

The formula for $g_k^{(\mu,\eta)} = \tilde{\gamma}_k^{(\mu,\eta)} \|r_{k-1}\|^2$ takes the form

$$g_k^{(\mu,\eta)} = \tilde{\gamma}_k^{(\mu,\eta)} \|r_{k-1}\|^2 = \|r_{k-1}\|^2 \frac{\left(\tilde{\gamma}_{k-1}^{(\mu)} - \gamma_{k-1}\right)\left(\tilde{\gamma}_{k-1}^{(\eta)} - \gamma_{k-1}\right)(\eta - \mu)}{\eta\left(\tilde{\gamma}_{k-1}^{(\eta)} - \gamma_{k-1}\right) - \mu\left(\tilde{\gamma}_{k-1}^{(\mu)} - \gamma_{k-1}\right)}$$

$$= \frac{\left(\|r_{k-1}\|^2\tilde{\gamma}_{k-1}^{(\mu)} - \|r_{k-1}\|^2\gamma_{k-1}\right)\left(\|r_{k-1}\|^2\tilde{\gamma}_{k-1}^{(\eta)} - \|r_{k-1}\|^2\gamma_{k-1}\right)(\eta - \mu)}{\eta\left(\|r_{k-1}\|^2\tilde{\gamma}_{k-1}^{(\eta)} - \|r_{k-1}\|^2\gamma_{k-1}\right) - \mu\left(\|r_{k-1}\|^2\tilde{\gamma}_{k-1}^{(\mu)} - \|r_{k-1}\|^2\gamma_{k-1}\right)}$$

$$= \frac{\left(g_{k-1}^{(\mu)} - g_{k-1}\right)\left(g_{k-1}^{(\eta)} - g_{k-1}\right)(\eta - \mu)}{\eta\left(g_{k-1}^{(\eta)} - g_{k-1}\right) - \mu\left(g_{k-1}^{(\mu)} - g_{k-1}\right)}.$$

Summarizing, starting with the formulas (4.1) we obtain for $k = 1, \ldots$ updating formulas taking the simple following form,

$$g_{k-1} = \gamma_{k-1}\|r_{k-1}\|^2, \qquad \Delta_{k-1}^{(\mu)} \equiv g_{k-1}^{(\mu)} - g_{k-1}, \qquad \Delta_{k-1}^{(\eta)} \equiv g_{k-1}^{(\eta)} - g_{k-1}, \tag{4.2}$$

and

$$g_k^{(\mu)} = \|r_k\|^2 \frac{\Delta_{k-1}^{(\mu)}}{\mu\Delta_{k-1}^{(\mu)} + \|r_k\|^2}, \quad g_k^{(\eta)} = \|r_k\|^2 \frac{\Delta_{k-1}^{(\eta)}}{\eta\Delta_{k-1}^{(\eta)} + \|r_k\|^2}, \tag{4.3}$$

$$g_k^{(\mu,\eta)} = (\eta - \mu)\frac{\Delta_{k-1}^{(\mu)}\Delta_{k-1}^{(\eta)}}{\eta\Delta_{k-1}^{(\eta)} - \mu\Delta_{k-1}^{(\mu)}}. \tag{4.4}$$

Now we have all the needed material to compute bounds. We distinguish three parts in the algorithm for running CG and obtaining bounds for the A-norm of the error.

1. The first part is simply the **CG-iteration** which computes two scalars and updates the vectors,

$$\begin{aligned}
\gamma_{k-1} &= \frac{r_{k-1}^T r_{k-1}}{p_{k-1}^T A p_{k-1}}, \\
x_k &= x_{k-1} + \gamma_{k-1} p_{k-1}, \\
r_k &= r_{k-1} - \gamma_{k-1} A p_{k-1}, \\
\delta_k &= \frac{r_k^T r_k}{r_{k-1}^T r_{k-1}}, \\
p_k &= r_k + \delta_k p_{k-1}.
\end{aligned}$$

2. The second part is called the **Quadrature part**. It computes the quantities g_{k-1}, $g_k^{(\mu)}$, $g_k^{(\eta)}$ and $g_k^{(\mu,\eta)}$ using the formulas (4.2), (4.3), and (4.4) if we are interested in computing the Gauss, Gauss-Radau and Gauss-Lobatto bounds.
3. The third part is called the **Estimates part** for iteration $k-d$. In this paper, we use the following way of constructing the estimates from g_{k-1}, $g_k^{(\mu)}$, $g_k^{(\eta)}$ and $g_k^{(\mu,\eta)}$. If $k > d$, then compute

$$Q_{k-d,d} = \sum_{j=k-d+1}^{k} g_j,$$

$$E_{k-d} = \sqrt{Q_{k-d,d}}, \qquad E_{k-d}^{(\mu)} = \sqrt{Q_{k-d,d} + g_k^{(\mu)}},$$

$$E_{k-d}^{(\eta)} = \sqrt{Q_{k-d,d} + g_k^{(\eta)}}, \qquad E_{k-d}^{(\mu,\eta)} = \sqrt{Q_{k-d,d} + g_k^{(\mu,\eta)}}$$

E_{k-d} is the Gauss lower bound. If $\mu < \lambda_1$ (resp. $\eta > \lambda_N$) $E_{k-d}^{(\mu)}$ (resp. $E_{k-d}^{(\eta)}$) is the Gauss-Radau upper (resp. lower) bound and $E_{k-d}^{(\mu,\eta)}$ is the Gauss-Lobatto upper bound. Note that d is a given positive integer indicating how many steps of CG should be precomputed to have the estimate at iteration $k-d$.

Algorithm 5 recalls the CGQL algorithm described in [21] and [13]. Algorithm 6 is the new version using the simpler formulas derived in this paper. We denote it as CGQ (in reference to CGQL, dropping the L because we do not use the Lanczos coefficients any longer). We see that computing the Gauss-Radau upper bound is almost as simple as computing the Gauss lower bound provided we have a $\mu < \lambda_1$.

In practical computations we usually only use the lower bound based on Gauss quadrature and the upper bound based on Gauss-Radau when a lower bound of the smallest eigenvalue of A is available. In that case we only need to compute g_k and $g_k^{(\mu)}$ using the formulas from Algorithm 6.

Note that in the same way we can also obtain formulas for the anti-Gauss quadrature rule [4, 20]. Defining

$$\hat{g}_k^{(c)} \equiv \|r_0\|^2 \left(\left[\left(\hat{T}_{k+1}^{(c)} \right)^{-1} \right]_{1,1} - \left(T_k^{-1} \right)_{1,1} \right)$$

and using the formula (3.16) we obtain

$$\hat{g}_k^{(c)} = c^2 \frac{g_k g_{k-1}}{g_{k-1} + (1-c^2) g_k} . \tag{4.5}$$

Algorithm 5 *CGQL* (Conjugate Gradients and Quadrature via Lanczos coefficients)

```
input $A, b, x_0, \mu$
$r_0 = b - Ax_0,\ p_0 = r_0$
$\delta_0 = 0, \gamma_{-1} = 1, c_1 = 1, \beta_0 = 0, d_0 = 1, \tilde{\alpha}_1^{(\mu)} = \mu, \tilde{\alpha}_1^{(\eta)} = \eta$
for $k = 1, \ldots,$ until convergence do
    CG-iteration ($k$)
    $$\boxed{\begin{aligned}
    \alpha_k &= \frac{1}{\gamma_{k-1}} + \frac{\delta_{k-1}}{\gamma_{k-2}},\ \beta_k^2 = \frac{\delta_k}{\gamma_{k-1}^2}\\
    d_k &= \alpha_k - \frac{\beta_{k-1}^2}{d_{k-1}},\ g_k = \|r_0\|^2 \frac{c_k^2}{d_k},\\
    \bar{d}_k^{(\mu)} &= \alpha_k - \tilde{\alpha}_k^{(\mu)},\ \tilde{\alpha}_{k+1}^{(\mu)} = \mu + \frac{\beta_k^2}{\bar{d}_k^{(\mu)}},\ g_k^{(\mu)} = \|r_0\|^2 \frac{\beta_k^2 c_k^2}{d_k\left(\tilde{\alpha}_{k+1}^{(\mu)} d_k - \beta_k^2\right)}\\
    \bar{d}_k^{(\eta)} &= \alpha_k - \tilde{\alpha}_k^{(\eta)},\ \tilde{\alpha}_{k+1}^{(\eta)} = \eta + \frac{\beta_k^2}{\bar{d}_k^{(\eta)}},\ g_k^{(\eta)} = \|r_0\|^2 \frac{\beta_k^2 c_k^2}{d_k\left(\tilde{\alpha}_{k+1}^{(\eta)} d_k - \beta_k^2\right)}\\
    \tilde{\alpha}_{k+1}^{(\mu,\eta)} &= \frac{\bar{d}_k^{(\mu)} \bar{d}_k^{(\eta)}}{\bar{d}_k^{(\eta)} - \bar{d}_k^{(\mu)}} \left(\frac{\eta}{\bar{d}_k^{(\mu)}} - \frac{\mu}{\bar{d}_k^{(\eta)}}\right),\ \left[\tilde{\beta}_k^{(\mu,\eta)}\right]^2 = \frac{\bar{d}_k^{(\mu)} \bar{d}_k^{(\eta)}}{\bar{d}_k^{(\eta)} - \bar{d}_k^{(\mu)}} (\eta - \mu)\\
    g_k^{(\mu,\eta)} &= \|r_0\|^2 \frac{\left[\tilde{\beta}_k^{(\mu,\eta)}\right]^2 c_k^2}{d_k\left(\tilde{\alpha}_{k+1}^{(\mu,\eta)} d_k - \left[\tilde{\beta}_k^{(\mu,\eta)}\right]^2\right)}\\
    c_{k+1}^2 &= \frac{\beta_k^2 c_k^2}{d_k^2}
    \end{aligned}}$$
    Estimates($k$,$d$)
end for
```

The anti-Gauss quadrature estimate can be then computed analogously,

$$\hat{E}_{k-d}^{(c)} = \sqrt{Q_{k-d,d} + \hat{g}_k^{(c)}}\,.$$

Algorithm 6 *CGQ* (Conjugate Gradients and Quadrature)

```
input $A, b, x_0, \mu, \eta$
$r_0 = b - Ax_0$, $p_0 = r_0$
$g_0^{(\mu)} = \frac{\|r_0\|^2}{\mu}$, $g_0^{(\eta)} = \frac{\|r_0\|^2}{\eta}$
for $k = 1, \ldots,$ until convergence do
    CG-iteration($k$)
    $g_{k-1} = \gamma_{k-1}\|r_{k-1}\|^2,$
    $\Delta_{k-1}^{(\mu)} = g_{k-1}^{(\mu)} - g_{k-1}\,, \qquad g_k^{(\mu)} = \frac{\|r_k\|^2 \Delta_{k-1}^{(\mu)}}{\mu\Delta_{k-1}^{(\mu)} + \|r_k\|^2}$
    $\Delta_{k-1}^{(\eta)} = g_{k-1}^{(\eta)} - g_{k-1}\,, \qquad g_k^{(\eta)} = \frac{\|r_k\|^2 \Delta_{k-1}^{(\eta)}}{\eta\Delta_{k-1}^{(\eta)} + \|r_k\|^2}$
    $g_k^{(\mu,\eta)} = (\eta - \mu)\frac{\Delta_{k-1}^{(\mu)}\Delta_{k-1}^{(\eta)}}{\eta\Delta_{k-1}^{(\eta)} - \mu\Delta_{k-1}^{(\mu)}}\,.$
    Estimates($k$,$d$)
end for
```

5 Error estimation in preconditioned CG

In the standard view of preconditioning, the CG method is thought of as being applied to a "preconditioned" system

$$\hat{A}\hat{x} = \hat{b}, \qquad \hat{A} = L^{-1}AL^{-T}, \quad \hat{b} = L^{-1}b, \tag{5.1}$$

where L represents a nonsingular (eventually lower triangular) matrix. Denoting the corresponding CG coefficients and vectors with hat and defining

$$x_k \equiv L^{-T}\hat{x}_k, \quad r_k \equiv L\hat{r}_k, \quad p_k \equiv L^{-T}\hat{p}_k, \quad z_k \equiv L^{-T}L^{-1}r_k \equiv P^{-1}r_k,$$

(here x_k and r_k represent the approximate solution and residual for the original problem $Ax = b$), we obtain the standard version of the preconditioned CG (PCG) method which involves only $P = LL^T$; for more details see, e.g. [22] or [29].

Since

$$\|\hat{r}_k\|^2 = r_k^T L^{-T} L^{-1} r_k = r_k^T P^{-1} r_k = (r_k, z_k)\,,$$

$$\|\hat{x} - \hat{x}_k\|_{\hat{A}} = \left(L^T x - L^T x_k\right)^T L^{-1}AL^{-T}\left(L^T x - L^T x_k\right) = \|x - x_k\|_A^2,$$

the A-norm of the error in PCG can be estimated similarly as in ordinary CG. One can compute the quadratures-based estimates of the A-norm of the error

Algorithm 7 Preconditioned CGQ (PCGQ) algorithm

```
input $A, b, x_0, P, \mu, \eta$
$r_0 = b - Ax_0,\ z_0 = P^{-1}r_0,\ p_0 = z_0$
$g_0^{(\mu)} = \frac{(r_0, z_0)}{\mu},\ g_0^{(\eta)} = \frac{(r_0, z_0)}{\eta}$
for $k = 1, \ldots, n$ until convergence do
    $\hat{\gamma}_{k-1} = \frac{z_{k-1}^T r_{k-1}}{p_{k-1}^T A p_{k-1}}$
    $x_k = x_{k-1} + \hat{\gamma}_{k-1} p_{k-1}$
    $r_k = r_{k-1} - \hat{\gamma}_{k-1} A p_{k-1}$
    $z_k = P^{-1} r_k$
    $\hat{\delta}_k = \frac{z_k^T r_k}{z_{k-1}^T r_{k-1}}$
    $p_k = z_k + \hat{\delta}_k p_{k-1}$
    $g_{k-1} = \hat{\gamma}_{k-1}(r_{k-1}, z_{k-1}),$
    $\Delta_{k-1}^{(\mu)} = g_{k-1}^{(\mu)} - g_{k-1}, \qquad g_k^{(\mu)} = \frac{(r_k, z_k)\Delta_{k-1}^{(\mu)}}{\mu\Delta_{k-1}^{(\mu)} + (r_k, z_k)}$
    $\Delta_{k-1}^{(\eta)} = g_{k-1}^{(\eta)} - g_{k-1}, \qquad g_k^{(\eta)} = \frac{(r_k, z_k)\Delta_{k-1}^{(\eta)}}{\eta\Delta_{k-1}^{(\eta)} + (r_k, z_k)}$
    $g_k^{(\mu,\eta)} = (\eta - \mu)\frac{\Delta_{k-1}^{(\mu)}\Delta_{k-1}^{(\eta)}}{\eta\Delta_{k-1}^{(\eta)} - \mu\Delta_{k-1}^{(\mu)}}.$
    Estimates($k$,$d$)
end for
```

using the PCG coefficients $\hat{\gamma}_{k-1}$ and inner products (r_k, z_k) (instead of using $\|\hat{r}_k\|^2$). The resulting Algorithm 7 is called PCGQ.

6 Numerical experiments

We present two examples where we demonstrate the effectivity of the new formula

$$g_k^{(\mu)} = \frac{\|r_k\|^2\left(g_{k-1}^{(\mu)} - g_{k-1}\right)}{\mu\left(g_{k-1}^{(\mu)} - g_{k-1}\right) + \|r_k\|^2}$$

that is key for computing the Gauss-Radau quadrature estimate (GR-estimate). Both examples are chosen such that many iterations are necessary for CG (or PCG) to converge, and such that the influence of rounding errors is substantial (in both examples we observe a significant delay of convergence).

The aim of numerical experiments is not to focus on the question of how to estimate the A-norm of the error and when to stop the algorithm, but to compare the new and the old formula for computing the existing estimate. For further discussion on estimating the A-norm of the error, stopping criteria, and numerical experiments we refer to [1, 4, 5, 22, 23, 28, 29]. The following experiments are performed in Matlab 7.13 (R2011b).

In the first numerical experiment we solve the system $Ax = b$ with the matrix `bcsstk01` (Harwell-Boeing collection) of order $n = 48$; see also numerical experiments in [23, Chapter 7]. The right-hand side b has been chosen such that b has equal components in the eigenvector basis, and such that $\|b\| = 1$. We choose $x_0 = 0$, $d = 1$, and $\mu = 3.417267e + 3$ (the smallest eigenvalue of the matrix is 3.417267562666500e+3).

In the upper part of Fig. 1 we plot the A-norm of the error (bold solid line), the Gauss quadrature estimate E_{k-1} (solid line, GQ-estimate), and the GR-estimate $E_{k-1}^{(\mu)}$, where $g_k^{(\mu)}$ is computed using CGQ (dashed line) and using CGQL (dotted line). Visually, it is not possible to distinguish between the estimate $E_{k-1}^{(\mu)}$ computed using CGQ (the new formula) and CGQL (the old formula). However, the lower part of Fig. 1 indicates that the new formula is not only simpler but also more accurate than the old one.

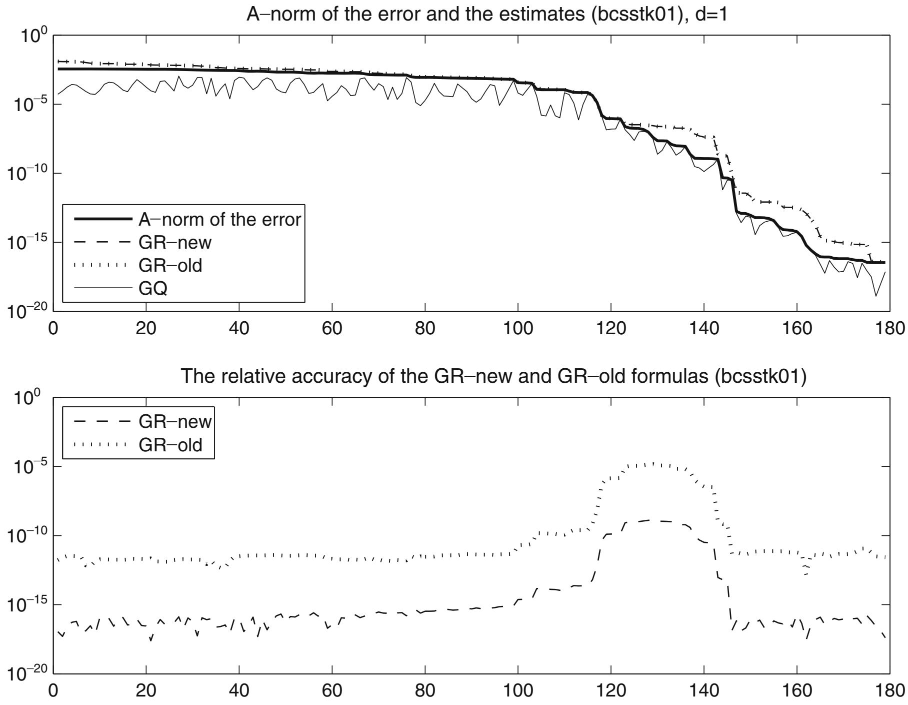


Fig. 1 The *upper part*: the A-norm of the error (*bold solid*), the GQ-estimate (*solid*) and the GR-estimate computed using CGQ (*dashed*) and CGQL (*dotted*). The *lower part*: relative accuracy of the GR-estimate computed using CGQ (*dashed*) and CGQL (*dotted*)

In the lower part of Fig. 1 we compare the relative accuracy of results computed using the two formulas. First, we compute the quantities $\|r_k\|^2$ and γ_k using the standard double precision CG. Second, we use variable-precision arithmetic in Matlab (Symbolic Toolbox, 64 digits) to get a reasonably accurate value of $E_{k-1}^{(\mu)}$ (computed from the double precision quantities $\|r_k\|^2$ and γ_k). Third, we compute $E_{k-1}^{(\mu)}$ using the standard double precision arithmetic and the two formulas; the corresponding computed values are denoted by $\hat{E}_{k-1}^{(\mu,new)}$ and $\hat{E}_{k-1}^{(\mu,old)}$. Finally, in the lower part of Fig. 1 we plot the quantities

$$\left|\frac{\hat{E}_{k-1}^{(\mu,new)} - E_{k-1}^{(\mu)}}{E_{k-1}^{(\mu)}}\right| \text{ (dashed)} \qquad \text{and} \qquad \left|\frac{\hat{E}_{k-1}^{(\mu,old)} - E_{k-1}^{(\mu)}}{E_{k-1}^{(\mu)}}\right| \text{ (dotted)}$$

that characterize the relative accuracy of the computed value of $E_{k-1}^{(\mu)}$. We can observe that the new formula is less sensitive to rounding error.

In the second numerical experiment we solve the system $Ax = b$ with the matrix `msc04515` (The University of Florida sparse matrix collection) of order $n = 4515$; see also numerical experiments in [23, Chapter 7]. The right-hand side b has again been chosen such that b has equal components in the

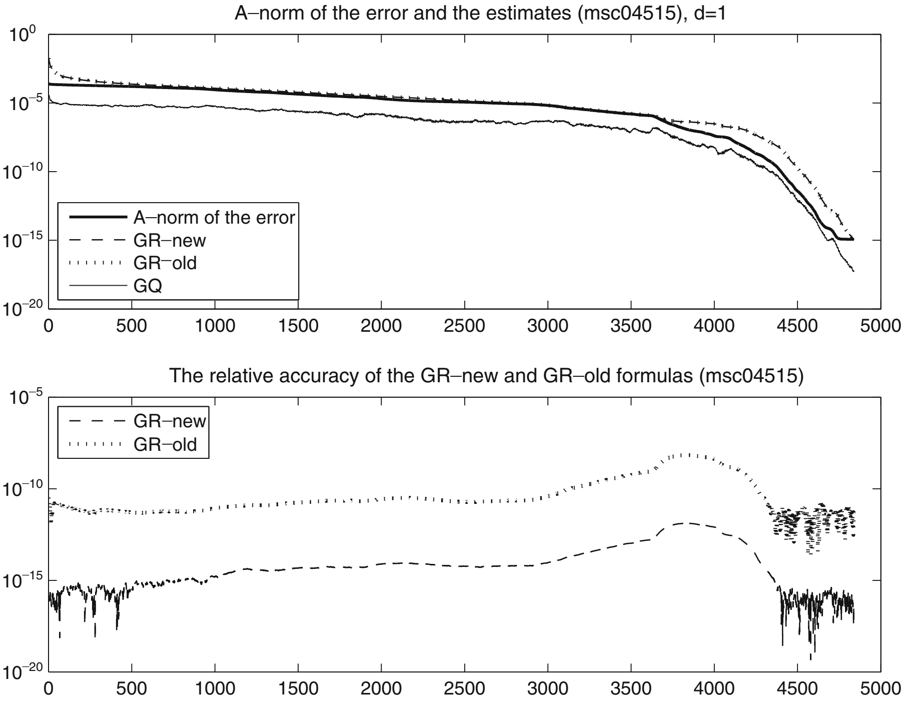


Fig. 2 The *upper part*: the A-norm of the error (bold solid), the GQ-estimate (*solid*) and the GR-estimate computed using CGQ (*dashed*) and CGQL (*dotted*). The *lower part*: relative accuracy of the GR-estimate computed using CGQ (*dashed*) and CGQL (*dotted*)

eigenvector basis, and $\|b\| = 1$. We choose $x_0 = 0$, $d = 1$, and use the diagonal preconditioner (the preconditioned matrix is diagonally normalized with 1's on the diagonal). The value of μ is given by $\mu = 1.75e - 6$ (the smallest eigenvalue of the preconditioned matrix is equal to 1.751795139099631e-6).

In Fig. 2 we observe more or less the same behaviour as for the first example; the GR-estimate computed using the new formula gives visually the same results, however, the new formula is simpler and more accurate that the old one.

7 Conclusion

In this paper we have described how the bounds based on Gauss quadrature rules for the CG A-norm of the error can be computed in a simple way. In particular, for the Gauss-Radau and Gauss-Lobatto bounds, the preceding implementations computed explicitly the entries of the (modified) Jacobi matrices and used them to compute the bounds. Here we exploited the fact that the LDL^T factorization of the corresponding Jacobi matrix is available in CG and showed how to update LDL^T factorizations of modified Jacobi matrices. The bounds are then computed directly from the known entries of LDL^T factorizations. The algebraic derivation of the new formulas is more difficult than it was when using Jacobi matrices but, in the end, the formulas are simpler. Obtaining simple formulas is a prerequisite for analyzing the behaviour of the bounds in finite precision arithmetic and also for a better understanding of their dependence on the auxiliary parameters μ and η which are lower and upper bounds (or estimates) of the smallest and largest eigenvalues. Numerical experiments predict that the new formulas are less prone to the growth of rounding errors. Therefore, we hope that these improvements will help the implementation of quadrature-based error bounds into existing and forthcoming CG codes.

References

1. Arioli, M.: A stopping criterion for the conjugate gradient algorithms in a finite element method framework. Numer. Math. **97**, 1–24 (2004)
2. Auchmuty, G.: A posteriori error estimates for linear equations. Numer. Math. **61**, 1–6 (1992)
3. Brezinski, C.: Error estimates for the solution of linear systems. SIAM J. Sci. Comput. **21**, 764–781 (1999)
4. Calvetti, D., Morigi, S., Reichel, L., Sgallari, F.: Computable error bounds and estimates for the conjugate gradient method. Numer. Algor. **25**, 75–88 (2000)
5. Calvetti, D., Morigi, S., Reichel, L., Sgallari, F.: An iterative method with error estimators. J. Comput. Appl. Math. **127**, 93–119 (2001)
6. Dahlquist, G., Eisenstat, S.C., Golub, G.H.: Bounds for the error of linear systems of equations using the theory of moments. J. Math. Anal. Appl. **37**, 151–166 (1972)
7. Dahlquist, G., Golub, G.H., Nash, S.G.: Bounds for the error in linear systems. In: Semi-infinite Programming (Proc. Workshop, Bad Honnef, 1978). Lecture Notes in Control and Information Sci., vol. 15, pp. 154–172. Springer, Berlin (1979)

8. Fischer, B., Golub, G.H.: On the error computation for polynomial based iteration methods. In: Recent Advances in Iterative Methods. IMA Vol. Math. Appl., vol. 60, pp. 59–67. Springer, New York (1994)
9. Gautschi, W.: Orthogonal Polynomials: Computation and Approximation. Oxford University Press, UK (2004)
10. Golub, G.H.: Some modified matrix eigenvalue problems. SIAM Rev. **15**, 318–334 (1973)
11. Golub, G.H., Meurant, G.: Matrices, moments and quadrature. In: Numerical Analysis 1993 (Dundee, 1993). Pitman Res. Notes Math. Ser., Longman Sci. Tech., vol. 303, pp. 105–156. Harlow (1994)
12. Golub, G.H., Meurant, G.: Matrices, moments and quadrature. II. How to compute the norm of the error in iterative methods. BIT **37**, 687–705 (1997)
13. Golub, G.H., Meurant, G.: Matrices, Moments and Quadrature with Applications. Princeton University Press, USA (2010)
14. Golub, G.H., Strakoš, Z.: Estimates in quadratic formulas. Numer. Algor. **8**, 241–268 (1994)
15. Golub, G.H., Van Loan, C.F.: Matrix Computations. Johns Hopkins Studies in the Mathematical Sciences, 3rd edn. Johns Hopkins University Press, Baltimore, MD (1996)
16. Golub, G.H., Welsch, J.H.: Calculation of Gauss quadrature rules. Math. Comp. **23**, 221–230 (1969) (addendum, ibid., **23**, A1–A10 (1969))
17. Hestenes, M.R., Stiefel, E.: Methods of conjugate gradients for solving linear systems. J. Research Nat. Bur. Standards **49**, 409–436 (1952)
18. Jiránek, P., Strakoš, Z., Vohralík, M.: A posteriori error estimates including algebraic error and stopping criteria for iterative solvers. SIAM J. Sci. Comput. **32**, 1567–1590 (2010)
19. Lanczos, C.: Solution of systems of linear equations by minimized iterations. J. Research Nat. Bur. Standards **49**, 33–53 (1952)
20. Laurie, D.P.: Anti-Gaussian quadrature formulas. Math. Comp. **65**, 739–747 (1996)
21. Meurant, G.: The computation of bounds for the norm of the error in the conjugate gradient algorithm. Numer. Algor. **16**, 77–87 (1998)
22. Meurant, G.: Numerical experiments in computing bounds for the norm of the error in the preconditioned conjugate gradient algorithm. Numer. Algor. **22**, 353–365 (1999)
23. Meurant, G.: The Lanczos and Conjugate Gradient Algorithms, from Theory to Finite Precision Computations. Software, Environments, and Tools, vol. 19. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (2006)
24. Meurant, G., Strakoš, Z.: The Lanczos and conjugate gradient algorithms in finite precision arithmetic. Acta Numer. **15**, 471–542 (2006)
25. Parlett, B.N.: The new qd algorithms. Acta Numer. **15**, 459–491 (1995)
26. Parlett, B.N., Dhillon, I.S.: Relatively robust representations of symmetric tridiagonals. In: Proceedings of the International Workshop on Accurate Solution of Eigenvalue Problems (University Park, PA, 1998), vol. 309, pp. 121–151 (2000)
27. Stoer, J., Bulirsch, R.: Introduction to Numerical Analysis, 2nd edn. Springer, Berlin, Germany (1983)
28. Strakoš, Z., Tichý, P.: On error estimation in the conjugate gradient method and why it works in finite precision computations. Electron. Trans. Numer. Anal. **13**, 56–80 (2002)
29. Strakoš, Z., Tichý, P.: Error estimation in preconditioned conjugate gradients. BIT **45**, 789–817 (2005)
30. Wilf, H.S.: Mathematics for the Physical Sciences. Wiley, New York, USA (1962)

Numer Algor
DOI 10.1007/s11075-014-9868-2

ERRATUM

Erratum to: On computing quadrature-based bounds for the A-norm of the error in conjugate gradients

Gérard Meurant · Petr Tichý

Erratum to: Numer Algor (2013) 62:163-191
DOI 10.1007/s11075-012-9591-9

1 The algorithm CGQL

In our paper [1] we found two typographical errors that can negatively influence the correct implementation of the algorithms by potential users. Therefore, we consider important to present this erratum.

The first typographical error in [1] appears in the main part of Algorithm 5 CGQL (surrounded by the frame) on page 185. To be consistent with the definition of g_k on page 182, the symbol g_k that appears only once in the main part of Algorithm 5 CGQL, should by replaced by g_{k-1}. Hence,

$$g_{k-1} = \|r_0\|^2 \frac{c_k^2}{d_k}$$

is the correct formula.

The online version of the original article can be found at http://dx.doi.org/10.1007/s11075-012-9591-9

G. Meurant
30 rue du sergent Bauchat, 75012 Paris, France
e-mail: gerard.meurant@gmail.com

P. Tichý (✉)
Institute of Computer Science, Academy of Sciences of the Czech Republic,
Pod Vodárenskou věží 2, 18207 Prague, Czech Republic
e-mail: tichy@cs.cas.cz

Published online: 28 May 2014

2 The Estimates part

The second typographical error is closely related with the first one and it is hidden in the “Estimates part” on page 184, that is common for both algorithms, CGQL and CGQ. It is again about wrong indexing of g_k, this time in the definition of $Q_{k-d,d}$. The corrected text on page 184 is the following: If $k \geq d$, then compute

$$Q_{k-d,d} = \sum_{j=k-d}^{k-1} g_j.$$

3 Final comments

We would like to assure the interested readers that these typographical errors did not appear in our Matlab codes so that the numerical experiments presented in our paper are correct. As it is clear form the above text, the first error arose in the CGQL algorithm where the CG related quantities are indexed from 0 while the tridiagonal matrices related quantities are indexed from 1. The first typographical error caused then the second one in the Estimates part.

References

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

Numer. Math. (2007) 107:147–174
DOI 10.1007/s00211-007-0078-x

Numerische Mathematik

On sensitivity of Gauss–Christoffel quadrature

Dianne P. O'Leary · Zdeněk Strakoš · Petr Tichý

Received: 8 September 2006 / Revised: 15 March 2007 / Published online: 14 April 2007

Abstract In numerical computations the question *how much does a function change under perturbations of its arguments* is of central importance. In this work, we investigate sensitivity of Gauss–Christoffel quadrature with respect to small perturbations of the distribution function. In numerical quadrature, a definite integral is approximated by a finite sum of functional values evaluated at given quadrature nodes and multiplied by given weights. Consider a sufficiently smooth integrated function uncorrelated with the perturbation of the distribution function. Then it seems natural that given the same number of function evaluations, the difference between the quadrature approximations is of the same order as the difference between the (original and perturbed) approximated integrals. That is perhaps one of the reasons why, to our knowledge, the sensitivity question has not been formulated and addressed in the literature, though

The work of the first author was supported by the National Science Foundation under Grants CCR-0204084 and CCF-0514213. The work of the other two authors was supported by the Program Information Society under project 1ET400300415 and by the Institutional Research Plan AV0Z100300504.
P. Tichý in the years 2003–2006 on leave at the Institute of Mathematics, TU Berlin, Germany.

D. P. O'Leary ( )
Department of Computer Science and Institute for Advanced Computer Studies,
University of Maryland, College Park, MD 20742, USA
e-mail: oleary@cs.umd.edu

Z. Strakoš
Institute of Computer Science and Faculty of Mathematics and Physics,
Academy of Sciences of the Czech Republic and Charles University, Prague, Czech Republic
e-mail: strakos@cs.cas.cz

P. Tichý
Institute of Computer Science, Academy of Sciences of the Czech Republic,
Prague, Czech Republic
e-mail: tichy@cs.cas.cz

several other sensitivity problems, motivated, in particular, by computation of the quadrature nodes and weights from moments, have been thoroughly studied by many authors. We survey existing particular results and show that *even a small perturbation of a distribution function can cause large differences in Gauss–Christoffel quadrature estimates*. We then discuss conditions under which the Gauss–Christoffel quadrature is insensitive under perturbation of the distribution function, present illustrative examples, and relate our observations to known conjectures on some sensitivity problems.

1 Introduction

The computation of orthogonal polynomials and Gauss–Christoffel quadrature draws upon several fields from classical analysis and approximation theory as well as modern numerical linear algebra. It has been intensively studied by many generations of mathematicians.

Here we consider linear functionals in the form of the Riemann–Stieltjes integral and restrict ourselves to distribution functions that are nondecreasing on a finite interval $[a, b]$ on the real line. By the k-point Gauss–Christoffel quadrature we mean the approximation of a given Riemann–Stieltjes integral

$$I_\omega(f) = \int_a^b f(x)\, d\omega(x) \tag{1}$$

by the discrete linear functional

$$I_\omega^k(f) = \sum_{j=1}^{k} \vartheta_j f(t_j)\,,$$

determined by nodes $a \le t_1 < \cdots < t_k \le b$ and positive weights $\{\vartheta_1, \ldots, \vartheta_k\}$ such that $I_\omega^k(f) = I_\omega(f)$ whenever f is a polynomial of degree at most $2k-1$ [6, Sect. 2.7], [18]. The recent encyclopedic book by Gautschi [23], his surveys [18,24] and the survey by Laurie [38] describe the state-of-the-art of Gauss–Christoffel quadrature computation, and can be recommended as fundamental reading for anyone interested in related problems.

In this paper we investigate sensitivity of Gauss–Christoffel quadrature with respect to small perturbations in the distribution function. Suppose we have two distribution functions $\omega(x)$ and $\tilde{\omega}(x)$ which are nondecreasing on the finite interval $[a, b]$ and close to each other. We are interested in estimating the two integrals

$$I_\omega = \int_a^b f(x)\, d\omega(x), \quad I_{\tilde{\omega}} = \int_a^b f(x)\, d\tilde{\omega}(x). \tag{2}$$

Although it seems natural to expect that the Gauss–Christoffel quadrature estimates of the same degree will be close when f is sufficiently smooth (and also uncorrelated with the difference between the given distribution functions), it is not clear that this is true. If we use Gauss–Christoffel quadrature to compute the estimates, then $\omega(x)$ and $\tilde{\omega}(x)$ induce different sequences of orthogonal polynomials. Therefore, the quadrature weights and nodes for the same degree of the quadrature might be different from each other and in fact can be sensitive to small perturbations to the distribution function. Indeed, in Sect. 2 we present an example in which small changes in the distribution function produce large changes in the nodes, weights and quadrature approximations, even though the value of the approximated integral does not change much. This motivates our further considerations.

In Sect. 3, we review particular subproblems arising from different methods for computing Gauss–Christoffel quadrature formulas, with the emphasis on the sensitivity of maps from (modified) moments to the nodes and weights of the computed quadrature. For earlier results, refer to [15, p. 252 and Sect. 2], and for recent analysis to [1,23,38]. Despite the vast literature on related subjects, the problem of sensitivity of the Gauss–Christoffel quadrature has, to our knowledge, not been posed or examined in the literature. That problem certainly is of theoretical importance, and it is desirable to investigate its relationship with the subproblems studied in the literature. Section 4 recalls some basics about the error in Gauss–Christoffel quadrature approximations. In Sect. 5 we present discussion and further examples that lead to some understanding of the sensitivity of Gauss–Christoffel quadrature approximations. Section 6 gives a summary and open questions.

Our interest in this problem originated in analysis of the conjugate gradient method for solving linear systems and of the Lanczos method for solving the symmetric eigenvalue problem. The close relationship of these methods of numerical linear algebra to Gauss–Christoffel quadrature of the Riemann–Stieltjes integral has been known since their introduction; see [31, Sect. 14–18], [57, Chap. III]. In particular, the conjugate gradient method generates a sequence of Gauss–Christoffel approximations to the piecewise constant distribution function that has jumps at the eigenvalues of the linear operator equal in magnitude to the squared components of the normalized initial residual along the corresponding eigenfunctions. Moreover, the size of the A-norm of the error at the kth step of the conjugate gradient method has a natural interpretation as the scaled remainder of the kth order Gauss–Christoffel quadrature approximation of the Riemann–Stieltjes integral; see [5] and [40, Sects. 2.2 and 3.3] for a recent review of related results and bibliography. There is also an interesting relationship of the sensitivity of Gauss–Christoffel quadrature to the convergence properties of the conjugate gradient and Lanczos algorithms in finite precision arithmetic. Its detailed investigation is, however, out of the scope of this paper.

All experiments in this paper were performed using MATLAB on a computer with machine precision $\approx 10^{-16}$.

2 Motivating examples

We now present an example of a nondecreasing discontinuous distribution function $\omega(x)$ with finite points of increase, and a perturbation of this function, for which the Gauss–Christoffel quadrature estimates can be quite sensitive. We use a distribution function from [51] with the value between a and the first point of increase zero, and points of increase $\lambda_1 < \cdots < \lambda_n$,

$$\lambda_i = \lambda_1 + \frac{i-1}{n-1}(\lambda_n - \lambda_1)\,\gamma^{n-i}, \quad i = 2, \ldots, n-1,$$

where $0 < a < \lambda_1$, $\lambda_n < b$ and $\gamma \in (0, 1)$ is a properly chosen parameter. The sizes of the individual jumps δ_i, $i = 1, \ldots, n$ are randomly generated using the MATLAB command `rand` and normalized so that

$$\int_a^b d\omega(x) = \sum_{i=1}^{n} \delta_i = 1.$$

We construct the related "perturbed" distribution function $\tilde{\omega}(x)$ to have two points of increase for each single point of increase of $\omega(x)$. Given a positive perturbation parameter ζ, where $\zeta \ll \lambda_1$ and $\zeta \ll \lambda_2 - \lambda_1$, we replace each point of increase λ_i of ω by two close points $\tilde{\lambda}_{2i-1} \equiv \lambda_i - \zeta$ and $\tilde{\lambda}_{2i} \equiv \lambda_i + \zeta$. We proportion the jumps $\tilde{\delta}_{2i-1}$ and $\tilde{\delta}_{2i}$ randomly (again using the MATLAB function `rand`), scaling so that $\tilde{\delta}_{2i-1} + \tilde{\delta}_{2i} = \delta_i$. For a small ζ the distribution functions ω and $\tilde{\omega}$ are close to each other.

We consider a smooth function $f(x) = x^{-1}$ and demonstrate that the difference between the Gauss–Christoffel quadrature estimates of the same degree for I_ω and $I_{\tilde{\omega}}$ can for some values of k become much larger than the difference between the integrals themselves.

In our experiment we take $\lambda_1 = 0.1$, $\lambda_n = 100$, $a = \lambda_1 - 10^{-5}$, $b = \lambda_n + 10^{-5}$, $n = 24$, $\gamma = 0.55$, and $\zeta = 10^{-8}$. The Jacobi matrices containing the recurrence coefficients of the corresponding orthogonal polynomials were computed from the spectral data using the algorithm of Gragg and Harrod implemented in the MATLAB routine `rkpw.m`; see [23,24,29].[1] The Gauss–Christoffel quadrature nodes and weights were computed as the eigenvalues and the squared first components of the corresponding normalized eigenvectors of the Jacobi matrices using the MATLAB routine `gauss.m`, see [23,24].[2]

In this first example, the Jacobi matrices could also be computed via the double-reorthogonalized Lanczos process (conjugate gradient algorithm) applied to the diagonal matrix $A = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ with the starting vector $v_1 = [(\delta_1)^{\frac{1}{2}}, \ldots, (\delta_n)^{\frac{1}{2}}]^{\mathrm{T}}$; see [29,52]. Similarly, one could use the Lanczos process (CG algorithm) on $\tilde{A} = \mathrm{diag}(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{2n})$ and $\tilde{v}_1 = [(\tilde{\delta}_1)^{\frac{1}{2}}, \ldots, (\tilde{\delta}_{2n})^{\frac{1}{2}}]^{\mathrm{T}}$ to compute the perturbed nodes and weights. In this way the close relationship between the Gauss–Christoffel quadrature and the Lanczos process (conjugate gradient method) could be exploited. For small values of n the computational cost is negligible and the cost of reorthogonalization,

[1] Please note that in [23,24] the same implementation is called `lanczos.m`. Since that might cause a confusion with the implementation of the Lanczos process, we use the original name from [29, p. 328].

[2] An interested reader can find all m-files used for generating our figures, including the extended precision implementations, at http://www.cs.cas.cz/mweb, section "Applications".

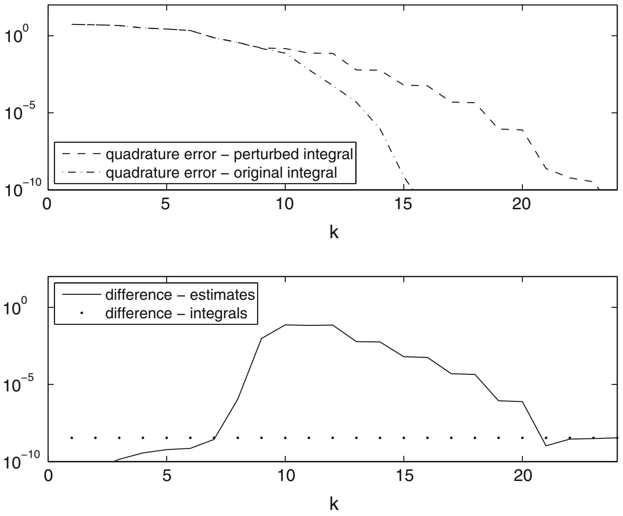


Fig. 1 Sensitivity of the Gauss–Christoffel quadrature for distribution functions with finite points of increase, $\zeta = 10^{-8}$. The top graph shows the error of the Gauss–Christoffel quadrature approximation for $f(x) = x^{-1}$ corresponding to the original stepwise distribution function ω (*dash-dotted line*) and to its perturbation $\tilde{\omega}$ with doubled points of increase (*dashed line*). The bottom graph displays the absolute value of difference in the estimates (*solid line*) and the difference between the approximated integrals (*dots*)

considered in [29, p. 325], does not play a role. In the MATLAB routine `pftoqd.m` we have also implemented the algorithm by Laurie which requires no subtractions; see [36]. We emphasize that the same sensitivity phenomenon can be observed, with differences which are here insignificant, using various computations of the recurrence coefficients from the spectral data.[3]

In the top of Fig. 1 we plot the error of the Gauss–Christoffel quadrature approximations $|E^k_\omega| \equiv |I_\omega - I^k_\omega|$ (dash-dotted line) and $|E^k_{\tilde{\omega}}| \equiv |I_{\tilde{\omega}} - I^k_{\tilde{\omega}}|$ (dashed line), and in the bottom we plot the difference between the Gauss–Christoffel approximations $|I^k_{\tilde{\omega}} - I^k_\omega|$ (solid line) and the difference between the approximated integrals $|\Delta| \equiv |I_\omega - I_{\tilde{\omega}}| \approx 3.443 \times 10^{-9}$ (dots). (Both $I_\omega \approx 5.50658692032301$ and $I_{\tilde{\omega}}$ were computed as finite sums of positive numbers to a relative accuracy close to machine precision). For $k \geq 8$ the Gauss–Christoffel approximations of the integrals $I_{\tilde{\omega}}$ and I_ω start to differ very dramatically, and the size of that difference exceeds 10^{-1} for $k = 10$. After that it is approximately equal to the error $|I_{\tilde{\omega}} - I^k_{\tilde{\omega}}|$ until that quantity drops below the size of the difference between the approximated integrals for $k = 21$.

This dramatic change in the estimates of the integral can be linked to a corresponding sensitivity in the orthogonal polynomials. Though the distribution functions ω and $\tilde{\omega}$ seem very close, the corresponding systems of orthogonal polynomials are quite different. This is illustrated in Fig. 2, which shows the entries of the Jacobi matrices

[3] That has been confirmed independently by Dirk Laurie, who computed, with the data from the motivating example, the Jacobi matrices, nodes and weights of the quadrature to full 16 digits of accuracy using his software package (D. Laurie, Personal communication, October 2006). Other valuable independent experiments were performed by Jarda Kautský (Personal communication, October 2006).

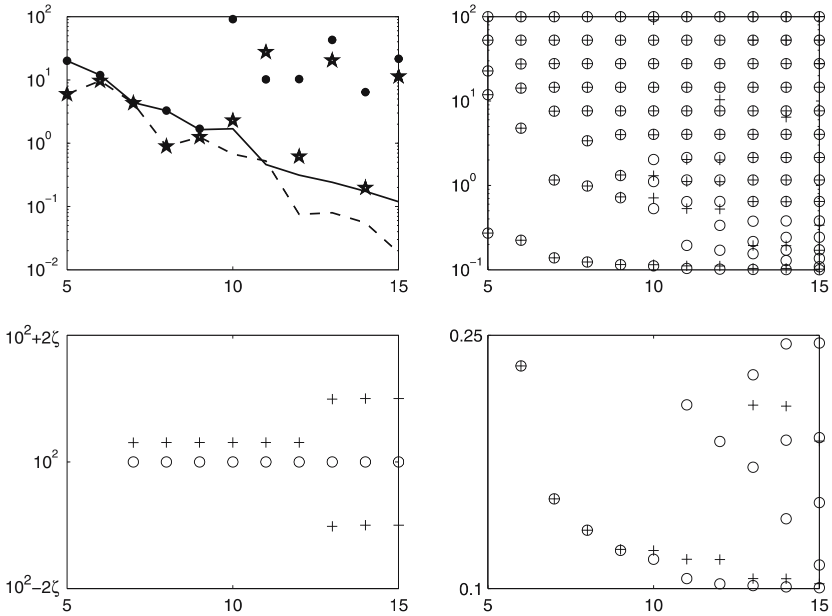


Fig. 2 Top left: Diagonal and off-diagonal entries of the Jacobi matrices corresponding to the distribution functions ω (*solid line and dashed line*, respectively) and $\tilde{\omega}$ (*dots and stars*, respectively). The other plots depict the quadrature nodes corresponding to the distribution function ω (*circles*) and $\tilde{\omega}$ (*pluses*) versus. the number of nodes k in the quadrature. Top right : all nodes. *Bottom left*: nodes near λ_n. *Bottom right*: nodes near λ_1

and the quadrature nodes (the zeros of the corresponding orthogonal polynomials) for ω, $\tilde{\omega}$ for iterations $k = 5, \dots, 15$. In the top left part the diagonal entries of the Jacobi matrices for ω, $\tilde{\omega}$ are plotted by the solid line and by dots, respectively. Similarly, the off-diagonal elements are plotted by the dashed line and by stars, respectively. Up to $k = 7$ the computed Jacobi matrices are very close with their difference close to the square root of the machine precision. For $k = 8, 9$ the difference grows very rapidly (though in the figure the entries are still graphically indistinguishable). The corresponding entries suddenly completely depart at $k = 10$. The same is true for some quadrature nodes. Up to $k = 8$ they are graphically indistinguishable. For $k = 9$ the nodes corresponding to ω (circles) and $\tilde{\omega}$ (plusses) close to λ_1 start to visually differ, and eventually there are many fewer nodes for $\tilde{\omega}$ near λ_1 than there are for ω. The missing nodes for $\tilde{\omega}$ can be found close to λ_n, where they lie in pairs near the nodes for ω. We can see that for $\tilde{\omega}$, $\tilde{\lambda}_{2n-1}$ and $\tilde{\lambda}_{2n}$ are approximated to full accuracy starting from $k = 13$. Results are similar for different values of ζ, providing that $\zeta \ll \lambda_1$, $\zeta \ll \lambda_2 - \lambda_1$.

This first example motivates our investigation. In this paper we ask when, as illustrated in Figs. 1 and 2, Gauss–Christoffel quadrature is sensitive to small perturbations of the distribution function, and under what conditions it is guaranteed to be insensitive. Such conditions exist, which can be verified using the following second example. Construct the perturbed distribution function $\tilde{\omega}(x)$ for $\omega(x)$ given above by placing *a* *single* (positive) $\tilde{\lambda}_i$ randomly in the interval of size 2ζ centered at λ_i, $i = 1, \ldots, n$, with $\tilde{\delta}_i = \delta_i$. (Here we do not specify the position of $\tilde{\lambda}_1$ and $\tilde{\lambda}_n$ relative to the centers of the intervals λ_1 and λ_n; it can be arbitrary.) Then, in contrast to the results shown in Fig. 1, the difference between the Gauss–Christoffel quadrature estimates for $f(x) = x^{-1}$ seems for all k bounded by the size of the difference between the approximated integrals $|I_\omega - I_{\tilde{\omega}}|$, independently of the choice of $0 < \zeta < 0.1$.

We will see that similar phenomena can be observed for continuous and even analytic distribution functions: Gauss–Christoffel quadrature can be highly sensitive to *some* small changes of a given distribution function, and insensitive to others. Next we describe such situations and relate them to theoretical results in the literature.

3 Literature review

As mentioned above, although the question on sensitivity of Gauss–Christoffel quadrature has not, to our knowledge, been addressed in the literature, some related problems have been thoroughly investigated. In this section we summarize what is known about the sensitivity of generating the coefficients of the three-term recurrence satisfied by polynomials orthogonal with respect to the integral (1) and then computing the quadrature nodes and weights from the recurrence coefficients. The richness of the mathematical roots of this field is evidenced in the fact that the same problems have been described independently in many different ways and analyzed using many different techniques in literature that has little cross-reference. It would be very useful to relate in detail all of the existing results, but in this section we give just a brief overview.

3.1 Sensitivity in computation of the recurrence coefficients

Analytic expressions for the recurrence coefficients are explicitly known for some classical distribution functions and the corresponding orthogonal polynomials; see, e.g., [24, p. 217], [38, p. 203], [53]. In practical applications, though, an analytic knowledge of the recurrence coefficients is exceptional, and one has to calculate them. Gautschi [24] presented four techniques. Using our terminology these are:

T1. A modified Chebyshev algorithm.
T2. Discretization of the distribution function.
T3. Computation of the recurrence coefficients for the discrete Riemann–Stieltjes integral.
T4. Computation of the recurrence coefficients for one distribution function from known coefficients for another distribution function.

The technique **T4** is not generally applicable, restricted to the case in which the original distribution function is multiplied by a rational nonnegative function [24, Sect. 2.5], [23, Sect. 2.4]. The problem of changes in orthogonal polynomials with respect to certain classes of modifications to the distribution function has been studied in many papers; see, e.g., [33,56], and [23, Sect. 2.4]. For a description of an old general result attributed to Markov concerning the dependence of the zeros of orthogonal polynomials on the parameter in the distribution function we refer to [55, Sect. 6.12, pp. 111–112]. Though such results are somewhat related to the problem of sensitivity of the Gauss–Christoffel quadrature, they are either of restricted applicability or merely qualitative. They do not lead to a general perturbation theory.

The modified Chebyshev algorithm **T1** represents an example of a more general approach based on knowledge of the recursion coefficients for some classical orthogonal polynomials determined by an auxiliary distribution function [24, Sect. 2.2]. Assume that the modified moments of the chosen (auxiliary) orthogonal polynomials with respect to the original distribution function can be determined *accurately*. From these moments and the *known* recurrence coefficients of the auxiliary polynomials, the modified Chebyshev algorithm determines the unknown recurrence coefficients of the desired orthogonal polynomials. The difficulties are the possibly large computational cost (not important in the context of our paper) and the possible inaccuracy in the computed results. The last difficulty has been thoroughly studied by Gautschi; see [15–17,19], and Sect. 2.1 of the book [23]. Subsect. 2.1.3 defines the following maps:

$\mathbf{K}_k$: the map from the modified moments to the recurrence coefficients;
$\mathbf{G}_k$: the map from the modified moments to the nodes and weights of the computed quadrature;
$\mathbf{H}_k$: the map from the nodes and weights of the computed quadrature to the recurrence coefficients.

Then $\mathbf{K}_k$ can be represented as a composition of the other two maps,

$$\mathbf{K}_k = \mathbf{H}_k \circ \mathbf{G}_k.$$

The *condition numbers* attributed to $\mathbf{G}_k$ and $\mathbf{K}_k$ were studied in [23, Sects. 2.1.4, 2.1.5 and 2.1.6, pp. 59–75]. If monomials are used as the auxiliary polynomials, the modified moments reduce to ordinary moments and the maps $\mathbf{G}_k$ and $\mathbf{K}_k$ are notoriously ill-conditioned. Even for a good choice of the auxiliary polynomials (such as the Chebyshev polynomials) and modified moments the situation is not simple. There are distribution functions for which the condition numbers are small, but there are other distribution functions for which the condition numbers grow exponentially with the number of nodes k. Moreover, the assumption that the modified moments are known accurately is difficult to satisfy; see [38, Sect. 3.2].

The general question of how to choose the auxiliary distribution function was analyzed by Beckermann and Bourreau in the remarkable paper [1]. They showed, among other results, that if the original and auxiliary distribution functions have different supports, i.e., the sets of all points of their increase, [23, p. 3], then the condition number of $\mathbf{K}_k$ grows exponentially with k; see [1, Theorem 11, p. 93]. The authors further conjectured on the same page that the condition number of $\mathbf{K}_k$ is linked with the condition numbers of the matrices of modified and mixed moments.

The map $\mathbf{H}_k$ is said to be generally well-conditioned in [23, p. 59], though *numerically stable computation* of the entries of the Jacobi matrix from the quadrature nodes and weights is not easy; see [23, Sect. 3.1.1, pp. 154–155, Sect. 3.5, pp. 253–254 and Notes to Sect. 1.3, p. 50] with references to [36, Theorem on p. 168], [37,38] and

[1, Theorem 1 and Corollary 8]. (See also the last two paragraphs of this section, which explain in detail history of that problem and the fact that the algorithmic construction of Laurie in [36] also gives the perturbation result.) Further results on the condition numbers of the map $\mathbf{H}_k$ (and also of its inverse $\mathbf{H}_k^{-1}$) can be found in [1, relation (7), Sect. 2 and Appendix], see also [14, Sect. 4, pp. 190–193]. (We address the map $\mathbf{H}_k^{-1}$, in particular sensitivity of the nodes and weights and their computation from the entries of the Jacobi matrix, in Sect. 3.2). The approach from [14] is based on a remarkable result by Nevai on modification of the recurrence coefficients when adding a single point of increase to the given distribution function; see [43, Sect. 7, Lemma 15, p. 131], [14, Sect. 3, Lemma 1, p. 187]. For an instructive algebraic description and application of the same idea we refer to [12,13]. It is interesting that *essentially the same problem* of sensitivity of the entries in the Jacobi matrix to small perturbations of the nodes and weights of the corresponding distribution function (i.e., the eigenvalues and the first components of the normalized eigenvectors respectively) has recently been studied in a different way (independently of the results mentioned above) in [11]; see also [42] and the earlier paper [59]. A related more general problem of sensitivity of the Lanczos reduction has been thoroughly investigated in [45], see also [4,34].

The maps $\mathbf{K}_k$, $\mathbf{H}_k$ and $\mathbf{G}_k$ are interesting to study. However, as we will see in Sect. 5, they do not represent a relevant tool for investigation of sensitivity of Gauss–Christoffel quadrature. Their detailed discussion has been included here in order to explain the differences between the sensitivity problems studied previously and the sensitivity question posed and investigated in this paper.

The techniques **T2** and **T3** couple into one approach. The basic idea behind the discretization methods (see [24, Sect. 2.4], [23, Sect. 2.2, p. 90]) is an approximation of the given distribution function by a suitable *discrete* distribution function, computation of the recurrence coefficients for the discrete distribution function, and approximation of the desired recurrence coefficients by the computed (discrete) ones. Gautschi identified in [23, p. 90] two important issues which must be considered: the appropriate choice of discretizations and convergence of the discrete orthogonal polynomials (recurrence coefficients) to the desired ones. Both issues are tightly related. In the simple case when the original distribution function is composed of several components for which the analytic formulas for the orthogonal polynomials (Legendre, Chebyshev, ...) are known, discretization by a suitable combination of the N-point Gauss-type quadratures (Gauss–Legendre, Gauss–Chebyshev, ...) for a sufficiently large $N \gg k$ gives the result. Assuming exact arithmetic, the first $N - 1$ polynomials orthogonal with respect to the original distribution function are then also orthogonal with respect to the discretized distribution function, and the desired recurrence coefficients are determined *accurately*; see [24, p. 222]. Practical cases can be much more complicated, and finding an appropriate discretization is a rather involved procedure [23, Sect. 2.2.4].

There is one additional very important issue not mentioned in [23,24]. Convergence $N \to \infty$ describes the limiting case. In order to evaluate the accuracy of the methods based on discretization, one must be able to estimate the discretization error for a *finite* N. In other words, one must investigate *how fast* the discrete orthogonal polynomials converge to the desired ones, or, in a more complex way, *sensitivity* of the Gauss–Christoffel quadrature under small perturbations of the original distribution function.

It seems that sensitivity is indeed a fundamental issue which cannot be omitted from consideration. If the Gauss–Christoffel quadrature is sensitive to small perturbations of the distribution function, then the computation based on the discretization may in general fail even if the discrete orthogonal polynomials, and, subsequently, the nodes and weights of the discrete quadrature, are determined accurately. A particular discretization procedure is not justified without proving that the results of the Gauss–Christoffel quadrature are insensitive with respect to the perturbation of the original distribution function represented by its discretization.

Finally, we discuss *computation* of the recurrence coefficients for the *discrete* Riemann–Stieltjes integral. This is an inverse problem: given nodes and weights of the N-point discrete Gauss–Christoffel quadrature formula, compute the entries of the corresponding Jacobi matrix.[4] In order to find the approximation to the desired k-point Gauss–Christoffel quadrature, we actually do not need the whole N by N Jacobi matrix, so we stop when we obtain its k by k left principal submatrix, $k \ll N$. In the classical language of orthogonal polynomials, the problem is solved by the discrete Stieltjes process [23, Sect. 2.2.3.1, p. 95]. In the language of numerical linear algebra, the Stieltjes process (implemented with modified Gram–Schmidt orthogonalization and normalization of the orthogonal polynomials) is equivalent to the Lanczos algorithm (see, e.g., [29, p. 322]), which is numerically unstable. This fact has been noted in the orthogonal polynomial literature (see, e.g., [19,20], [14, Sect. 2]), and reorthogonalization has been rejected as too costly [29, p. 325]. When k is small, however, the cost of reorthogonalization is negligible. Moreover, the analysis of the Lanczos algorithm behavior in finite precision arithmetic by Paige, Parlett, Scott, Simon, Greenbaum and others (reviewed, for example, in [40]) is almost unknown in the literature of orthogonal polynomial community, despite some notable work [2,3,22,26,28,33] which emphasizes the interplay between the classical polynomial and vector algebraic formulations. The analysis can supply, at least, very convincing examples for illustrating and testing numerical instabilities.

In order to overcome the numerical instability of the Lanczos algorithm, Gragg and Harrod suggested in their beautiful paper [29] a new algorithm based on ideas of Rutishauser. For an interesting experimental comparison, see [48, Sect. 2]. An alternative approach, based on the above mentioned results of Nevai [43], along with an experimental comparison, can be found in [14]. From numerical results Gragg and Harrod spotted a curious phenomenon: close nodes and weights can give two very different $k \times k$ Jacobi matrices. They concluded that the problem of reconstructing a Jacobi matrix from the weights and nodes is ill-conditioned [29, p. 330 and 332]. This conclusion has been examined by Laurie [36], who pointed out that the negative statement is linked to the use of the max-norm for vectors. He suggested instead measuring the perturbation of the weights in the componentwise relative sense [36, p. 179], [38, Sect. 6]. The main part of [36] is devoted to the constructive proof of the following statement [36, Theorem on p. 168]: given the weights and the $N-1$ positive differences between the consecutive nodes, the main diagonal entries of the corresponding Jacobi matrix (shifted by the smallest node) and the off-diagonal entries can be computed in

[4] Note that the inverse problem corresponds in the literature to the map $\mathbf{H}_k$, not to $\mathbf{H}_k^{-1}$.

$\frac{9}{2}N^2 + O(N)$ arithmetic operations, all of which can involve only addition, multiplication and division of positive numbers. Consequently, in finite precision arithmetic they can be computed to a *relative accuracy* no worse than $\frac{9}{2}N^2\varepsilon + O(N\varepsilon)$, where ε denotes machine precision. This result bounds also the conditioning of the problem. If the weights and the $N - 1$ positive differences between the consecutive nodes are perturbed, with the size of the relative perturbations of the individual entries bounded by some small ϵ, then such perturbation can cause a relative change of the individual entries of the shifted main diagonal and of the individual off-diagonal entries of the Jacobi matrix not larger than $\frac{9}{2}N^2\epsilon + O(N\epsilon)$. The resulting algorithm combines ideas from earlier works from approximation theory, orthogonal polynomials, and numerical linear algebra.

3.2 Sensitivity and computing of the quadrature nodes and weights

Computing the quadrature nodes and weights is of great interest on its own. If the recurrence coefficients are used to construct a symmetric tridiagonal matrix with positive subdiagonals (Jacobi matrix), then, as mentioned above, the quadrature nodes are the eigenvalues and the weights are the first components of the normalized eigenvectors; see, e.g., [23, Sect. 3.1.1.1, pp. 152–154; Sect. 3.5, pp. 253–254]. In some special cases such as Gauss–Legendre quadrature, it is useful to consider also different ways of computing the quadrature nodes and weights; see [54]. It should be noted, however, that the comparison given in [54] does not refer to the recent developments in eigensolvers for Jacobi matrices recalled below. In most cases, computing the quadrature nodes and weights reduces to computing eigenvalues and the first components of eigenvectors of Jacobi matrices.

It is well known that two Jacobi matrices that are close to each other also have close eigenvalues and eigenvectors in the *absolute sense*, where the closeness is measured by the absolute values of the differences between the corresponding individual eigenvalues and the corresponding individual eigenvectors; see, e.g., [27, Chap. 8], [33, p. 454], [11, p. 104], and [1, relation (7) and Appendix] mentioned above. For eigenvectors, the proportionality constant depends on the relative gaps between the eigenvalues of the unperturbed matrix. However, two close Jacobi matrices do not necessarily have eigenvalues that are close in a *relative* sense. A small perturbation of the entries of the Jacobi matrix can cause a large *relative* change in the eigenvalues and the eigenvector entries; see [8, pp. 71–72] and [39]. It is worth noting that Kahan has shown that small relative changes in the entries of the Cholesky factors of a *positive definite Jacobi matrix* do cause small relative changes in the eigenvalues of the Jacobi matrix [46, p. 123]; see also [7]. The thesis [8] gives also a comparison of different numerically stable algorithms for computing eigenvalues and eigenvectors of Jacobi matrices; see also the survey and comments in [38, Sect. 2], and the recent work [9,10,30,58]. We can conclude that the computation and perturbation theory of quadrature nodes and weights from the recurrence coefficients is well understood. The main difficulty in perturbation analysis and in computation of the Gauss–Christoffel quadrature lies in generating the recurrence coefficients.

Given this vast literature and our motivating example from the previous section, we focus our attention on the sensitivity of the quadrature formulas to changes in the distribution function.

3.3 Application to motivating examples: when larger support matters

The main difference between the first example at the beginning and the second example at the end of Sect. 2 consists in whether or not the number of points of increase (i.e., the 'size' of the support) is changed when ω is perturbed to form $\tilde{\omega}$. We will show that if there is no change in the number of points of increase, then a result by Laurie [36] explains the observed insensitivity of Gauss-quadrature for small enough perturbations.

Suppose we perturb the (discrete) ω of Sect. 2, resulting in $\tilde{\omega}$ with the *same number* of points of increase. Then, by Laurie's result, the corresponding shifted Jacobi matrices are close to each other in the componentwise *relative* sense. Using a classical perturbation result for eigensystems of symmetric matrices, the resulting Gauss–Christoffel quadrature nodes and weights for ω and $\tilde{\omega}$ must also be close to each other, with individual differences proportional to the perturbation parameter ζ. Consequently, for the (smooth and monotonic) function $f(x) = x^{-1}$ with ζ sufficiently small, the difference between the quadrature estimates must be proportional to the difference between the approximated integrals $|I_\omega - I_{\tilde{\omega}}|$.

There are two limitations of this argument. First, it does not apply to the first motivating example, since Laurie's result cannot be applied when the number of points of increase changes. Second, it does not apply to the second motivating example either, since the value of $\zeta = 10^{-8}$ was chosen too large. It does, however, provide *quantitative* sensitivity results for smaller ζ, or when ω and $\tilde{\omega}$ coincide at all points of increase, except for, say, λ_n, which is well separated from $\lambda_1, \ldots, \lambda_{n-1}$. We next prove a result that does predict the difference in behavior of our two examples.

4 Quadrature differences in terms of approximation error

We present a slight generalization of a result found in the classic textbook of Isaacson and Keller [32] in Theorem 3 (p. 329) and in the second line of the identity (6) on p. 334.

The standard approach to Gauss quadrature of the Riemann integral and to Gauss–Christoffel quadrature of the Riemann–Stieltjes integral is based on Hermite interpolation and is attributed to Markov; see, e.g., [18, p. 82]. Here we take advantage of results based on Lagrange interpolation. This allows us to retain k free parameters in the remainder term for the kth order quadrature, which will later prove convenient in evaluation of the quadrature differences. In our exposition we follow the presentation of Gauss–Christoffel quadrature given by Lanczos in [35, Chap. VI, Sect. 10], cf. also [21, Theorem 3.2.1].

Choose k distinct points $x_1, \ldots, x_k$ inside the interval $[a, b]$, and let $q_k(x) = (x - x_1) \ldots (x - x_k)$. Then the Lagrange polynomial interpolating $f(x)$ at the points

$x_1, \dots, x_k$ can be written as

$$\mathcal{L}_k(x) = \sum_{j=1}^{k} f(x_j) \frac{q_k(x)}{q_k'(x_j)(x - x_j)},$$

and

$$f(x) = \mathcal{L}_k(x) + q_k(x) f[x_1, \dots, x_k, x],$$

where $f[x_1, \dots, x_k, x]$ is the kth divided difference of f with respect to $x_1, \dots, x_k, x$; see e.g., [32, Sect. 6.1]. We can derive a corresponding interpolatory quadrature formula

$$\int_a^b f(x)\, d\omega(x) = \sum_{j=1}^{k} \vartheta_j f(x_j) + \int_a^b q_k(x) f[x_1, \dots, x_k, x]\, d\omega(x), \tag{3}$$

where the last term represents the error and

$$\vartheta_j = \frac{1}{q_k'(x_j)} \int_a^b \frac{q_k(x)}{(x - x_j)}\, d\omega(x), \quad j = 1, \dots, k. \tag{4}$$

Up to now $x_1, \dots, x_k$ were arbitrary distinct nodes inside $[a, b]$. The beauty of the Gauss–Christoffel quadrature is in setting the interpolatory nodes equal to the roots of the kth orthogonal polynomial corresponding to $\omega(x)$. Then we can consider k additional distinct nodes inside $[a, b]$ *which we need not even know* and show that the interpolatory quadrature on k nodes is as accurate as if $2k$ nodes had been used. This elegant consequence is summarized in the following theorem.

Theorem 1 *Consider a nondecreasing function $\omega(x)$ on a finite interval $[a, b]$. Let $p_k(x) = (x - t_1) \dots (x - t_k)$ be the kth monic orthogonal polynomial with respect to the inner product defined by the Riemann–Stieltjes integral on the interval $[a, b]$ with the distribution function $\omega(x)$. Choose k arbitrary distinct points $\mu_1, \dots, \mu_k$ in $[a, b]$. Let*

$$I_\omega = \int_a^b f(x)\, d\omega(x), \tag{5}$$

where f'' is continuous on $[a, b]$, and let I_ω^k be the approximation to I_ω obtained from the k-point Gauss–Christoffel quadrature rule. Then for $m = 1, \dots, k$, the error of *this approximation is given by*

$$E_\omega^k(f) \equiv I_\omega - I_\omega^k = \int_a^b p_k(x) f[t_1, \dots, t_k, x]\, d\omega(x) \tag{6}$$

$$= \int_a^b p_k(x)(x-\mu_1)\dots(x-\mu_m) f[t_1, \dots, t_k, \mu_1, \dots, \mu_m, x]\, d\omega(x), \tag{7}$$

where $f[t_1, \dots, t_k, \mu_1, \dots, \mu_m, x]$ is the $(k+m)$th divided difference of the function $f(x)$ with respect to the nodes $t_1, \dots, t_k, \mu_1, \dots, \mu_m, x$.

Proof Assume, for the moment, that the nodes $\mu_1, \dots, \mu_k$ are distinct from the nodes $t_1, \dots, t_k$. If we derive the quadrature rule (3), (4) using $t_1, \dots, t_k$, then we have

$$\int_a^b f(x) d\omega(x) = \sum_{j=1}^k \vartheta_j f(t_j) + \int_a^b p_k(x) f[t_1, \dots, t_k, x]\, d\omega(x),$$

where the continuity of f' guarantees the finiteness of the divided difference as x varies. If $f(x)$ is a polynomial of degree at most $2k-1$, then $f[t_1, \dots, t_k, x]$ is a polynomial in x of degree at most $k-1$ and the rule is exact, since the orthogonality of $p_k(x)$ to all such polynomials makes the error term equal to zero. Consequently, the resulting interpolatory quadrature represents the Gauss–Christoffel quadrature. If we derive a quadrature rule using the points t_j plus the new nodes μ_i, then for $m = 1, \dots, k$,

$$\begin{aligned}\int_a^b f(x)\, d\omega(x) &= \sum_{j=1}^k \hat{\vartheta} f(t_j) + \sum_{i=1}^m \hat{\xi}_i f(\mu_i) \\ &+ \int_a^b p_k(x)\,(x-\mu_1)\dots(x-\mu_m) f[t_1, \dots, t_k, \mu_1, \dots, \mu_m, x]\, d\omega(x).\end{aligned}$$

We observe from (4) that for $i = 1, \dots, m$ the weight $\hat{\xi}_i$ of each additional node is proportional to

$$\int_a^b p_k(x) r_i(x)\, d\omega(x) = 0,$$

where $r_i(x) = (x-\mu_1)\dots(x-\mu_m)/(x-\mu_i)$ is a polynomial of degree at most $k-1$, and therefore the orthogonality of $p_k(x)$ to all such polynomials results in a zero weight. Consequently, the contribution of the additional nodes $\mu_1, \dots, \mu_m$ to the integration formula vanishes, i.e.,

$$\sum_{i=1}^{m} \hat{\hat{\xi}}_i f(\mu_i) = 0.$$

It follows from uniqueness of the Gauss–Christoffel quadrature rules that $\vartheta_j = \hat{\vartheta}_j$ and the statement is proved.

If some μ_i is equal to some t_j, then replacing the Lagrange interpolant by the Hermite interpolant (cf. [41, p. 175], [32, p. 330]), and using the continuity of f'' finishes the proof in an analogous way. □

For analytic functions $f(x)$ it is possible to express the error of the Gauss–Christoffel quadrature rule without using derivatives or divided differences. Letting $p_k(x)$ be as above, the function

$$\rho_k(z) = \int_a^b \frac{p_k(x)}{z - x}\, d\omega(x)$$

is analytic in the complex plane outside the interval $[a, b]$. Suppose that $f(z)$ is analytic in a simply connected domain containing $[a, b]$ in its interior, and let Γ be a simple closed positively oriented curve in that domain encircling $[a, b]$. Then

$$E_\omega^k(f) = \frac{1}{2\pi\sqrt{-1}} \int_\Gamma K_k(z) f(z)\, dz, \quad K_k(z) = \frac{\rho_k(z)}{p_k(z)}; \tag{8}$$

see [23, Theorem 2.48], [6, p. 303, relation (4.6.18)]. This identity has been applied to estimate the error and to study its decrease with k for some particular classes of distribution functions $\omega(x)$ [18,23,25], [6, Sect. 4.6]. The kernel $K_k(z)$ depends through $p_k(z)$ and $\rho_k(z)$ on the given distribution function $\omega(x)$. The question of sensitivity of $E_\omega^k(f)$ with respect to perturbations of the distribution function $\omega(x)$ is thus reduced to the question of sensitivity of $K_k(z)$, where z lies on a properly chosen curve Γ in the complex plane, with respect to small perturbations of $\omega(x)$.

An application of Theorem 1 gives the following important result, an expression for the difference between the Gauss–Christoffel quadrature approximations.

Theorem 2 *Let $p_k(x) = (x - x_1) \dots (x - x_k)$ be the kth orthogonal polynomial with respect to $d\omega$ on $[a, b]$, and let $\tilde{p}_k(x) = (x - \tilde{x}_1) \dots (x - \tilde{x}_k)$ be the kth orthogonal polynomial with respect to $d\tilde{\omega}$. Denote by $\hat{p}_s(x) = (x - \xi_1) \dots (x - \xi_s)$ the least common multiple of the polynomials $p_k(x)$ and $\tilde{p}_k(x)$. If f'' is continuous on $[a, b]$,* *then the difference between the approximation I_ω^k to I_ω and the approximation $I_{\tilde\omega}^k$ to $I_{\tilde\omega}$, obtained from the k-point Gauss–Christoffel quadrature rule, is bounded as*

$$|I_\omega^k - I_{\tilde\omega}^k| \le \left| \int_a^b \hat{p}_s(x) f[\xi_1,\dots,\xi_s,x]\, d\omega(x) - \int_a^b \hat{p}_s(x) f[\xi_1,\dots,\xi_s,x] d\tilde\omega(x) \right| + \left| \int_a^b f(x) d\omega(x) - \int_a^b f(x) d\tilde\omega(x) \right|. \tag{9}$$

Proof Consider the difference between the two Gauss quadrature approximations:

$$I_\omega^k - I_{\tilde\omega}^k = I_\omega - E_\omega^k - (I_{\tilde\omega} - E_{\tilde\omega}^k) = (E_{\tilde\omega}^k - E_\omega^k) + (I_\omega - I_{\tilde\omega}). \tag{10}$$

Let the polynomials $p_k(x)$ and $\tilde{p}_k(x)$ have $k - m$ common zeros, numbered so that $x_{m+1} = \tilde{x}_{m+1}, \dots, x_k = \tilde{x}_k$. Let $s = k + m$ and use the last equality in Theorem 1 twice. For E_ω^k set the points $t_1, \dots, t_k$ in the theorem to be the zeros of $p_k(x)$, and set the points $\mu_1, \dots, \mu_m$ to be the first m zeros $\tilde{x}_1, \dots, \tilde{x}_m$ of $\tilde{p}_k(x)$. For $E_{\tilde\omega}^k$, set the points $t_1, \dots, t_k$ to be the zeros of $\tilde{p}_k(x)$, and set the points $\mu_1, \dots, \mu_m$ to be the first m zeros $x_1, \dots, x_m$ of $p_k(x)$. The statement will immediately follow. □

Note that from (10) the difference between the Gauss–Christoffel quadrature approximations is of order of the difference between the integrals (*or smaller*) if and only if the first term in the bound (9) is of order of the second term or smaller. Please note that the integrands in the first term in the bound (9) are *identical*. This simplifies the situation in comparison with a possible use of the standard quadrature error formulas known from the literature, from which it seems very difficult to get insight into the sensitivity phenomenon.

We state an analogous result for the weighted Riemann integral with nonnegative weight function that is (for simplicity) continuous on the finite interval $[a, b]$. The continuity assumption is not essential but simplifies the exposition.

Corollary 1 *Let $w(x)$ and $\tilde{w}(x)$ be nonnegative and continuous functions on the finite interval $[a, b]$; let*

$$\omega(x) = \int_a^x w(t)\, dt, \quad \tilde\omega(x) = \int_a^x \tilde{w}(t)\, dt, \quad x \in [a, b]$$

be the corresponding distribution functions. Then the integrals I_ω and $I_{\tilde\omega}$ in (2) represent the weighted Riemann integrals. Using the notation and assumptions of Theorem 2,

$$|I_\omega^k - I_{\tilde{\omega}}^k| \leq \left| \int_a^b \hat{p}_s(x) f[\xi_1, \dots, \xi_s, x](w(x) - \tilde{w}(x))\, dx \right| + \left| \int_a^b f(x)(w(x) - \tilde{w}(x))\, dx \right| . \tag{11}$$

Proof The statement follows immediately as a special case of Theorem 2. □

If $f(x)$ is analytic, we can get identities which do not contain divided differences. Using the kernel expression of the error (8),

$$\begin{aligned} |I_\omega^k - I_{\tilde{\omega}}^k| &\leq \frac{1}{2\pi} \left| \int_\Gamma (K_k(z) - \tilde{K}_k(z)) f(z) dz \right| + |I_\omega - I_{\tilde{\omega}}| \\ &= \frac{1}{2\pi} \left| \int_\Gamma \frac{\rho_k(z)\tilde{p}_k(z) - \tilde{\rho}_k(z) p_k(z)}{p_k(z)\tilde{p}_k(z)} f(z)\, dz \right| + |I_\omega - I_{\tilde{\omega}}| . \end{aligned} \tag{12}$$

5 Discussion and numerical illustrations

The previous section presents simple bounds for the size of the difference $|I_\omega^k - I_{\tilde{\omega}}^k|$ between the results of the k-point Gauss–Christoffel quadrature which immediately follow from the identity (10) and the quadrature error formulas. As shown in Theorem 2, the crucial first term on the right hand side of (10) represents a difference between two integrals with the same integrand $\hat{p}_s(x) f[\xi_1, \dots, \xi_s, x]$ and different distribution functions ω and $\tilde{\omega}$. We will explain why for some distribution functions ω and nearby $\tilde{\omega}$, with f sufficiently smooth and uncorrelated with the difference $\omega - \tilde{\omega}$, this term must inevitably become large, while for slightly different nearby distribution functions the term remains small. We will start with a closer look at our motivating examples from Sect. 2.

5.1 Discrete distribution functions: motivating example revisited

First, for clarity of exposition, we simplify the motivating examples from Sect. 2: in both examples keep the first $n-1$ points of increase of $\tilde{\omega}(x)$ equal to $\lambda_1, \dots, \lambda_{n-1}$, with the corresponding weights $\delta_1, \dots, \delta_{n-1}$. Thus, in both examples, $\tilde{\omega}(x)$ differs from $\omega(x)$ only near λ_n. In the first example, λ_n is replaced by *two* points of increase $\tilde{\lambda}_n = \lambda_n - \zeta$ and $\tilde{\lambda}_{n+1} = \lambda_n + \zeta$, with the (positive) weights $\tilde{\delta}_n$, respectively $\tilde{\delta}_{n+1}$, $\tilde{\delta}_n + \tilde{\delta}_{n+1} = \delta_n$. In the second example λ_n is perturbed to $\tilde{\lambda}_n = \lambda_n + \zeta$ with $\tilde{\delta}_n = \delta_n$.

For $f(x) = x^{-1}$ we get $f[\xi_1, \dots, \xi_s, x] = (-1)^s (x\, \xi_1 \dots \xi_s)^{-1}$, which holds, by induction, for any $s \leq 2k$. Therefore the integrand in the first part of the bound (9) for the k-point quadrature simplifies to

$$g^k(x) \equiv \hat{p}_s(x) f[\xi_1, \ldots, \xi_s, x] = \frac{\hat{p}_s(x)}{x\, \hat{p}_s(0)} = f(x)\, \frac{\hat{p}_s(x)}{\hat{p}_s(0)}, \tag{13}$$

where the last term represents a polynomial having value one at zero. Using (7) we find that

$$E^k_{\tilde{\omega}} - E^k_{\omega} \equiv h^k$$

where

$$h^k \equiv \tilde{\delta}_n\, (g^k(\tilde{\lambda}_n) - g^k(\lambda_n)) \; + \tilde{\delta}_{n+1}\, (g^k(\tilde{\lambda}_{n+1}) - g^k(\lambda_n)).$$

Therefore, using (10) we have

$$I^k_{\omega} - I^k_{\tilde{\omega}} = h^k \; + \Delta, \quad \Delta = I_{\omega} - I_{\tilde{\omega}}.$$

In the second example, the second term in h^k is nonexistent.

In the first example, $E^k_{\tilde{\omega}} - E^k_{\omega} = h^k$ corresponds to the replacement of the *single* λ_n by *two* nearby points $\tilde{\lambda}_n$ and $\tilde{\lambda}_{n+1}$. With the given distribution functions ω and $\tilde{\omega}$ and for some values of $k \ll n$, the term h^k becomes much larger in magnitude than $|\Delta|$.

For small k, the Gauss–Christoffel quadrature approximation $I^k_{\tilde{\omega}}$ does not recognize $\tilde{\lambda}_n$ and $\tilde{\lambda}_{n+1}$ as two distinct points, and h^k is small. For larger k, λ_n becomes closely approximated by the largest node from the Gauss–Christoffel quadrature approximation I^k_{ω} of I_{ω}, and $g^k(\lambda_n)$ becomes very small. At the same time, $\tilde{\lambda}_n$ and $\tilde{\lambda}_{n+1}$ are approximated by a *single* quadrature node from $I^k_{\tilde{\omega}}$, placed in between them. Then $g^k(x)$ has in between $\tilde{\lambda}_n$ and $\tilde{\lambda}_{n+1}$ *two* roots, with one of them very close to λ_n. As k grows, this will soon become not enough to keep h^k small, since $g^k(\tilde{\lambda}_n)$ and $g^k(\tilde{\lambda}_{n+1})$ will grow in magnitude and are of the same sign, while $g^k(\lambda_n)$ is small due to the closeness of the quadrature node for I^k_{ω} to λ_n. Consequently, the differences $g^k(\tilde{\lambda}_n) - g^k(\lambda_n)$ and $g^k(\tilde{\lambda}_{n+1}) - g^k(\lambda_n)$ will also grow in magnitude and are of the same sign. Inevitably, for some value of k, $I^k_{\tilde{\omega}}$ has to place a *second* node, so that both $\tilde{\lambda}_n$ and $\tilde{\lambda}_{n+1}$ are sufficiently closely approximated and the size of the term h^k is kept under control. For that k, then, compared to the quadrature formula for I^k_{ω}, the quadrature formula for $I^k_{\tilde{\omega}}$ has one fewer node in some other part of the interval of integration. Therefore $|I^k_{\tilde{\omega}} - I^k_{\omega}|$ will suddenly become large. The missing node appears in the $(k+1)$st step of the Gauss–Christoffel quadrature approximation. Therefore, from then on, although $|I^k_{\tilde{\omega}} - I^k_{\omega}|$ may not be small, the difference shifted by one step, i.e., $|I^{k+1}_{\tilde{\omega}} - I^k_{\omega}|$, is small.

The situation is illustrated in Fig. 3. In the top part the quadrature errors E^k_{ω} and $E^k_{\tilde{\omega}}$ are plotted by the solid and dashed line, respectively. They cannot be visually distinguished until $k = 9$. Starting from $k = 11$, the convergence of $I^k_{\tilde{\omega}}$ is *delayed*

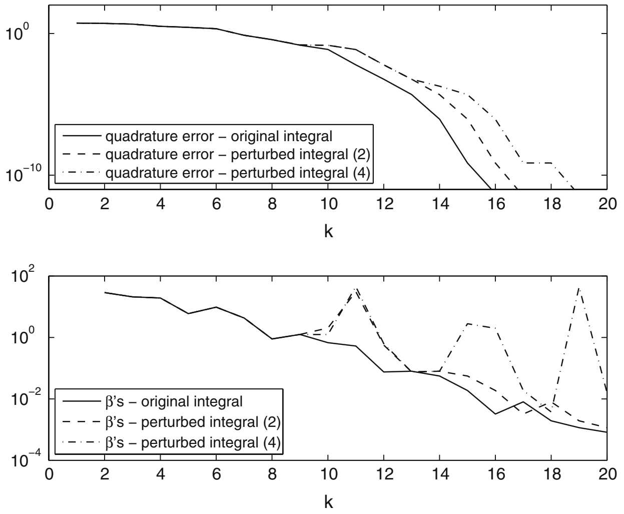


Fig. 3 Sensitivity of the Gauss–Christoffel quadrature for distribution functions with finite points of increase which differ only near λ_n, $\zeta = 10^{-8}$. The *top graph* shows the error of the Gauss–Christoffel quadrature approximation for $f(x) = x^{-1}$ corresponding to the original stepwise distribution function ω (*solid line*), to its perturbation $\tilde{\omega}$ with two points of increase near λ_n (*dashed line*), and to its perturbation $\tilde{\omega}$ with four points of increase near λ_n (*dash-dotted line*). The *bottom graph* displays the off-diagonal entries of the corresponding Jacobi matrices

by one step in comparison to I^k_ω. Entries of the corresponding Jacobi matrices behave in an interesting way, which is illustrated by plotting the off-diagonal entries in the bottom part of the figure, with the solid line corresponding to I^k_ω and the dashed line to $I^k_{\tilde{\omega}}$. Until $k = 9$ the lines coincide. For $k = 10, 11$ the corresponding entries separate, and, starting from $k = 12$, the dashed line is just delayed (shifted to the right) by one step.

The dash-dotted lines in both parts of Fig. 3 correspond to an additional example where λ_n is replaced by four close points $\tilde{\lambda}_n = \lambda_n - \zeta$, $\tilde{\lambda}_{n+1} = \lambda_n - \zeta/3$, $\tilde{\lambda}_{n+2} = \lambda_n + \zeta/3$, $\tilde{\lambda}_{n+3} = \lambda_n + \zeta$, while the new points share the original weight δ_n. The situation is fully analogous. Starting from $k = 14$ and $k = 18$, the convergence of $I^k_{\tilde{\omega}}$ is delayed by two and three steps, respectively.

The behavior of $g^k(x)$ is illustrated in Fig. 4. For clarity we plot $\operatorname{sign}(g^k(x)) \log_{10}(1 + |g^k(x)|)$. The left part plots the behavior in the whole interval of integration for $k = 10$. The right part plots the behavior near λ_n for $k = 9, 11, 12$. For $k = 9$ the line is close to the horizontal axis. For $k = 11$ we can observe the increasing gradient of $g^k(x)$ ($\zeta = 10^{-8}$), and for $k = 12$ both $\tilde{\lambda}_n$ and $\tilde{\lambda}_{n+1}$ are closely approximated by quadrature nodes of $I^k_{\tilde{\omega}}$.

This phenomenon is closely related to the fact that the presence of close eigenvalues affects the rate of convergence of the conjugate gradient method; see the beautiful explanation given by van der Sluis and van der Vorst [49,50]. Similarly, it is closely related to the convergence of the Rayleigh quotient in the power method and to the so-called ‘misconvergence phenomenon’ in the Lanczos method see [44,47]. In exact

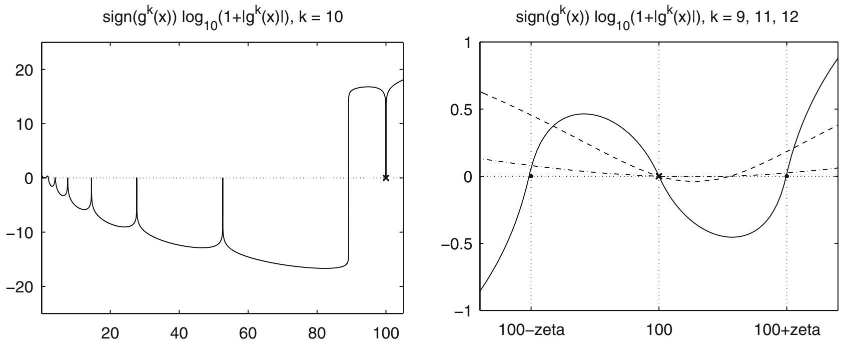


Fig. 4 The behavior of $g^k(x)$, see (13) (for better graphical view we plot $\operatorname{sign}(g^k(x))\log_{10}(1+|g^k(x)|)$). The left part shows the behavior in the whole interval of integration for $k=10$. The points of increase $\tilde{\lambda}_n$ and $\tilde{\lambda}_{n+1}$ are approximated by a single node of $I^k_{\tilde{\omega}}$ between them (the node close to 90 is still far away). The right part displays the behavior near λ_n for $k=9$ (*dash-dotted line*), $k=11$ (*dashed line*) and $k=12$ (*solid line*). For $k=12$ both nodes $\tilde{\lambda}_n$ and $\tilde{\lambda}_{n+1}$ are very closely approximated by the nodes of $I^k_{\tilde{\omega}}$

arithmetic in the presence of very close eigenvalues, a Ritz value in the Lanczos and the CG method initially converges to the cluster as fast as if the cluster were replaced by a single eigenvalue with the combined weight. Within a few further steps it converges very fast to one of the eigenvalues, with another Ritz value converging simultaneously to approximate the rest of the cluster. In the presence of more than two eigenvalues in a cluster, the story repeats until all eigenvalues in a cluster are approximated by individual Ritz values.

Now we consider the second modified example, where λ_n is perturbed to $\tilde{\lambda}_n = \lambda_n + \zeta$, $\tilde{\delta}_n = \delta_n$. Then $I^k_{\tilde{\omega}}$ converges to $I_{\tilde{\omega}}$ with the same speed as I^k_{ω} to I_{ω}, there is no delay, and the fact that $|E^k_{\omega} - E^k_{\tilde{\omega}}|$ is small can be *proved* using the result by Laurie [36]; see Sect. 3.3.

In the original motivating examples from Sect. 2 the situation is quite analogous, with the effects described on the simplified examples now taking place (for different values of k) near $\lambda_n, \lambda_{n-1}, \ldots$. A steep increase of $|I^k_{\omega} - I^k_{\tilde{\omega}}|$ significantly above $|\Delta|$ is well pronounced in the presence of well-separated rightmost points $\lambda_n, \lambda_{n-1}, \ldots,$ because they are fast approximated to high accuracy by the quadrature nodes. The phenomenon is almost independent of the position of the eigenvalues within the individual clusters (here $0 < \zeta < 0.1$ in order to ensure $\tilde{\lambda}_1 > 0$); see a similar statement in [49, Sect. 6.7, point (d), p. 559]. When λ_n is well-separated, the phenomenon *must* take place even for very small ζ.

The sensitivity of the Gauss–Christoffel quadrature is a consequence of the fact that $\tilde{\omega}$ has *more points of increase* (here two) close to the single points of increase of ω. The Gauss–Christoffel quadrature is sensitive because the number of points $\{\tilde{\lambda}_1, \ldots, \tilde{\lambda}_m\}$ in the support of $\tilde{\omega}$ is *larger* than the number in the support $\{\lambda_1, \ldots, \lambda_n\}$ of ω. More precisely, $\tilde{\omega}$ has more points of increase in the area where the gradient of $g^k(x)$ becomes very large as k increases. The second example, with the same number of points of increase, shows that moving each point of increase slightly does not cause sensitivity if the number of points is kept the same.

5.2 A continuous analog of the motivating example

Consider the analytic function $\Phi(x;\sigma) \equiv \sum_{i=1}^{n} \delta_i \varphi(x;\sigma,\lambda_i)$, where $0 < a < \lambda_1 < \cdots < \lambda_n < b$, $\delta_1, \ldots, \delta_n$ are as above and

$$\varphi(x;\sigma,t) \equiv \left[1 + e^{-\frac{x-t}{\sigma}}\right]^{-1} \tag{14}$$

is the strictly increasing sigmoid function with values between 0 and 1. Define the distribution function

$$\Omega(x;\sigma) \equiv c_0\, \Phi(x;\sigma), \qquad \int_a^b d\Omega(x;\sigma) = 1, \tag{15}$$

where c_0 is the normalization constant. Clearly, $\Omega(x;\sigma)$ approximates the step function from the motivating example:

$$\lim_{\sigma\to 0} \Omega(x;\sigma) = \omega(x),$$

for all $a \le x \le b$ except for $x = \lambda_i$, $i = 1, \ldots, n$, and the value of the parameter σ determines how closely $\Omega(x;\sigma)$ approximates $\omega(x)$.

In order make our computations accurate, we use the following linearization of $\Omega(x;\sigma)$. Divide the interval $[t - 10\sigma, t + 10\sigma]$ into $2m$ equal subintervals, with $m = 50$. Define $\hat{\varphi}(x;\sigma,t)$ to be the piecewise linear continuous function that interpolates $\Omega(x;\sigma)$ at the endpoints of the subintervals and is constant on $(-\infty, t - 10\sigma]$ and $[t + 10\sigma, \infty)$. Then, using $\hat{\Phi}(x;\sigma) \equiv \sum_{i=1}^{n} \delta_i \hat{\varphi}(x;\sigma,\lambda_i)$, we obtain a linearized distribution function

$$\hat{\Omega}(x;\sigma) \equiv c_1\, \hat{\Phi}(x;\sigma), \qquad \int_a^b d\hat{\Omega}(x;\sigma) = 1, \tag{16}$$

with c_1 the normalization constant.

The Riemann–Stieltjes integral $I_{\hat{\Omega}}(x^{-1}) = \int_a^b x^{-1} d\hat{\Omega}(x;\sigma)$ can be computed analytically. The recurrence coefficients of the orthogonal polynomials were computed by the double-reorthogonalized Lanczos process, with the corresponding integrals computed numerically. Using the partitioning described above and the fact that $\hat{\Omega}(x;\sigma)$ is linear on each subinterval, we conveniently use on each subinterval the Gauss–Legendre quadrature of sufficient order, implemented in MATLAB by Laurie in the file `r_jacobi.m`; see [24, Sect. 2.1]. For determining the quadrature nodes and weights we then use the standard approach implemented in the file `gauss.m` by Gautschi [23, pp. 153–154], [24, Sect. 2.4]. We use $\sigma = 10^{-8}$ and $\sigma = 10^{-6}$, $a = \lambda_1 - 10^{-5} = 10^{-1} - 10^{-5}$ and $b = \lambda_n + 10^{-5} = 100 + 10^{-5}$. Results for the original distribution function $\hat{\Omega}(x;10^{-8})$ and its *perturbation* $\hat{\Omega}(x;10^{-6})$, analogous to Figs. 1 and 2, are presented in Figs. 5 and 6. We can observe the same phenomena

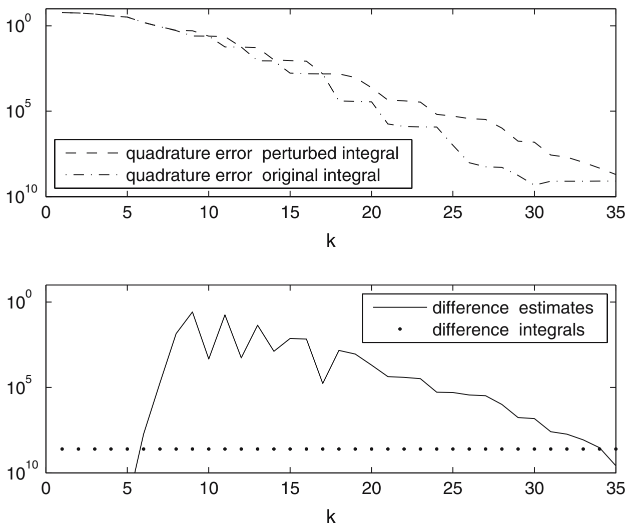


Fig. 5 Sensitivity of the Gauss–Christoffel quadrature for the continuous distribution function $\hat{\Omega}(x; \sigma)$. The *top graph* shows the error of the Gauss–Christoffel quadrature approximation for $f(x) = x^{-1}$ corresponding to the original distribution function with $\sigma = 10^{-8}$ (*dash-dotted line*) and to its perturbation with $\sigma = 10^{-6}$ (*dashed line*). The *bottom graph* displays the absolute value of difference in the estimates (*solid line*) and the difference between the approximated integrals (*dots*)

as in the motivating example, and the explanation is analogous. Since now the distribution functions are continuous, *many* quadrature nodes are eventually placed close to the rightmost $\lambda_n, \lambda_{n-1}, \ldots$ for both $\sigma = 10^{-8}$ and $\sigma = 10^{-6}$.

We emphasize that the observed Gauss–Christoffel quadrature sensitivity is a consequence of the fact that the support of $\hat{\Omega}(x; 10^{-6})$, which is the union of intervals of length 2×10^{-5} around the points λ_i, is *larger* than the corresponding support of $\hat{\Omega}(x; 10^{-8})$. If the supports were different (with the difference of a similar scale as before) but of the *same size*, no sensitivity would occur. Indeed, computation confirms that if $\tilde{\Omega}(x; 10^{-8})$ is a *perturbation* of the original distribution function $\hat{\Omega}(x; 10^{-8})$ obtained by shifting the individual sigmoids randomly 10^{-6} to the left or right, with subsequent normalization, then the quadrature nodes and weights change proportionally to the shifts of the individual sigmoids. The size of the difference between the Gauss–Christoffel quadrature estimates for $f(x) = x^{-1}$ and $\hat{\Omega}(x; 10^{-8})$ and $\tilde{\Omega}(x; 10^{-8})$ remains below or close to the size of the difference between the estimated integrals. In short, in agreement with our discussion above, no sensitivity of the Gauss–Christoffel quadrature appears.

5.3 Discussion: relationship to modified moments

We will explain that the sensitivity of the Gauss–Christoffel quadrature described above cannot be analyzed by investigation of modified moments. Our point is that the Gauss–Christoffel quadrature can be highly sensitive to some small changes of the original distribution function but insensitive to others, and this principal difference

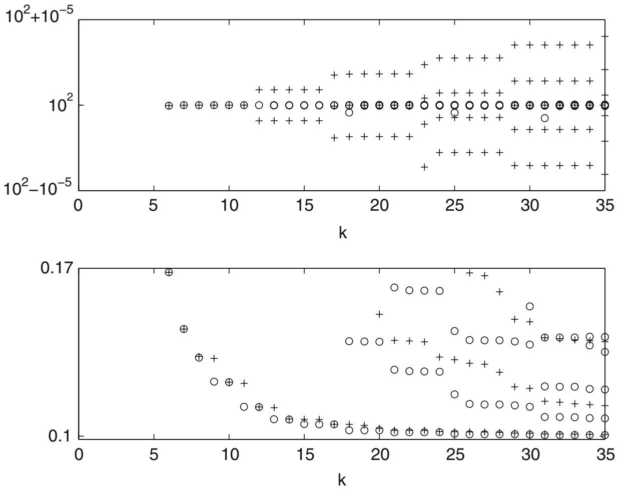


Fig. 6 Quadrature nodes corresponding to the distribution function $\hat{\Omega}(x;\sigma)$ with $\sigma = 10^{-8}$ (*circles*) and $\sigma = 10^{-6}$ (*pluses*) in two subintervals close to λ_n (*top*) and λ_1 (*bottom*). The *horizontal axis* is the number of nodes k in the quadrature

cannot be captured by the conditioning of the map $\mathbf{K}_k$ from the modified moments to the recurrence coefficients studied by Gautschi [23] and Beckermann and Bourreau [1]; see Sect. 3.1. In order to justify our claim, we will use the example with continuous distribution functions given above.

Using the previous notation, consider the original distribution function $\Omega_0(x) \equiv \hat{\Omega}(x;10^{-8})$ and two *perturbations* $\Omega_1(x) \equiv \hat{\Omega}(x;10^{-6})$, $\Omega_2(x) \equiv \tilde{\Omega}(x;10^{-8})$. We will now consider $\Omega_1(x)$ and $\Omega_2(x)$ two different *auxiliary distribution functions* in the sense of the modified Chebyshev algorithm; see Sect. 3.1. We know that the Gauss–Christoffel quadrature is sensitive to change from Ω_0 to Ω_1, and insensitive to change from Ω_0 to Ω_2. We might intuitively expect that the sensitivity in the first case would be reflected by the ill-conditioning of the map $\mathbf{K}_k^{(1)}$, which corresponds to the original distribution function Ω_0 and the auxiliary distribution function Ω_1, and that the insensitivity is in the second case would perhaps be accompanied by well-conditioning of the map $\mathbf{K}_k^{(2)}$, which corresponds to the original distribution function Ω_0 and the auxiliary distribution function Ω_2. But this is not true. The support of Ω_0 is different from the supports of Ω_1 and Ω_2, and using [1, Theorem 11, p. 93], we find that *both* maps $\mathbf{K}_k^{(1)}$ and $\mathbf{K}_k^{(2)}$ are notoriously ill-conditioned.

In order to illustrate this numerically, we consider the conjecture [1, p. 93] that there is a link between the condition number of $\mathbf{K}_k$ and that of the matrix of mixed moments of the polynomials orthogonal with respect to the original and auxiliary distribution functions (where the mixed moments are computed using the original distribution function; see [1, the matrix of transmission coefficients $T_n(\sigma, s)$ on p. 93]. The mixed moments appear in the modified Chebyshev algorithm as intermediate quantities; see [23, p. 76, relation (2.1.101)]). With a reference to the habilitation thesis of Beckermann, it is argued that the condition number GM_k of the matrix of modified

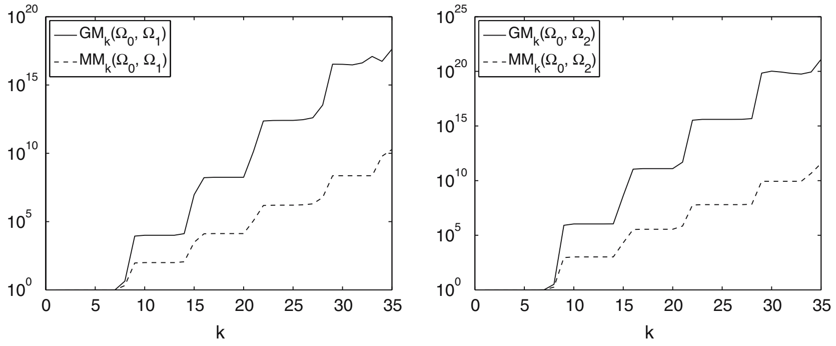


Fig. 7 Condition numbers of the matrix of the modified moments (GM_k, *solid line*) and of the matrix of mixed moments (MM_k, *dashed line*). The *left graph* corresponds to the distribution functions Ω_0 and Ω_1, and the *right graph* to the distribution functions Ω_0 and Ω_2

moments and the condition number MM_k of the matrix of mixed moments grow exponentially if the supports of the original and the auxiliary distribution functions do not coincide. This is illustrated in Fig. 7. Here GM_k is plotted by the solid line, MM_k by the dashed line. The left part corresponds to the distribution functions Ω_0 and Ω_1, the right part to Ω_0 and Ω_2. We can see the condition numbers GM_k and MM_k are growing essentially exponentially with k in *both* cases. (The staircase character of the plots is yet to be analyzed.) Using the conjecture in [1, p. 93], the fast growth of MM_k can be linked with the ill-conditioning of the maps $\mathbf{K}_k^{(1)}$ and $\mathbf{K}_k^{(2)}$.

In conclusion, the Gauss–Christoffel quadrature for a given distribution function can be *insensitive* to some perturbations despite the corresponding large MM_k and the corresponding ill-conditioning of the map $\mathbf{K}_k$.

5.4 Analytic distribution functions with different support

The phenomena described above can also be observed with analytic distribution functions. We present experiments with distribution function $\Omega(x;\sigma)$; see (15). The recurrence coefficients of the corresponding orthogonal polynomials are again computed by the double reorthogonalized Lanczos process, where for the numerical computation of the required integrals we use the MATLAB adaptive Lobatto quadrature `quadl`. The quadrature nodes and weights are then determined as above using the code `gauss.m`. We set $a = 0.1$ and $b = 100.2$. In order to reduce numerical errors below a noticeable level we take $\lambda_1 = 0.3$, $\lambda_n = 100$, $n = 4$, $\gamma = 0.55$, and consider the original distribution function $\Omega(x; 0.04)$ and its *perturbation* $\Omega(x; 0.08)$. Figure 8 shows results of the Gauss–Christoffel quadrature estimates for $k = 1, \ldots, 10$, $f(x) = x^{-1}$ (top) and $f(x) = 1 + \sin(x)$ (bottom). The sensitivity of the Gauss–Christoffel quadrature is here less pronounced than before. Still it is observable.

5.5 Analytic distribution functions with the same support

For slightly perturbed analytic functions with the same support the difference between the Gauss quadrature approximations $|I_{\tilde{\omega}}^k - I_{\omega}^k|$ is typically of the order $|I_{\tilde{\omega}} - I_{\omega}|$. For

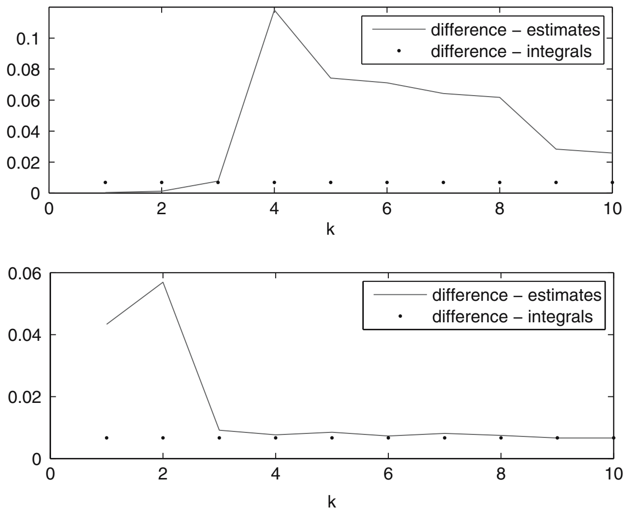


Fig. 8 Sensitivity of the Gauss–Christoffel quadrature for the analytic distribution function $\Omega(x;\sigma)$. The figure shows the absolute value of the difference in the quadrature estimates (*solid line*) and the difference between the approximated integrals (*dots*) for $f(x) = x^{-1}$ (*top*) and $f(x) = 1 + \sin(x)$ (*bottom*), corresponding to the original distribution function with $\sigma = 0.04$ and to its perturbation with $\sigma = 0.08$

small values of k, the errors of the corresponding estimates are much larger than the difference between them. Eventually the estimates must separate because they aim at approximating different integrals. The value k for which the two estimates separate is essentially determined by the difference between the approximated integrals. The roots of the corresponding orthogonal polynomials are typically very stable. We performed experiments, e.g., for the weight function $w(x) = \sqrt{x - x^2}$ for the shifted Chebyshev polynomials of the second kind, for the highly oscillatory weight function $w(x) = 1 + \cos(10\pi x)$, and for the Jacobi weight functions $w(x) = (1 - x)^{\alpha}(1 + x)^{\beta}$ with various values of the exponents and various perturbations.

We observed two characteristics in these experiments. First, the rate of decrease of the quadrature error was exponential, which can be explained using the Cauchy integrating kernels; see (8). Second, when perturbation of the distribution function preserves its support (here the whole interval), the quadrature is not sensitive. For an interesting example where the preservation of the support is linked with the analysis of the conditioning of the map $\mathbf{K}_k$ we refer to [1, Example 15, p. 96].

6 Conclusions

Literature about Gauss–Christoffel quadrature and about its computational aspects is extensive. This paper raises the following points which seem, however, new:

1. Gauss–Christoffel quadrature for a small number of quadrature nodes can be highly sensitive to small changes in the distribution function. In particular, the difference between the corresponding quadrature approximations (using the same number

of quadrature nodes) can be many orders of magnitude larger than the difference between the integrals being approximated.
2. This sensitivity in Gauss–Christoffel quadrature can be observed for discontinuous, continuous, and even analytic distribution functions, and for analytic integrands uncorrelated with changes in the distribution functions and with no singularity close to the interval of integration.
3. The sensitivity of the Gauss–Christoffel quadrature illustrated in this paper is related to the difference in the *size* of the support of the original and of the perturbed distribution functions. For a discrete distribution function, the size is the number of points of increase, and for a continuous distribution function it is the length (measure) of the union of intervals containing points at which the distribution function increases. In general, different supports of the *same size* do not exhibit sensitivity in quadrature results.
4. The sensitivity of Gauss–Christoffel quadrature cannot be explained using existing analysis based on modified moments. In our examples, if the support of the original distribution function differs in size from the support of the auxiliary (perturbed) distribution function, then the matrices of both modified and mixed moments become highly ill-conditioned. The same is true if the supports are different but of the same size. But only in the case of different size of the supports are the recurrence coefficients (i.e., the entries of the Jacobi matrix) and the Gauss–Christoffel quadrature estimates highly sensitive to the perturbation.

Many open questions remain. We give several *examples* of sensitivity of the Gauss–Christoffel quadrature. It would certainly be of great interest to describe the *classes of problems* for which the Gauss–Christoffel quadrature is sensitive to small perturbations of the distribution function, and determine which of them are of practical importance. Application of these results to theory of the conjugate gradient and Lanczos methods in finite precision arithmetic will be considered in our future work. Another highly relevant question is how to measure differences between distribution functions.

Acknowledgments Numerous comments and insightful suggestions of the anonymous referees were very helpful in revising the original version of the manuscript. The authors are indebted to Dirk Laurie and Jarda Kautský for their interest in this work, for their questions, thoughts and for performing independent numerical experiments. We also gratefully acknowledge the groundbreaking work in this field by Walter Gautschi and Gene H. Golub.

References

1. Beckermann, B., Bourreau, E.: How to choose modified moments?. J. Comput. Appl. Math. **98**(1), 81–98 (1998)
2. Boley, D., Golub, G.H.: A survey of matrix inverse eigenvalue problems. Inverse Probl **3**(41), 595–622 (1987)
3. de Boor, C., Golub, G.H.: The numerically stable reconstruction of a Jacobi matrix from spectral data. Linear Algebra Appl. **21**(3), 245–260 (1978)
4. Carpraux, J.F., Godunov, S.K., Kuznetsov, S.V.: Condition number of the Krylov bases and subspaces. Linear Algebra Appl. **248**, 137–160 (1996)

5. Dahlquist, G., Golub, G.H., Nash, S.G.: Bounds for the error in linear systems. In: Semi-infinite programming (Proc. Workshop, Bad Honnef, 1978). Lecture Notes in Control and Information Sci., vol. 15, pp. 154–172. Springer, Berlin (1979)
6. Davis, P.J., Rabinowitz, P.: Methods of numerical integration, 2nd edn. Computer Science and Applied Mathematics. Academic, Orlando (1984)
7. Demmel, J., Kahan, W.: Accurate singular values of bidiagonal matrices. SIAM J. Sci. Stat. Comput. **11**(5), 873–912 (1990)
8. Dhillon, I.S.: A new O(n^2) algorithm for the symmetric tridiagonal eigenvalue/eigenvector problem. PhD Thesis, University of California, Berkeley (1997)
9. Dhillon, I.S., Parlett, B.N.: Orthogonal eigenvectors and relative gaps. SIAM J. Matrix Anal. Appl. (Electronic) **25**(3), 858–899 (2003)
10. Dhillon, I.S., Parlett, B.N.: Multiple representations to compute orthogonal eigenvectors of symmetric tridiagonal matrices. Linear Algebra Appl. **387**, 1–28 (2004)
11. Druskin, V., Borcea, L., Knizhnerman, L.: On the sensitivity of Lanczos recursions to the spectrum. Linear Algebra Appl. **396**, 103–125 (2005)
12. Elhay, S., Golub, G.H., Kautsky, J.: Updating and downdating of orthogonal polynomials with data fitting applications. SIAM J. Matrix Anal. Appl. **12**(2), 327–353 (1991)
13. Elhay, S., Golub, G.H., Kautsky, J.: Jacobi matrices for sums of weight functions. BIT **32**(1), 143–166 (1992)
14. Fischer, H.J.: On generating orthogonal polynomials for discrete measures. Z. Anal. Anwendungen **17**(1), 183–205 (1998)
15. Gautschi, W.: Construction of Gauss-Christoffel quadrature formulas. Math. Comp. **22**, 251–270 (1968)
16. Gautschi, W.: On the construction of Gaussian quadrature rules from modified moments. Math. Comp. **24**, 245–260 (1970)
17. Gautschi, W.: Questions on numerical condition related to polynomials. In: Recent Advances in Numerical Analysis (Madison, 1978), pp. 45–72. Academic, New York (1978)
18. Gautschi, W.: A survey of Gauss–Christoffel quadrature formulae. In: Butzer, P.L., Fehér, F. (eds.) E.B. Christoffel: The Influence of His Work on Mathematics and Physics, pp. 72–147. Birkhäuser, Basel (1981)
19. Gautschi, W.: On generating orthogonal polynomials. SIAM J. Sci. Stat. Comput. **3**(3), 289–317 (1982)
20. Gautschi, W.: Is the recurrence relation for orthogonal polynomials always stable? BIT **33**(2), 277–284 (1993)
21. Gautschi, W.: Numerical analysis: an introduction. Birkhäuser Boston Inc., Boston (1997)
22. Gautschi, W.: The interplay between classical analysis and (numerical) linear algebra—a tribute to Gene H. Golub. Electron. Trans. Numer. Anal. (Electronic) **13**, 119–147 (2002)
23. Gautschi, W.: Orthogonal polynomials: computation and approximation. Numerical Mathematics and Scientific Computation. Oxford University Press, New York (2004)
24. Gautschi, W.: Orthogonal polynomials (in Matlab). J. Comput. Appl. Math. **178**(1–2), 215–234 (2005)
25. Gautschi, W., Varga, R.S.: Error bounds for Gaussian quadrature of analytic functions. SIAM J. Numer. Anal. **20**(6), 1170–1186 (1983)
26. Golub, G.H.: Matrix computation and the theory of moments. In: Proceedings of the International Congress of Mathematicians, vol. 1, 2 (Zürich, 1994), pp. 1440–1448. Birkhäuser, Basel (1995)
27. Golub, G.H., Van Loan, C.F.: Matrix computations, Johns Hopkins Series in the Mathematical Sciences, vol. 3, 2nd edn. Johns Hopkins University Press, Baltimore (1989)
28. Golub, G.H., Welsch, J.H.: Calculation of Gauss quadrature rules. Math. Comp. 23 (1969), 221-230; addendum, ibid. **23**(106, loose microfiche suppl), A1–A10 (1969)
29. Gragg, W.B., Harrod, W.J.: The numerically stable reconstruction of Jacobi matrices from spectral data. Numer. Math. **44**(3), 317–335 (1984)
30. Grosser, B., Lang, B.: On symmetric eigenproblems induced by the bidiagonal SVD. SIAM J. Matrix Anal. Appl. (Electronic) **26**(3), 599–620 (2005)
31. Hestenes M.R., Stiefel, E.: Methods of conjugate gradients for solving linear systems. J. Res. Nat. Bur. Standards **49**, 409–436 (1953) (1952)
32. Isaacson, E., Keller, H.B.: Analysis of numerical methods. Wiley, New York (1966)
33. Kautský, J., Golub, G.H.: On the calculation of Jacobi matrices. Linear Algebra Appl. **52/53**, 439–455 (1983)

34. Kuznetsov, S.V.: Perturbation bounds of the Krylov bases and associated Hessenberg forms. Linear Algebra Appl. **265**, 1–28 (1997)
35. Lanczos, C.: Applied analysis. Prentice Hall, Inc., Englewood Cliffs (1956)
36. Laurie, D.P.: Accurate recovery of recursion coefficients from Gaussian quadrature formulas. J. Comput. Appl. Math. **112**(1–2), 165–180 (1999)
37. Laurie, D.P.: Questions related to Gaussian quadrature formulas and two-term recursions. In: Applications and computation of orthogonal polynomials (Oberwolfach, 1998). Int. Ser. Numer. Math. **131**, 133–144 (1999)
38. Laurie, D.P.: Computation of Gauss-type quadrature formulas. J. Comput. Appl. Math. **127**(1–2), 201–217 (2001)
39. Li, R.C.: Relative perturbation theory. III. More bounds on eigenvalue variation. Linear Algebra Appl. **266**, 337–345 (1997)
40. Meurant, G., Strakoš, Z.: Lanczos and conjugate gradient algorithms in finite precision arithmetic. Acta Numer. **15**, 471–542 (2006)
41. Milne-Thomson, L.M.: The Calculus of Finite Differences. MacMillan & Co., New York (1960)
42. Natterer, F.: Discrete Gel'fand–Levitan theory, www page of the author (1999)
43. Nevai, P.G.: Orthogonal polynomials. Mem. Am. Math. Soc. **18**(213), v+185 (1979)
44. O'Leary, D.P., Stewart, G.W., Vandergraft, J.S.: Estimating the largest eigenvalue of a positive definite matrix. Math. Comp. **33**(148), 1289–1292 (1979)
45. Paige, C.C., Van Dooren, P.: Sensitivity analysis of the Lanczos reduction. Numer. Linear Algebra Appl. **6**(1), 29–50 (1999). Czech-US Workshop in Iterative Methods and Parallel Computing, Part I (Milovy, 1997)
46. Parlett, B.N., Dhillon, I.S.: Relatively robust representations of symmetric tridiagonals. In: Proceedings of the International Workshop on Accurate Solution of Eigenvalue Problems (University Park, PA, 1998), vol. 309, pp. 121–151 (2000)
47. Parlett, B.N., Simon, H., Stringer, G.: On estimating the largest eigenvalue with the Lanczos algorithm. Math. Comp. **38**, 153–165 (1982)
48. Reichel, L.: Fast QR decomposition of Vandermonde-like matrices and polynomial least squares approximation. SIAM J. Matrix Anal. Appl. **12**(3), 552–564 (1991)
49. van der Sluis, A., van der Vorst, H.: The rate of convergence of conjugate gradients. Numer. Math. **48**, 543–560 (1986)
50. van der Sluis, A., van der Vorst, H.: The convergence behavior of Ritz values in the presence of close eigenvalues. Linear Algebra Appl. **88**, 651–694 (1987)
51. Strakoš, Z.: On the real convergence rate of the conjugate gradient method. Linear Algebra Appl. **154–156**, 535–549 (1991)
52. Strakoš, Z., Tichý, P.: On error estimation in the conjugate gradient method and why it works in finite precision computations. Electron. Trans. Numer. Anal. (Electronic) **13**, 56–80 (2002)
53. Stroud, A.H., Secrest, D.: Gaussian quadrature formulas. Prentice-Hall Inc., Englewood Cliffs (1966)
54. Swarztrauber, P.N.: On computing the points and weights for Gauss-Legendre quadrature. SIAM J. Sci. Comput. (Electronic) **24**(3), 945–954 (2002)
55. Szegö, G.: Orthogonal Polynomials. Colloquium Publications, vol. XXIII. American Mathematical Society, New York (1939)
56. Uvarov, V.B.: The connection between systems of polynomials that are orthogonal with respect to different distribution functions. Ž. Vyčisl. Mat. i Mat. Fiz. **9**, 1253–1262 (1969)
57. Vorobyev, Y.V.: Methods of moments in applied mathematics. Translated from the Russian by Bernard Seckler. Gordon and Breach Science Publishers, New York (1965)
58. Willems, P., Lang, B., Vömel, C.: Computing the bidiagonal SVD using multiple relatively robust representations. SIAM J. Matrix Anal. Appl. **28**(4), 907–926 (2006)
59. Xu, S.F.: A stability analysis of the Jacobi matrix inverse eigenvalue problem. BIT **33**(4), 695–702 (1993)

Electronic Transactions on Numerical Analysis.
Volume 13, pp. 56-80, 2002.
Copyright © 2002, Kent State University.
ISSN 1068-9613.

ETNA
Kent State University
etna@mcs.kent.edu

ON ERROR ESTIMATION IN THE CONJUGATE GRADIENT METHOD AND WHY IT WORKS IN FINITE PRECISION COMPUTATIONS [*]

ZDENĚK STRAKOŠ[†] AND PETR TICHÝ[*]

Abstract. In their paper published in 1952, Hestenes and Stiefel considered the conjugate gradient (CG) method an iterative method which terminates in at most n steps if no rounding errors are encountered [24, p. 410]. They also proved identities for the A-norm and the Euclidean norm of the error which could justify the stopping criteria [24, Theorems 6:1 and 6:3, p. 416]. The idea of estimating errors in iterative methods, and in the CG method in particular, was independently (of these results) promoted by Golub; the problem was linked to Gauss quadrature and to its modifications [7], [8]. A comprehensive summary of this approach was given in [15], [16]. During the last decade several papers developed error bounds algebraically without using Gauss quadrature. However, we have not found any reference to the corresponding results in [24]. All the existing bounds assume exact arithmetic. Still they seem to be in a striking agreement with finite precision numerical experiments, though in finite precision computations they estimate quantities which can be orders of magnitude different from their exact precision counterparts! For the lower bounds obtained from Gauss quadrature formulas this nontrivial phenomenon was explained, with some limitations, in [17].

In our paper we show that the lower bound for the A-norm of the error based on Gauss quadrature ([15], [17], [16]) is mathematically equivalent to the original formula of Hestenes and Stiefel [24]. We will compare existing bounds and we will demonstrate necessity of a proper rounding error analysis: we present an example of the well-known bound which can fail in finite precision arithmetic. We will analyse the simplest bound based on [24, Theorem 6:1], and prove that it is numerically stable. Though we concentrate mostly on the lower bound for the A-norm of the error, we describe also an estimate for the Euclidean norm of the error based on [24, Theorem 6:3]. Our results are illustrated by numerical experiments.

Key words. conjugate gradient method, Gauss quadrature, evaluation of convergence, error bounds, finite precision arithmetic, rounding errors, loss of orthogonality.

AMS subject classifications. 15A06, 65F10, 65F25, 65G50.

1. Introduction.

Consider a symmetric positive definite matrix $A \in \mathbb{R}^{n\times n}$ and a right-hand side vector $b \in \mathbb{R}^n$ (for simplicity of notation we will assume A, b real; generalization to complex data will be obvious). This paper investigates numerical estimation of errors in iterative methods for solving linear systems

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

In particular, we focus on the conjugate gradient method (CG) of Hestenes and Stiefel [24] and on the lower estimates of the A-norm (also called the energy norm) of the error, which has important meaning in physics and quantum chemistry, and plays a fundamental role in evaluating convergence [1], [2].

Starting with the initial approximation x_0, the conjugate gradient approximations are determined by the condition

$$\begin{aligned} x_j &\in x_0 + \mathcal{K}_j(A, r_0) \\ \|x - x_j\|_A &= \min_{u \in x_0 + \mathcal{K}_j(A,r_0)} \|x - u\|_A, \end{aligned} \tag{1.2}$$

i.e. they minimize the A-norm of the error

$$\|x - x_j\|_A = \big((x - x_j), A(x - x_j)\big)^{\frac{1}{2}}$$

[*] Received June 7, 2001. Accepted for publication August 18, 2002. Recommended by L. Reichel.

[†] Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07 Praha 8, Czech Republic. E-mail: strakos@cs.cas.cz and petr.tichy@centrum.cz. This research was supported by the Grant Agency of the Czech Republic under grant No. 201/02/0595.

over all methods generating approximations in the manifold $x_0 + \mathcal{K}_j(A, r_0)$. Here

$$\mathcal{K}_j(A, r_0) = \mathrm{span}\{r_0, Ar_0, \dots A^{j-1} r_0\}$$

is the j-th Krylov subspace generated by A with the initial residual r_0, $r_0 = b - Ax_0$, and x is the solution of (1.1). The standard implementation of the CG method was given in [24, (3:1a)-(3:1f)]:

Given x_0, $r_0 = b - Ax_0$, $p_0 = r_0$, and for $j = 1, 2, \dots,$ let

$$\begin{aligned} \gamma_{j-1} &= (r_{j-1}, r_{j-1})/(p_{j-1}, Ap_{j-1}), \\ x_j &= x_{j-1} + \gamma_{j-1}\, p_{j-1}, \\ r_j &= r_{j-1} - \gamma_{j-1}\, Ap_{j-1}, \\ \delta_j &= (r_j, r_j)/(r_{j-1}, r_{j-1}), \\ p_j &= r_j + \delta_j\, p_{j-1}. \end{aligned} \tag{1.3}$$

The residual vectors $\{r_0, r_1, \dots, r_{j-1}\}$ form an orthogonal basis and the direction vectors $\{p_0, p_1, \dots, p_{j-1}\}$ an A-orthogonal basis of the j-th Krylov subspace $\mathcal{K}_j(A, r_0)$.

In [24] Hestenes and Stiefel considered CG as an iterative procedure. They presented relations [24, (6:1)-(6:3) and (6:5), Theorems 6:1 and 6:3] as justifications of a possible stopping criterion for the algorithm. In our notation these relations become

$$\|x - x_{j-1}\|_A^2 - \|x - x_j\|_A^2 = \gamma_{j-1}\|r_{j-1}\|^2, \tag{1.4}$$

$$\|x - x_j\|_A^2 - \|x - x_k\|_A^2 = \sum_{i=j}^{k-1} \gamma_i \|r_i\|^2, \quad 0 \le j < k \le n, \tag{1.5}$$

$$x_j = x_0 + \sum_{l=0}^{j-1} \gamma_l p_l = x_0 + \sum_{l=0}^{j-1} \frac{\|x - x_l\|_A^2 - \|x - x_j\|_A^2}{\|r_l\|^2}\, r_l, \tag{1.6}$$

$$\|x - x_{j-1}\|^2 - \|x - x_j\|^2 = \frac{\|x - x_{j-1}\|_A^2 + \|x - x_j\|_A^2}{\mu(p_{j-1})}, \qquad \mu(p_{j-1}) = \frac{(p_{j-1}, Ap_{j-1})}{\|p_{j-1}\|^2}. \tag{1.7}$$

Please note that (1.5) represents an identity describing the decrease of the A-norm of the error in terms of quantities available in the algorithm, while (1.7) describes decrease of the Euclidean norm of the error in terms of the A-norm of the error in the given steps.

Hestenes and Stiefel did not give any particular stopping criterion. They emphasized, however, that while the A-norm of the error and the Euclidean norm of the error had to decrease monotonically at each step, the residual norm oscillated and might even increase in each but the last step. An example of this behaviour was used in [23].

The paper [24] is frequently referenced, but some of its results has not been paid much attention. Residual norms have been (and still are) commonly used for evaluating convergence of CG. The possibility of using (1.4)–(1.7) for constructing a stopping criterion has not been, to our knowledge, considered.

An interest in estimating error norms in the CG method reappeared with works of Golub and his collaborators. Using some older results [7], Dahlquist, Golub and Nash [8] related error bounds to Gauss quadrature (and to its modifications). The approach presented in that paper became a basis for later developments. It is interesting to note that the relationship of the CG method to the Riemann-Stieltjes integral and Gauss quadrature was described in detail in [24, Section 14], but without any link to error estimation. The work of Golub and his collaborators was independent of [24].

The paper [8] brought also into attention an important issue of rounding errors. The authors noted that in order to guarantee the numerical stability of the computed Gauss quadrature nodes and weights, the computed basis vectors had to be reorthogonalized. That means that the authors of that paper were from the very beginning aware of the fact that rounding errors might play a significant role in the application of their bounds to practical computations. In the numerical experiments used in [8] the effect of rounding errors were, however, not noticeable. This can be explained using the results by Paige ([33], [34], [35] and [36]). Due to the distribution of eigenvalues of the matrix used in [8] Ritz values do not converge to the eigenvalues until the last few steps. Before this convergence takes place there is no significant loss of orthogonality and the effects of rounding errors are not visible.

Error bounds in iterative methods were intensively studied or used in many later papers and in several books, see, e.g. [9], [10], [12], [15], [17], [16], [11], [21], [28], [29], [30], [4], [6]. Except for [17], effects of rounding errors were not analysed in these publications.

Frommer and Weinberg [13] pointed out the problem of applying exact precision formulas to finite precision computations, and proposed to use interval arithmetic for computing verified error bounds. As stated in [13, p. 201], this approach had serious practical limitations. Axelsson and Kaporin [3] considered preconditioned conjugate gradients and presented (1.5) independently of [24]. Their derivation used (global) mutual A-orthogonality among the direction vectors $p_j, j = 0, 1, \ldots, n-1$. They noticed that the numerical values found from the resulting estimate were identical to those obtained from Gauss quadrature, but did not prove this coincidence. They also noticed the potential difficulty due to rounding errors. They presented an observation that loss of orthogonality did not destroy applicability of their estimate. Calvetti et al. [5] presented several bounds and estimates for the A-norm of the error, and addressed a problem of cancellation in their computations [5, relation (46)].

In our paper we briefly recall some error estimates published after (and independently of) (1.4)-(1.7) in [24]. For simplicity of our exposition we will concentrate mostly on the A-norm of the error $\|x - x_j\|_A$. We will show that the simplest possible estimate for $\|x - x_j\|_A$, which follows from the relation (1.4) published in the original paper [24], is mathematically (in exact arithmetic) equivalent to the corresponding bounds developed later. In finite precision arithmetic, rounding errors in *the whole computation, not only in the computation of the convergence bounds*, must be taken into account. We emphasize that rounding error analysis of formulas for computation of the convergence bounds represents in almost all cases a simple and unimportant part of the problem. Almost all published convergence bounds (including those given in [5]) can be computed accurately (i.e. computation of the bounds using given formulas is not significantly affected by rounding errors). But this does not prove that these bounds give anything reasonable when they are applied to finite precision CG computations. We will see an example of the accurately computed bound which gives no useful information about the convergence of CG in finite precision arithmetic in Section 6.

An example of rounding error analysis for the bounds based on Gauss quadrature was presented in [17]. The results from [17] rely on the work by Paige and Greenbaum ([36], [19] and [22]). Though [17] gives a strong qualitative justification of the bounds in finite precision arithmetic, this justification is applicable only until $\|x - x_j\|_A$ reaches the square root of the machine precision. Moreover, quantitative expressions for the rounding error terms are very complicated. They contain factors which are not tightly estimated (see [19], [22]). Here we complement the analysis from [17] by substantially stronger results. We prove that the simplest possible lower bound for $\|x - x_j\|_A$ based on (1.4) works also for numerically computed quantities till $\|x - x_j\|_A$ reaches its ultimate attainable accuracy.

The paper is organized as follows. In Section 2 we briefly describe relations between the CG and Lanczos methods. Using the orthogonality of the residuals, these algorithms are related to sequences of orthogonal polynomials, where the inner product is defined by a Riemann-Stieltjes integral with some particular distribution function $\omega(\lambda)$. The value of the j-th Gauss quadrature approximation to this Riemann-Stieltjes integral for the function $1/\lambda$ is the complement to the error in the j-th iteration of the CG method measured by $\|x - x_j\|_A^2/\|r_0\|^2$. In Section 3 we reformulate the result of the Gauss quadrature using quantities that are at our disposal during the CG iterations. In Section 4 we use the identities from Section 3 for estimation of the A-norm of the error in the CG method, and we compare the main existing bounds. Section 5 describes delay of convergence due to rounding errors. Section 6 explains why applying exact precision convergence estimates to finite precision CG computations represents a serious problem which must be properly addressed. Though exact precision CG and finite precision CG can dramatically differ, some exact precision bounds seem to be in good agreement with the finite precision computations. Sections 7–10 explain this paradox. The individual terms in the identities which the convergence estimates are based on can be strongly affected by rounding errors. *The identities as a whole, however, hold true (with small perturbations) also in finite precision arithmetic*. Numerical experiments are presented in Section 11.

When it will be helpful we will use the word “ideally” (or “mathematically”) to refer to a result that would hold using exact arithmetic, and “computationally” or “numerically” to a result of a finite precision computation.

2. Method of conjugate gradients and Gauss quadrature.

For A and r_0 the Lanczos method [27] generates ideally a sequence of orthonormal vectors $v_1, v_2, \ldots$ via the recurrence

Given $v_1 = r_0/\|r_0\|$, $\beta_1 \equiv 0$, and for $j = 1, 2, \ldots,$ let

$$\begin{aligned} \alpha_j &= (Av_j - \beta_j v_{j-1}, v_j), \\ w_j &= Av_j - \alpha_j v_j - \beta_j v_{j-1}, \\ \beta_{j+1} &= \|w_j\|, \\ v_{j+1} &= w_j/\beta_{j+1}. \end{aligned} \tag{2.1}$$

Denoting by $V_j = [v_1, \ldots, v_j]$ the n by j matrix having the Lanczos vectors $\{v_1, \ldots, v_j\}$ as its columns, and by T_j the symmetric tridiagonal matrix with positive subdiagonal

$$T_j = \begin{pmatrix} \alpha_1 & \beta_2 & & \\ \beta_2 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_j \\ & & \beta_j & \alpha_j \end{pmatrix} \tag{2.2}$$

the formulas (2.1) are written in the matrix form

$$AV_j = V_j T_j + \beta_{j+1} v_{j+1} e_j^T, \tag{2.3}$$

where e_j is the j-th column of the n by n identity matrix. Comparing (1.3) with (2.1) gives

$$v_{j+1} = (-1)^j \frac{r_j}{\|r_j\|}, \tag{2.4}$$

and also relations between the recurrence coefficients:

$$\begin{aligned}\alpha_j &= \frac{1}{\gamma_{j-1}} + \frac{\delta_{j-1}}{\gamma_{j-2}}, \qquad \delta_0 \equiv 0,\ \gamma_{-1} \equiv 1, \\ \beta_{j+1} &= \frac{\sqrt{\delta_j}}{\gamma_{j-1}}\,. \end{aligned} \tag{2.5}$$

Finally, using the change of variables

$$x_j = x_0 + V_j\, y_j, \tag{2.6}$$

and the orthogonality relation between r_j and the basis $\{v_1, v_2, \dots, v_j\}$ of $\mathcal{K}_j(A, r_0)$, we see that

$$\begin{aligned} 0 = V_j^T r_j &= V_j^T (b - Ax_j) = V_j^T (r_0 - AV_j\, y_j) \\ &= e_1 \|r_0\| - V_j^T A V_j\, y_j = e_1 \|r_0\| - T_j\, y_j\,. \end{aligned}$$

Ideally, the CG approximate solution x_j can therefore be determined by solving

$$T_j\, y_j = e_1 \|r_0\|\,, \tag{2.7}$$

with subsequent using of (2.6).

Orthogonality of the CG residuals creates the elegance of the CG method which is represented by its link to the world of classical orthogonal polynomials. Using (1.3), the j-th error resp. residual can be written as a polynomial in the matrix A applied to the initial error resp. residual,

$$x - x_j = \varphi_j(A)\ (x - x_0), \qquad r_j = \varphi_j(A)\ r_0, \qquad \varphi_j \in \Pi_j, \tag{2.8}$$

where Π_j denotes the class of polynomials of degree at most j having the property $\varphi(0) = 1$ (that is, the constant term equal to one). Consider the eigendecomposition of the symmetric matrix A in the form

$$A = U \Lambda U^T, \qquad U U^T = U^T U = I, \tag{2.9}$$

where $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$ and $U = [u_1, \dots, u_n]$ is the matrix having the normalized eigenvectors of A as its columns. Substituting (2.9) and (2.8) into (1.2) gives

$$\begin{aligned} \|x - x_j\|_A = \|\varphi_j(A)(x - x_0)\|_A &= \min_{\varphi \in \Pi_j} \|\varphi(A)(x - x_0)\|_A = \min_{\varphi \in \Pi_j} \|\varphi(A) r_0\|_{A^{-1}} \\ &= \min_{\varphi \in \Pi_j} \left\{ \sum_{i=1}^{n} \frac{(r_0, u_i)^2}{\lambda_i}\, \varphi^2(\lambda_i) \right\}^{1/2} . \end{aligned} \tag{2.10}$$

Consequently, for A symmetric positive definite the rate of convergence of CG is determined by the distribution of eigenvalues of A and by the size of the components of r_0 in the direction of the individual eigenvectors.

Similarly to (2.8), v_{j+1} is linked with some monic polynomial ψ_j,

$$v_{j+1} = \psi_j(A)\, v_1 \cdot \frac{1}{\beta_2 \beta_3 \dots \beta_{j+1}}\,. \tag{2.11}$$

Using the orthogonality of v_{j+1} to $v_1, \ldots, v_j$, the polynomial ψ_j is determined by the minimizing condition

$$\|\psi_j(A)v_1\| = \min_{\psi \in \mathcal{M}_j} \|\psi(A)v_1\| = \min_{\psi \in \mathcal{M}_j} \left\{ \sum_{i=1}^{n} (v_1, u_i)^2 \, \psi^2(\lambda_i) \right\}^{1/2} , \tag{2.12}$$

where $\mathcal{M}_j$ denotes the class of monic polynomials of degree j.

We will explain what we consider the essence of the CG and Lanczos methods.

Whenever the CG or the Lanczos method (defined by (1.3) resp. by (2.1)) is considered, there is a sequence $1, \psi_1, \psi_2, \ldots$ of the monic orthogonal polynomials determined by (2.12). These polynomials are orthogonal with respect to the discrete inner product

$$(f, g) = \sum_{i=1}^{n} \omega_i f(\lambda_i) g(\lambda_i) \, , \tag{2.13}$$

where the weights ω_i are determined as

$$\omega_i = (v_1, u_i)^2, \qquad \sum_{i=1}^{n} \omega_i = 1 \, , \tag{2.14}$$

$(v_1 = r_0/\|r_0\|)$. For simplicity of notation we assume that all the eigenvalues of A are distinct and increasingly ordered (an extension to the case of multiple eigenvalues will be obvious). Let ζ, ξ be such that $\zeta \leq \lambda_1 < \lambda_2 < \ldots < \lambda_n \leq \xi$. Consider the distribution function $\omega(\lambda)$ with the finite points of increase $\lambda_1, \lambda_2, \ldots, \lambda_n$,

$$\begin{aligned} \omega(\lambda) &= 0 && \text{for} && \lambda < \lambda_1 \, , \\ \omega(\lambda) &= \sum_{l=1}^{i} \omega_l && \text{for} && \lambda_i \leq \lambda < \lambda_{i+1} \, , \\ \omega(\lambda) &= 1 && \text{for} && \lambda_n \leq \lambda \, , \end{aligned} \tag{2.15}$$

see Fig. 2.1, and the corresponding Riemann-Stieltjes integral

$$\int_{\zeta}^{\xi} f(\lambda) \, d\omega(\lambda) = \sum_{i=1}^{n} \omega_i f(\lambda_i) \, . \tag{2.16}$$

Then (2.12) can be rewritten as

$$\psi_j = \arg \min_{\psi \in \mathcal{M}_j} \left\{ \int_{\zeta}^{\xi} \psi^2(\lambda) \, d\omega(\lambda) \right\} , \qquad j = 0, 1, 2, \ldots, n \, . \tag{2.17}$$

The j steps of the CG resp. the Lanczos method starting with $\|r_0\| v_1$ resp. v_1 determine a symmetric tridiagonal matrix (with a positive subdiagonal) T_j (2.2). Consider, analogously to (2.9), the eigendecomposition of T_j in the form

$$T_j = S_j \Theta_j S_j^T \, , \qquad S_j^T S_j = S_j S_j^T = I, \tag{2.18}$$

$\Theta_j = \operatorname{diag}(\theta_1^{(j)}, \ldots, \theta_j^{(j)})$, $S_j = [s_1^{(j)}, \ldots, s_j^{(j)}]$. Please note that we can look at T_j also as determined by the CG or the Lanczos method applied to the j-dimensional problem $T_j y_j = e_1 \|r_0\|$ resp. T_j with initial residual $e_1 \|r_0\|$ resp. starting vector e_1. Clearly, we can construct

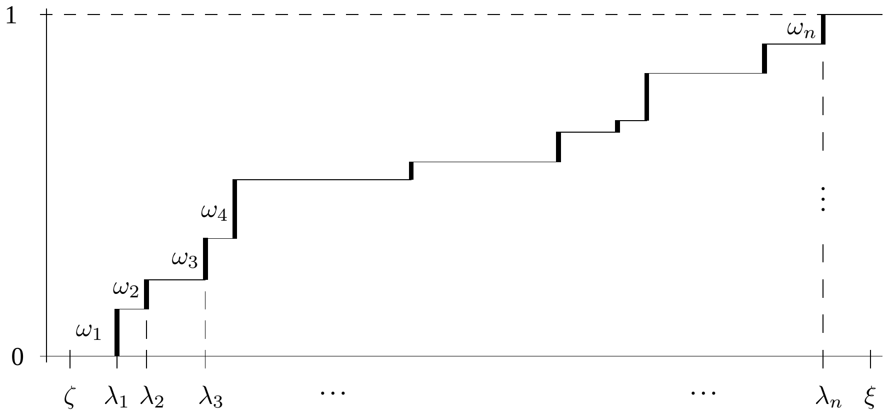


FIG. 2.1. *Distribution function* $\omega(\lambda)$

Riemann-Stieltjes integral for this j-dimensional problem similarly as above. Let $\zeta \le \theta_1^{(j)} < \theta_2^{(j)} < \ldots < \theta_j^{(j)} \le \xi$ be the eigenvalues of T_j (Ritz values, they must be distinct, see, e.g. [38, Chapter 7]). Let

$$\omega_i^{(j)} = (e_1, s_i^{(j)})^2, \qquad \sum_{i=1}^{j} \omega_i^{(j)} = 1 \tag{2.19}$$

be the weights determined by the squared size of the components of e_1 in the direction of T_j's eigenvectors, and

$$\begin{aligned} \omega^{(j)}(\lambda) &= 0 && \text{for} && \lambda < \theta_1^{(j)}\,, \\ \omega^{(j)}(\lambda) &= \sum_{l=1}^{i} \omega_l^{(j)} && \text{for} && \theta_i^{(j)} \le \lambda < \theta_{i+1}^{(j)}, \\ \omega^{(j)}(\lambda) &= 1 && \text{for} && \theta_j^{(j)} \le \lambda\,. \end{aligned}$$

Then the first j polynomials from the set $\{1, \psi_1, \ldots, \psi_n\}$ determined by (2.17) are also determined by the condition based on the Riemann-Stieltjes integral with the distribution function $\omega^{(j)}(\lambda)$

$$\psi_l = \arg \min_{\psi \in \mathcal{M}_l} \left\{ \int_\zeta^\xi \psi^2(\lambda)\, d\omega^{(j)}(\lambda) \right\}, \qquad l = 0, 1, \ldots, j\,, \tag{2.20}$$

(we can look at the subsequence $\{1, \psi_1, \ldots, \psi_j\}$ as determined by the CG or the Lanczos method applied to the j-dimensional problem described above). The integral

$$\int_\zeta^\xi f(\lambda)\, d\omega^{(j)}(\lambda) = \sum_{i=1}^{j} \omega_i^{(j)} f(\theta_i^{(j)}) \tag{2.21}$$

is the well-known j-th Gauss quadrature approximation of the integral (2.16), see, e.g., [14]. Thus, the CG and Lanczos methods determine the sequence of distribution functions $\omega^{(1)}(\lambda), \omega^{(2)}(\lambda), \ldots, \omega^{(j)}(\lambda), \ldots$ approximating in an optimal way (in the sense of Gauss quadrature, i.e. $\omega^{(l)}(\lambda)$ ensures that for any polynomial of degree less than of equal to $2l - 1$ the value of the original integral (2.16) is approximated by (2.21) exactly) the original distribution function $\omega(\lambda)$, cf. [26], [46, Chapter XV], [45].

All this is well-known. Gauss quadrature represents a classical textbook material and the connection of CG to Gauss quadrature was pointed out in the original paper [24]. This connection is, however, a key to understanding both mathematical properties and finite precision behaviour of the CG method.

Given A and r_0, (2.16) and its Gauss quadrature approximations (2.21) are for $j = 1, 2, \ldots, n$ uniquely determined (remember we assumed that the eigenvalues of A are positive and distinct). Conversely, the distribution function $\omega^{(j)}(\lambda)$ uniquely determines the symmetric tridiagonal matrix T_j, and, through (2.7) and (2.6), the CG approximation x_j. With $f(\lambda) = \lambda^{-1}$ we have from (2.10)

$$\|x - x_0\|_A^2 = \|r_0\|^2 \sum_{i=1}^{n} \frac{\omega_i}{\lambda_i} = \|r_0\|^2 \int_\zeta^\xi \lambda^{-1}\, d\omega(\lambda)\,, \tag{2.22}$$

and, using (2.3) with $j = n$,

$$\|x - x_0\|_A^2 = (r_0, A^{-1} r_0) = \|r_0\|^2 (e_1, T_n^{-1} e_1) \equiv \|r_0\|^2\, (T_n^{-1})_{11}\,.$$

Consequently,

$$\int_\zeta^\xi \lambda^{-1}\, d\omega(\lambda) = (T_n^{-1})_{11}\,. \tag{2.23}$$

Repeating the same considerations using the CG method for T_j with the initial residual $\|r_0\| e_1$, or the Lanczos method for T_j with e_1

$$\int_\zeta^\xi \lambda^{-1}\, d\omega^{(j)}(\lambda) = (T_j^{-1})_{11}\,. \tag{2.24}$$

Finally, applying the j-point Gauss quadrature to (2.16) gives

$$\int_\zeta^\xi f(\lambda)\, d\omega(\lambda) = \int_\zeta^\xi f(\lambda)\, d\omega^{(j)}(\lambda) + R_j(f), \tag{2.25}$$

where $R_j(f)$ stands for the (truncation) error in the Gauss quadrature. In the next section we present several different ways of expressing (2.25) with $f(\lambda) = \lambda^{-1}$.

3. Basic Identities. Multiplying the identity (2.25) by $\|r_0\|^2$ gives

$$\|r_0\|^2 \int_\zeta^\xi f(\lambda)\, d\omega(\lambda) \;=\; \|r_0\|^2 \int_\zeta^\xi f(\lambda)\, d\omega^{(j)}(\lambda) + \|r_0\|^2 R_j(f). \tag{3.1}$$

Using (2.22), (2.23) and (2.24), (3.1) can for $f(\lambda) = \lambda^{-1}$ be written as

$$\|x - x_0\|_A^2 \;=\; \|r_0\|^2 (T_n^{-1})_{11} \;=\; \|r_0\|^2 (T_j^{-1})_{11} + \|r_0\|^2 R_j(\lambda^{-1})\,.$$

In [17, pp. 253-254] it was proved that for $f(\lambda) = \lambda^{-1}$ the truncation error in the Gauss quadrature is equal to

$$R_j(\lambda^{-1}) = \frac{\|x - x_j\|_A^2}{\|r_0\|^2}\,,$$

which gives

$$\|x - x_0\|_A^2 \;=\; \|r_0\|^2 (T_j^{-1})_{11} + \|x - x_j\|_A^2 \,. \tag{3.2}$$

Summarizing, the value of the j-th Gauss quadrature approximation to the integral (2.23) is the complement of the error in the j-th CG iteration measured by $\|x - x_j\|_A^2 / \|r_0\|^2$,

$$\frac{\|x - x_0\|_A^2}{\|r_0\|^2} \;=\; j\text{-point Gauss quadrature} \;+\; \frac{\|x - x_j\|_A^2}{\|r_0\|^2}. \tag{3.3}$$

This relation was developed in [8] in the context of moments; it was a subject of extensive work motivated by estimation of the error norms in CG in the papers [12], [15] and [17]. Work in this direction continued and led to the papers [16], [28], [30], [5].

An interesting form of (3.2) was noticed by Warnick in [47]. In the papers mentioned above the values of $\|x - x_0\|_A^2 / \|r_0\|^2 = (T_n^{-1})_{11}$ and $(T_j^{-1})_{11}$ were approximated from the actual Gauss quadrature calculations (or from the related recurrence relations). Using (2.7) and (2.6), the identities

$$\begin{aligned}\|r_0\|^2 (T_j^{-1})_{11} &= \|r_0\| \, e_1^T \, T_j^{-1} \, e_1 \|r_0\| \\ &= \|r_0\| \, v_1^T V_j \, T_j^{-1} e_1 \|r_0\| = (\|r_0\| v_1)^T \left(V_j T_j^{-1} e_1 \|r_0\|\right) \\ &= r_0^T (x_j - x_0)\end{aligned}$$

show that $(T_j^{-1})_{11}$ is given by a simple inner product. Indeed,

$$\|x - x_0\|_A^2 \;=\; r_0^T (x_j - x_0) + \|x - x_j\|_A^2 \,. \tag{3.4}$$

This remarkable identity was pointed out to us by Saylor [41], [40]. Please note that derivation of the identity (3.4) from the Gauss quadrature-based (3.2) uses the orthogonality relation $v_1^T V_j = e_1$. In finite precision computations this orthogonality relation does not hold. Consequently, (3.4) does not hold in finite precision arithmetic. We will return to this point in Section 6.

A mathematically equivalent identity can be derived by simple algebraic manipulations without using Gauss quadrature,

$$\begin{aligned}(x - x_0)^T A (x - x_0) &= (x - x_j + x_j - x_0)^T A (x - x_0) \\ &= (x - x_j)^T A (x - x_0) + (x_j - x_0)^T A (x - x_0) \\ &= (x - x_j)^T A (x - x_j + x_j - x_0) + (x_j - x_0)^T r_0 \\ &= \|x - x_j\|_A^2 + (x - x_j)^T A (x_j - x_0) + r_0^T (x_j - x_0) \\ &= \|x - x_j\|_A^2 + r_j^T (x_j - x_0) + r_0^T (x_j - x_0),\end{aligned}$$

hence

$$\|x - x_0\|_A^2 \;=\; r_j^T (x_j - x_0) + r_0^T (x_j - x_0) + \|x - x_j\|_A^2. \tag{3.5}$$

The right-hand side of (3.5) contains, in comparison with (3.4), the additional term $r_j^T (x_j - x_0)$. This term is in exact arithmetic equal to zero, but it has an important correction effect in finite precision computations (see Section 6).

Relations (3.2), (3.4) and (3.5) represent various mathematically equivalent forms of (3.1). While in (3.2) the j-point Gauss quadrature is evaluated as $(T_j^{-1})_{11}$, in (3.4) and (3.5) this quantity is computed using inner products of the vectors that are at our disposal during the iteration process. But, as mentioned in Introduction, there is much simpler identity (1.5) mathematically equivalent to (3.1). It is very surprising that, though (1.5) is present in the Hestenes and Stiefel paper [24, Theorem 6.1, relation (6:2), p. 416], this identity has (at least to our knowledge) never been related to Gauss quadrature. Its derivation is very simple. Using (1.3)

$$\begin{aligned}\|x-x_i\|_A^2 - \|x-x_{i+1}\|_A^2 &= \|x-x_{i+1}+x_{i+1}-x_i\|_A^2 - \|x-x_{i+1}\|_A^2\\ &= \|x_{i+1}-x_i\|_A^2 + 2(x-x_{i+1})^T A(x_{i+1}-x_i)\\ &= \gamma_i^2\, p_i^T A p_i + 2r_{i+1}^T(x_{i+1}-x_i)\\ &= \gamma_i \|r_i\|^2 \,. \qquad (3.6)\end{aligned}$$

Consequently, for $0 \le l < j \le n$,

$$\|x-x_l\|_A^2 - \|x-x_j\|_A^2 = \sum_{i=l}^{j-1}\left(\|x-x_i\|_A^2 - \|x-x_{i+1}\|_A^2\right) = \sum_{i=l}^{j-1}\gamma_i\|r_i\|^2, \qquad (3.7)$$

and (3.1) can be written in the form

$$\|x-x_0\|_A^2 \;=\; \sum_{i=0}^{j-1}\gamma_i\|r_i\|^2 + \|x-x_j\|_A^2. \qquad (3.8)$$

The numbers $\gamma_i\|r_i\|^2$ are trivially computable; both γ_i and $\|r_i\|^2$ are available at every iteration step. Please note that in the derivation of (3.7) we used the local orthogonality among the consecutive residuals and direction vectors only. We avoided using mutual orthogonality among the vectors with generally different indices. This fact will be very important in the rounding error analysis of the finite precision counterparts of (3.7) in Sections 7–10.

4. Estimating the A-norm of the error. Using $\|x-x_0\|_A^2 = \|r_0\|^2(T_n^{-1})_{11}$, (3.2) is written in the form

$$\|x-x_j\|_A^2 \;=\; \|r_0\|^2\left[(T_n^{-1})_{11} - (T_j^{-1})_{11}\right] .$$

As suggested in [17, pp. 28–29], the unknown value $(T_n^{-1})_{11}$ can be replaced, at a price of $m-j$ extra steps, by a computable value $(T_m^{-1})_{11}$ for some $m > j$. The paper [17], however, did not properly use this idea and did not give a proper formula for computing the difference $(T_m^{-1})_{11} - (T_j^{-1})_{11}$ without cancellation, which limited the applicability of the proposed result. Golub and Meurant cleverly resolved this trouble in [16] and proposed an algorithm for estimating the A-norm of the error in the CG method called CGQL. This section will briefly summarize several important estimates.

Consider, in general, (3.1) for j and $j+d$, where d is some positive integer. The idea is simply to eliminate the unknown term $\int_\zeta^\xi f(\lambda)\,d\omega(\lambda)$ by subtracting the identities for j and $j+d$ which results in

$$\|r_0\|^2 R_j(f) = \|r_0\|^2\left(\int_\zeta^\xi f(\lambda)\,d\omega^{(j+d)}(\lambda) - \int_\zeta^\xi f(\lambda)\,d\omega^{(j)}(\lambda)\right) + \|r_0\|^2 R_{j+d}(f).$$

In particular, using (3.2), (3.4), (3.5), and (3.8) we obtain the mathematically equivalent identities

$$\|x-x_j\|_A^2 \;=\; \|r_0\|^2\left[(T_{j+d}^{-1})_{11} - (T_j^{-1})_{11}\right] + \|x-x_{j+d}\|_A^2\,, \qquad (4.1)$$

$$\|x-x_j\|_A^2 \;=\; r_0^T(x_{j+d}-x_j) + \|x-x_{j+d}\|_A^2\;, \qquad (4.2)$$

$$\begin{aligned}\|x-x_j\|_A^2 \;=\; & r_0^T(x_{j+d}-x_j) - r_j^T(x_j-x_0) + r_{j+d}^T(x_{j+d}-x_0)\\ & + \|x-x_{j+d}\|_A^2\;, \qquad (4.3)\end{aligned}$$

and

$$\|x - x_j\|_A^2 \;=\; \sum_{i=j}^{j+d-1} \gamma_i \|r_i\|^2 + \|x - x_{j+d}\|_A^2\,. \tag{4.4}$$

Now recall that the A-norm of the error is in the CG method strictly decreasing. If d is chosen such that

$$\|x - x_j\|_A^2 \gg \|x - x_{j+d}\|_A^2\,, \tag{4.5}$$

then neglecting $\|x - x_{j+d}\|_A^2$ on the right-hand sides of (4.1), (4.2), (4.3) and (4.4) gives lower bounds (all mathematically equal) for the squared A-norm of the error in the j-th step. Under the assumption (4.5) these bounds are reasonably tight (their inaccuracy is given by $\|x - x_{j+d}\|_A^2$). We denote them

$$\eta_{j,d} \;=\; \|r_0\|^2\;[(T_{j+d}^{-1})_{11} - (T_j^{-1})_{11}], \tag{4.6}$$

where the difference $(T_{j+d}^{-1})_{11} - (T_j^{-1})_{11}$ is computed by the algorithm CGQL from [16],

$$\mu_{j,d} \;=\; r_0^T(x_{j+d} - x_j), \tag{4.7}$$

which refers to the original bound due to Warnick,

$$\vartheta_{j,d} \;=\; r_0^T(x_{j+d} - x_j) - r_j^T(x_j - x_0) + r_{j+d}^T(x_{j+d} - x_0), \tag{4.8}$$

which is the previous bound modified by the correction terms and

$$\nu_{j,d} \;=\; \sum_{i=j}^{j+d-1} \gamma_i \|r_i\|^2. \tag{4.9}$$

Clearly, the last bound, which is a direct consequence of [24, Theorem 6:1], see (1.5), is much simpler than the others.

Mathematically (in exact arithmetic)

$$\eta_{j,d} = \mu_{j,d} = \vartheta_{j,d} = \nu_{j,d}\;. \tag{4.10}$$

In finite precision computations (4.10) does not hold in general, and the different bounds may give substantially different results. *Does any of the identities (4.1)–(4.4) have any relevance for the quantities computed in finite precision arithmetic?* The work described in this subsection and the papers published on this subject would be of little practical use without answering this question.

5. Delay of convergence. For more than 20 years the effects of rounding errors to the Lanczos and CG methods seemed devastating. Orthogonality among the computed vectors $v_1, v_2, \ldots$ was usually lost very quickly, with a subsequent loss of linear independence. Consequently, the finite termination property was lost. Still, despite a total loss of orthogonality among the vectors in the Lanczos sequence $v_1, v_2, \ldots,$ and despite a possible regular appearance of Lanczos vectors which were linearly dependent on the vectors computed in preceding iterations, the Lanczos and the CG methods produced reasonable results.

A fundamental work which brought light into this darkness was done by Paige. He proved that loss of orthogonality among the computed Lanczos vectors $v_1, v_2, \ldots$ was possible only in the directions of the converged Ritz vectors $z_l^{(j)} \equiv V_j s_l^{(j)}$. For more details see [33], [34], [35], [36], the review paper [44, Section 3.1] and the works quoted there (in particular [38], [39], [32] and [45]). Little was known about rounding errors in the Krylov subspace methods before the Ph.D. thesis of Paige [33], and almost all results (with the exception of works on ultimate attainable accuracy) published on the subject after this thesis and the papers [34], [35], [36] were based on them.

Another step, which can compete in originality with that of Paige, was made by Greenbaum in [19]. If CG is used to solve a linear symmetric positive definite system $Ax = b$ on a computer with machine precision ε, then [19] shows that the A-norms of the errors $\|x - x_l\|_A$, $l = 1, 2, \ldots, j$ are very close to the $\overline{A}$-norms of the errors $\|\overline{x} - \overline{x}_l\|_{\overline{A}}$, $l = 1, 2, \ldots, j$ determined by the *exact* CG applied to some particular symmetric positive definite system $\overline{A}(j)\overline{x}(j) = \overline{b}(j)$ (see [19, Theorem 3, pp. 26-27]). This system and the initial approximation $\overline{x}_0(j)$ depend on the iteration step j. The matrix $\overline{A}(j)$ is larger than the matrix A. Its eigenvalues must lie in tiny intervals about the eigenvalues of A, and there must be at least one eigenvalue of $\overline{A}(j)$ close to each eigenvalue of A (the last result was proved in [43]). Moreover, for each eigenvalue λ_i of A, $i = 1, \ldots, n$ (similarly to Section 2 we assume, with no loss of generality, that the eigenvalues of A are distinct), the weight $\omega_i = (v_1, u_i)^2$ closely approximates the sum of weights corresponding to the eigenvalues of $\overline{A}(j)$ clustered around λ_i (see [19, relation (8.21) on p. 60]).

The quantitative formulations of the relationships between A, b, x_0 and $\overline{A}(j)$, $\overline{b}(j)$, $\overline{x}_0(j)$ contains some terms related in various complicated ways to machine precision ε (see [19], [43] and [17, Theorems 5.1–5.3 and the related discussion on pp. 257–260]). The actual size of the terms given in the quoted papers documents much more difficulties of handling accurately peculiar technical problems of rounding error analysis than it says about the accuracy of the described relationships. The fundamental concept to which the (very often weak) rounding error bounds lead should be read: the first j steps of a *finite precision* CG computation for $Ax = b$ can be viewed as the first j steps of the *exact* CG computation for some particular $\overline{A}(j)\overline{x}(j) = \overline{b}(j)$. This relationship was developed and proved theoretically. Numerical experiments show that its tightness is much better than the technically complicated theoretical calculations in [19] would suggest. We will not continue with describing the results of the subsequent work [22]. We do not need it here. Moreover, a rigorous theoretical description of the model from [22] in the language of Riemann-Stieltjes integral and Gauss quadrature still needs some clarification. We hope to return to that subject elsewhere.

As a consequence of the loss of orthogonality caused by rounding errors, convergence of the CG method is delayed. In order to illustrate this important point numerically, we plot in Fig. 5.1 results of the CG method (1.3) for the matrix $A = Q\Lambda Q^T$, where Q is the orthogonal matrix obtained from the Matlab QR-decomposition of the randomly generated matrix (computed by the Matlab command `randn(n)`), and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix with the eigenvalues

$$\lambda_i = \lambda_1 + \frac{i-1}{n-1}(\lambda_n - \lambda_1)\, \rho^{n-i}, \quad i = 2, \ldots, n-1, \tag{5.1}$$

see [43]. We have used $n = 48$, $\lambda_1 = 0.1$, $\lambda_n = 1000$, $\rho = 0.9$, $x = (1, \ldots, 1)^T$, $b = Ax$, and $x_0 = (0, \ldots, 0)^T$. We have simulated the exact arithmetic values by double reorthogonalization of the residual vectors (see [22]). The quantities obtained from the CG implementation with the double reorthogonalized residuals will be denoted by (E). Fig. 5.1 shows that when the double reorthogonalization is applied, the corresponding A-norm of the error (dash-dotted line) can be very different from the A-norm of the error of the ordinary finite precision (FP) CG implementation (solid line). Without reorthogonalization, the orthogonality among the (FP) Lanczos vectors, measured by the Frobenius norm $\|I - V_j^T V_j\|_F$ (dotted line), is

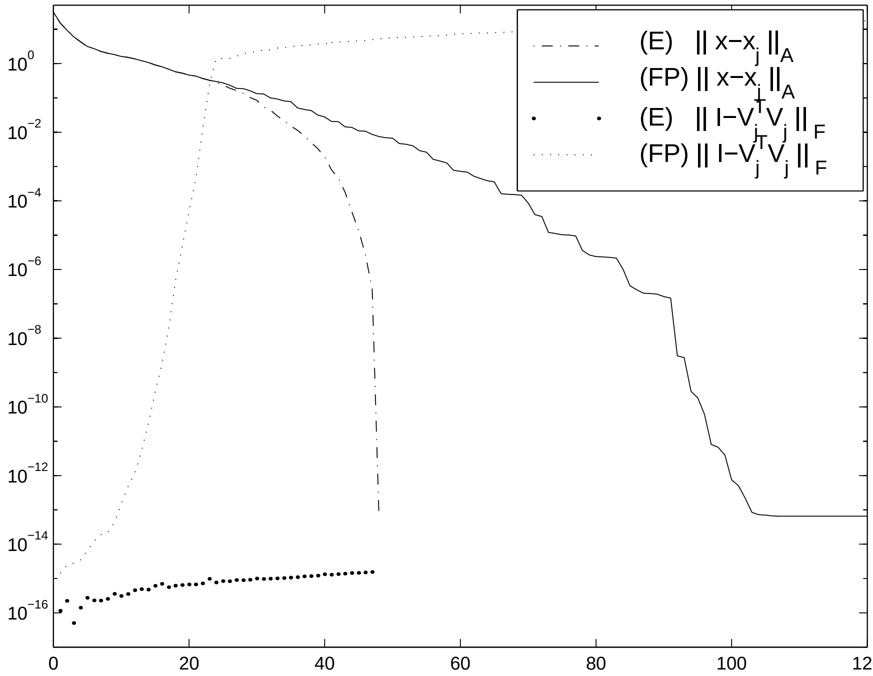


FIG. 5.1. *The A-norm of the error for the CG implementation with the double reorthogonalized residuals* (E) (*dashed-dotted line*) *is compared to the A-norm of the error of the ordinary finite precision CG implementation* (FP) (*solid line*)*. The corresponding loss of orthogonality among the normalized residuals is plotted by the dots resp. the dotted line.*

lost after a few iterations. With double reorthogonalization the orthogonality is kept close to machine precision (dots). Experiments were performed using Matlab 5.1 on a personal computer with machine precision $\varepsilon \sim 10^{-16}$.

We see that the delay of convergence due to loss of orthogonality can be very substantial. Consider now application of the estimates (4.6)–(4.9) to finite precision computations. In derivation of all these estimates we assumed exact arithmetic. Consequently, in these derivations we did not count for any loss of orthogonality and delay of convergence. For the example presented above, the bounds can therefore be expected to give good results for the double reorthogonalized CG (dash-dotted convergence curve). Should they give anything reasonable also for the ordinary (FP) CG implementation (solid convergence curve)? If yes, then why? The following section explains that this question is of fundamental importance.

6. Examples.

Indeed, without a proper rounding error analysis of the identities (4.1)–(4.4) there is no justification that the estimates derived assuming exact arithmetic will work in finite precision arithmetic. For example, when the significant loss of orthogonality occurs, the bound $\mu_{j,d}$ given by (4.7) does not work!

This fact is demonstrated in Fig. 6.1 which presents experimental results for the problem described in the previous section (see Fig. 5.1). It plots the computed estimate $|\mu_{j,d}|^{1/2}$ (dashed line) and demonstrates the importance of the correction term

$$c_{j,d} = -r_j^T(x_j - x_0) + r_{j+d}^T(x_{j+d} - x_0), \tag{6.1}$$

($|c_{j,d}|^{1/2}$ is plotted by dots). Fig. 6.1 shows clearly that when the global orthogonality (measured by $\|I - V_j^T V_j\|_F$ and plotted by a dotted line) grows greater than $\|x - x_j\|_A$ (solid

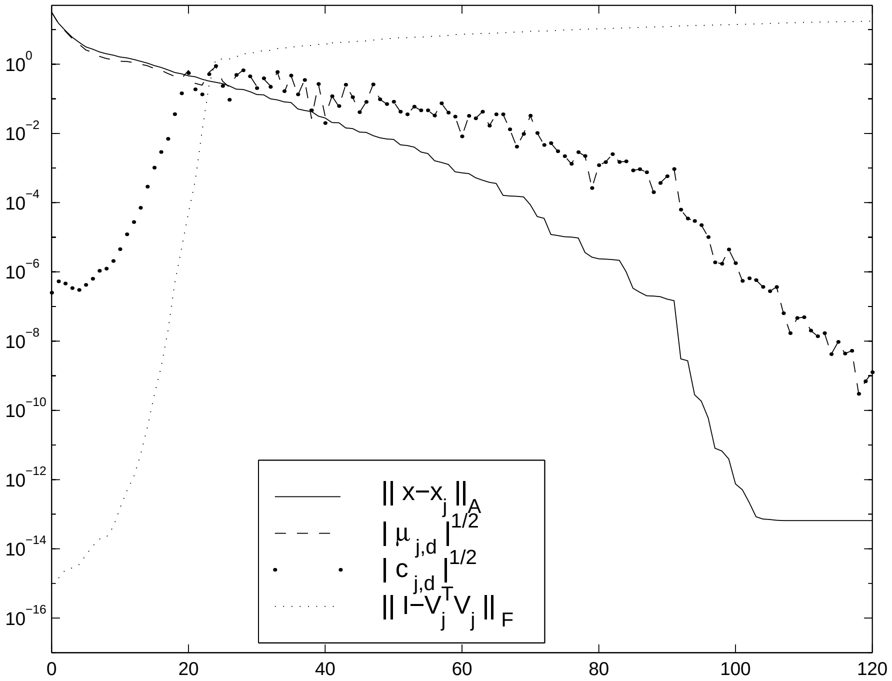


FIG. 6.1. *Error estimate $\mu_{j,d}^{1/2}$ can fail. The computed estimate $|\mu_{j,d}|^{1/2}$ (dashed line) for the A-norm of the error (solid line) gives useful information about convergence only until the loss of orthogonality (dotted line) crosses the convergence curve. After that point $\mu_{j,d}$ can even become negative, and must be modified by adding the correction term $c_{j,d}$ ($|c_{j,d}|^{1/2}$ is plotted by dots). We used $d = 4$.*

line), the bound ${\mu_{j,d}}^{1/2}$, which is based on global orthogonality, ceases to give any useful information about convergence ($\mu_{j,d}$ may even become negative, therefore we plot the second root of its absolute value). Adding the correction term $c_{j,d}$ to $\mu_{j,d}$ gives $\vartheta_{j,d}$, see (4.8), which gives estimates comparable to $\eta_{j,d}$ and $\nu_{j,d}$ (see Section 11). In this experiment we used $d = 4$.

It is important to understand that the additional rounding errors in computing $\eta_{j,d}$ $\mu_{j,d}$, $\vartheta_{j,d}$ and $\nu_{j,d}$ from the given formulas (the algorithm CGQL and (4.7)–(4.9)) do not affect significantly the values of the computed bounds and do not represent a problem. The problem is in the fact, that when the orthogonality is significantly lost, the input quantities used in the algorithm CGQL and in the formulas (4.7)–(4.9) are significantly different from their exact precision counterparts. These quantities affected by the loss of orthogonality are plugged into the formulas which assume, in their derivation, exact orthogonality.

In order to stress the previous point and to underline the necessity of rounding error analysis of the identities (4.7)–(4.9), we present the following analogous example. In the Lanczos method the eigenvalues $\theta_1^{(j)} < \theta_2^{(j)} < \ldots < \theta_j^{(j)}$ of T_j (Ritz values) are considered approximations to the eigenvalues of the matrix A (see Section 2). Let $\theta_l^{(j)}$, $z_l^{(j)} = V_j s_l^{(j)}$ (where $s_l^{(j)}$ is the normalized eigenvector of T_j corresponding to $\theta_l^{(j)}$) represents an approximate eigenpair of A. In exact arithmetic we have the following bound for the distance of $\theta_l^{(j)}$ to the nearest eigenvalue of A

$$\min_i |\lambda_i - \theta_l^{(j)}| \le \frac{\|Az_l^{(j)} - \theta_l^{(j)} z_l^{(j)}\|}{\|z_l^{(j)}\|} = \|Az_l^{(j)} - \theta_l^{(j)} z_l^{(j)}\|, \tag{6.2}$$

where $\|z_l^{(j)}\| = 1$ due to the orthonormality of the Lanczos vectors $v_1, \ldots, v_j$. Using (2.3), $\|Az_l^{(j)} - \theta_l^{(j)} z_l^{(j)}\| = \beta_{j+1}(e_j, s_l^{(j)})$, which gives

$$\min_i |\lambda_i - \theta_l^{(j)}| \;\leq\; \beta_{j+1}(e_j, s_l^{(j)}) \equiv \delta_{lj}, \tag{6.3}$$

see, e.g., [38], [36]. Consequently, in exact arithmetic, if δ_{lj} is small, then $\theta_l^{(j)}$ must be close to some λ_i. In finite precision arithmetic loss of orthogonality has, among the others, a very unpleasant effect: we cannot guarantee, in general, that $z_l^{(j)}$, which is a linear combination of $v_1, \ldots, v_j$ has a nonvanishing norm. We can still compute δ_{lj} from β_{j+1} and T_j; the effect of rounding errors in this additional computation is negligible. We can therefore say, similarly to the analogous statements published about computation of the convergence estimates in the CG method, that δ_{lj} is in the presence of rounding errors computed "accurately". Does δ_{lj} computed in finite precision arithmetic tell anything about convergence of $\theta_l^{(j)}$ to some λ_i? Yes, it does! But this affirmative answer is based neither on the exact precision formulas (6.2) and (6.3), nor on the fact that δ_{lj} is computed "accurately". It is based on an ingenious analysis due to Paige, who have shown that the orthogonality can be lost in the directions of the well approximated eigenvectors only. For the complicated details of this difficult result we refer to [33], [37] and to the summary given in [44, Theorem 2]. We see that even in finite precision computations small δ_{lj} guarantees that $\theta_l^{(j)}$ approximates some λ_i to high accuracy. It is very clear, however, that this conclusion is the result of the rounding error analysis of the Lanczos method given by Paige, and no similar statement could be made without this analysis.

In the following three sections we present rounding error analysis of the bound $\nu_{j,d}$ given by (4.4) and (4.9). We concentrate on $\nu_{j,d}$ because it is the simplest of all the others. If $\nu_{j,d}$ is proved numerically stable, then there is a small reason for using the other bounds $\eta_{j,d}$ or $\vartheta_{j,d}$ in practical computations.

7. Finite precision CG computations. In the analysis we assume the standard model of floating point arithmetic with machine precision ε, see, e.g. [25, (2.4)],

$$\text{fl}[a \circ b] = (a \circ b)(1 + \delta), \qquad |\delta| \leq \varepsilon, \tag{7.1}$$

where a and b stands for floating-point numbers and the symbol $\circ$ stands for the operations addition, subtraction, multiplication and division. We assume that this model holds also for the square root operation. Under this model, we have for operations involving vectors v, w, a scalar α and the matrix A the following standard results [18], see also [20], [35]

$$\|\alpha\, v - \text{fl}[\alpha\, v]\| \leq \varepsilon\, \|\alpha\, v\|, \tag{7.2}$$

$$\|v + w - \text{fl}[v + w]\| \leq \varepsilon\, (\|v\| + \|w\|), \tag{7.3}$$

$$|(v, w) - \text{fl}[(v, w)]| \leq \varepsilon\, n\, (1 + O(\varepsilon))\, \|v\|\, \|w\|, \tag{7.4}$$

$$\|Av - \text{fl}[Av]\| \leq \varepsilon\, c\, \|A\| \|v\|. \tag{7.5}$$

When A is a matrix with at most h nonzeros in any row and if the matrix-vector product is computed in the standard way, $c = hn^{1/2}$. In the following analysis we count only for the terms linear in the machine precision epsilon ε and express the higher order terms as $O(\varepsilon^2)$. By $O(const)$ where $const$ is different from ε^2 we denote $const$ multiplied by a bounded positive term of an insignificant size which is independent of the $const$ and of any other variables present in the bounds.

Numerically, the CG iterates satisfy

$$x_{j+1} = x_j + \gamma_j p_j + \varepsilon z_j^x, \tag{7.6}$$

$$r_{j+1} = r_j - \gamma_j A p_j + \varepsilon z_j^r, \tag{7.7}$$

$$p_{j+1} = r_{j+1} + \delta_{j+1} p_j + \varepsilon z_j^p, \tag{7.8}$$

where εz_j^x, εz_j^r and εz_j^p account for the local roundoff ($r_0 = b - Ax_0 - \varepsilon f_0$, $\varepsilon \|f_0\| \leq \varepsilon\{\|b\| + \|Ax_0\| + c\|A\|\|x_0\|\} + O(\varepsilon^2)$). The local roundoff can be bounded according to the standard results (7.2)–(7.5) in the following way

$$\varepsilon\, \|z_j^x\| \leq \varepsilon\, \{\|x_j\| + 2\, \|\gamma_j p_j\|\} + O(\varepsilon^2) \leq \varepsilon\, \{3\|x_j\| + 2\|x_{j+1}\|\} + O(\varepsilon^2), \tag{7.9}$$

$$\varepsilon\, \|z_j^r\| \leq \varepsilon\, \{\|r_j\| + 2\, \|\gamma_j A p_j\| + c\, \|A\|\|\gamma_j p_j\|\} + O(\varepsilon^2), \tag{7.10}$$

$$\varepsilon\, \|z_j^p\| \leq \varepsilon\, \{\|r_{j+1}\| + 2\, \|\delta_{j+1} p_j\|\} + O(\varepsilon^2) \leq \varepsilon\, \{3\|r_{j+1}\| + 2\|p_{j+1}\|\} + O(\varepsilon^2). \tag{7.11}$$

Similarly, the computed coefficients γ_j and δ_j satisfy

$$\gamma_j = \frac{\|r_j\|^2}{p_j^T A p_j} + \varepsilon \zeta_j^\gamma, \quad \delta_j = \frac{\|r_j\|^2}{\|r_{j-1}\|^2} + \varepsilon \zeta_j^\delta. \tag{7.12}$$

Assuming $n\varepsilon \ll 1$, the local roundoff $\varepsilon \zeta_j^\delta$ is bounded, according to (7.1) and (7.4), by

$$\varepsilon |\zeta_j^\delta| \leq \varepsilon\, \frac{\|r_j\|^2}{\|r_{j-1}\|^2}\, O(n) + O(\varepsilon^2). \tag{7.13}$$

Using (7.2)–(7.5) and $\|A\|\|p_j\|^2/(p_j, Ap_j) \leq \kappa(A)$,

$$\begin{aligned} \mathrm{fl}[(p_j, Ap_j)] &= (p_j, Ap_j) + \varepsilon\, \|Ap_j\|\|p_j\| O(n) + \varepsilon\, \|A\|\|p_j\|^2 O(c) + O(\varepsilon^2) \\ &= (p_j, Ap_j)\big(1 + \varepsilon\, \kappa(A) O(n + c)\big) + O(\varepsilon^2). \end{aligned}$$

Assuming $\varepsilon(n + c)\, \kappa(A) \ll 1$, the local roundoff $\varepsilon \zeta_j^\gamma$ is bounded by

$$\varepsilon |\zeta_j^\gamma| \leq \varepsilon\, \kappa(A) \frac{\|r_j\|^2}{(p_j, Ap_j)}\, O(n + c) + O(\varepsilon^2). \tag{7.14}$$

It is well-known that in finite precision arithmetic the true residual $b - Ax_j$ differs from the recursively updated residual vector r_j,

$$r_j = b - Ax_j - \varepsilon f_j. \tag{7.15}$$

This topic was studied in [42] and [20]. The results can be written in the following form

$$\|\varepsilon f_j\| \leq \varepsilon\, \|A\|\, (\|x\| + \max_{0 \leq i \leq j} \|x_i\|)\, O(jc), \tag{7.16}$$

$$\|r_j\| = \|b - Ax_j\|\, (1 + \varepsilon F_j), \tag{7.17}$$

where εF_j is bounded by

$$|\varepsilon F_j| = \frac{|\|r_j\| - \|b - Ax_j\||}{\|b - Ax_j\|} \leq \frac{\|r_j - (b - Ax_j)\|}{\|b - Ax_j\|} = \frac{\varepsilon \|f_j\|}{\|b - Ax_j\|}. \tag{7.18}$$

Rounding errors affect results of CG computations in two main ways: they delay convergence (see Section 5) and limit the ultimate attainable accuracy. Here we are primarily interested in estimating the convergence rate. We therefore assume that the final accuracy level has not been reached yet and εf_j is, in comparison to the size of the true and iterative residuals, small. In the subsequent text we will relate the numerical inaccuracies to the

A-norm of the error $\|x - x_j\|_A$. The following inequalities derived from (7.18) will prove useful,

$$\lambda_1^{1/2} \, \|x - x_j\|_A \, (1 + \varepsilon \, F_j) \le \|r_j\| \le \lambda_n^{1/2} \, \|x - x_j\|_A \, (1 + \varepsilon \, F_j). \tag{7.19}$$

The monotonicity of the A-norm and of the Euclidean norm of the error is in CG preserved (with small additional inaccuracy) also in finite precision computations (see [19], [22]). Using this fact we get for $j \ge i$

$$\varepsilon \, \frac{\|r_j\|}{\|r_i\|} \le \varepsilon \, \frac{\lambda_n^{1/2}}{\lambda_1^{1/2}} \cdot \frac{\|x - x_j\|_A}{\|x - x_i\|_A} \cdot \frac{(1 + \varepsilon \, F_j)}{(1 + \varepsilon \, F_i)} \le \varepsilon \, \kappa(A)^{1/2} + O(\varepsilon^2). \tag{7.20}$$

This bound will be used later.

8. Finite precision analysis – basic identity. The bounds (4.6)–(4.9) are mathematically equivalent. We will concentrate on the simplest one given by $\nu_{j,d}$ (4.9) and prove that it gives (up to a small term) correct estimates also in finite precision computations. In particular, we prove that the ideal (exact precision) identity (4.4) changes numerically to

$$\|x - x_j\|_A^2 = \nu_{j,d} + \|x - x_{j+d}\|_A^2 + \widetilde{\nu}_{j,d}, \tag{8.1}$$

where $\widetilde{\nu}_{j,d}$ is as small as it can be (the analysis here will lead to much stronger results than the analysis of the finite precision counterpart of (4.1) given in [17]). Please note that the difference between (4.4) and (8.1) *is not trivial*. The ideal and numerical counterparts of each individual term in these identities may be orders of magnitude different! Due to the facts that rounding errors in computing $\nu_{j,d}$ numerically from the quantities γ_i, r_i are negligible and that $\widetilde{\nu}_{j,d}$ will be related to $\varepsilon \, \|x - x_j\|_A$, (8.1) will justify the estimate $\nu_{j,d}$ in finite precision computations.

From the identity for the numerically computed approximate solution

$$\begin{aligned} \|x - x_j\|_A^2 &= \|x - x_{j+1} + x_{j+1} - x_j\|_A^2 \\ &= \|x - x_{j+1}\|_A^2 + 2\,(x - x_{j+1})^T A (x_{j+1} - x_j) + \|x_{j+1} - x_j\|_A^2, \end{aligned}$$

we obtain easily

$$\|x - x_j\|_A^2 - \|x - x_{j+1}\|_A^2 = \|x_{j+1} - x_j\|_A^2 + 2\,(x - x_{j+1})^T A (x_{j+1} - x_j). \tag{8.2}$$

Please note that (8.2) represents an identity for the computed quantities. In order to get the desired form leading to (8.1), we will develop the right hand side of (8.2). In this derivation we will rely on local properties of the finite precision CG recurrences (7.6)–(7.8) and (7.12).

Using (7.6), the first term on the right hand side of (8.2) can be written as

$$\begin{aligned} \|x_{j+1} - x_j\|_A^2 &= (\gamma_j p_j + \varepsilon \, z_j^x)^T A (\gamma_j p_j + \varepsilon \, z_j^x) \\ &= \gamma_j^2 \, p_j^T A p_j + 2\varepsilon \, \gamma_j p_j^T A z_j^x + O(\varepsilon^2) \\ &= \gamma_j^2 \, p_j^T A p_j + 2\varepsilon \, (x_{j+1} - x_j)^T A z_j^x + O(\varepsilon^2). \end{aligned} \tag{8.3}$$

Similarly, the second term on the right hand side of (8.2) transforms, using (7.15), to the form

$$\begin{aligned} 2\,(x - x_{j+1})^T A (x_{j+1} - x_j) &= 2\,(r_{j+1} + \varepsilon \, f_{j+1})^T (x_{j+1} - x_j) \\ &= 2\, r_{j+1}^T (x_{j+1} - x_j) + 2\varepsilon \, f_{j+1}^T (x_{j+1} - x_j). \end{aligned} \tag{8.4}$$

Combining (8.2), (8.3) and (8.4),

$$\begin{aligned} \|x - x_j\|_A^2 - \|x - x_{j+1}\|_A^2 &= \gamma_j^2\, p_j^T A p_j + 2\, r_{j+1}^T (x_{j+1} - x_j) \\ &\quad + 2\varepsilon\, (f_{j+1} + A z_j^x)^T (x_{j+1} - x_j) + O(\varepsilon^2). \end{aligned} \tag{8.5}$$

Substituting for γ_j from (7.12), the first term in (8.5) can be written as

$$\gamma_j^2\, p_j^T A p_j = \gamma_j \|r_j\|^2 + \varepsilon\, \gamma_j\, p_j^T A p_j\, \zeta_j^\gamma = \gamma_j \|r_j\|^2 + \varepsilon\, \gamma_j \|r_j\|^2 \left\{ \zeta_j^\gamma \frac{p_j^T A p_j}{\|r_j\|^2} \right\}.$$

Consequently, the difference between the squared A-norms of the error in the consecutive steps can be written in the form convenient for the further analysis

$$\begin{aligned} \|x - x_j\|_A^2 - \|x - x_{j+1}\|_A^2 &= \gamma_j \|r_j\|^2 + \varepsilon\, \gamma_j \|r_j\|^2 \left\{ \zeta_j^\gamma \frac{p_j^T A p_j}{\|r_j\|^2} \right\} \\ &\quad + 2\, r_{j+1}^T (x_{j+1} - x_j) \\ &\quad + 2\varepsilon\, (f_{j+1} + A z_j^x)^T (x_{j+1} - x_j) + O(\varepsilon^2). \end{aligned} \tag{8.6}$$

The goal of the following analysis is to show that until $\|x - x_j\|_A$ reaches its ultimate attainable accuracy level, the terms on the right hand side of (8.6) are, except for $\gamma_j \|r_j\|^2$, insignificant. Bounding the second term will not represent a problem. The norm of the difference $x_{j+1} - x_j = (x - x_j) - (x - x_{j+1})$ is bounded by $2\|x - x_j\|_A / \lambda_1^{1/2}$. Therefore the size of the fourth term is proportional to $\varepsilon\, \|x - x_j\|_A$. The third term is related to the line-search principle. Ideally (in exact arithmetic), the $(j+1)$-th residual is orthogonal to the difference between the $(j+1)$-th and j-th approximation (which is a multiple of the j-th direction vector). This is equivalent to the line-search: ideally the $(j+1)$-th CG approximation minimizes the A-norm of the error along the line determined by the j-th approximation and the j-th direction vector. Here the term $r_{j+1}^T (x_{j+1} - x_j)$, with r_{j+1}, x_j and x_{j+1} computed numerically, examines how closely the line-search holds in finite precision arithmetic. In fact, bounding the local orthogonality $r_{j+1}^T (x_{j+1} - x_j)$ represents the technically most difficult part of the remaining analysis.

9. Local orthogonality in the Hestenes and Stiefel implementation.

Since the classical work of Paige it is well-known that in the three-term Lanczos recurrence local orthogonality is preserved close to the machine epsilon (see [35]). We will derive an analogy of this for the CG algorithm, and state it as an independent result.

The local orthogonality term $r_{j+1}^T (x_{j+1} - x_j)$ can be written in the form

$$r_{j+1}^T (x_{j+1} - x_j) = r_{j+1}^T (\gamma_j p_j + \varepsilon\, z_j^x) = \gamma_j r_{j+1}^T p_j + \varepsilon\, r_{j+1}^T z_j^x. \tag{9.1}$$

Using the bound $\|r_{j+1}\| \le \lambda_n^{1/2} \|x - x_{j+1}\|_A (1 + \varepsilon\, F_{j+1}) \le \lambda_n^{1/2} \|x - x_j\|_A (1 + \varepsilon\, F_{j+1})$, see (7.19), the size of the second term in (9.1) is proportional to $\varepsilon\, \|x - x_j\|_A$. The main step consist of showing that the term $r_{j+1}^T p_j$ is sufficiently small. Multiplying the recurrence (7.7) for r_{j+1} by the column vector p_j^T gives (using (7.8) and (7.12))

$$\begin{aligned} p_j^T r_{j+1} &= p_j^T r_j - \gamma_j p_j^T A p_j + \varepsilon\, p_j^T z_j^r \\ &= (r_j + \delta_j p_{j-1} + \varepsilon\, z_{j-1}^p)^T r_j - \left(\frac{\|r_j\|^2}{p_j^T A p_j} + \varepsilon\, \zeta_j^\gamma \right) p_j^T A p_j + \varepsilon\, p_j^T z_j^r \\ &= \delta_j\, p_{j-1}^T r_j + \varepsilon\, \{ r_j^T z_{j-1}^p - \zeta_j^\gamma p_j^T A p_j + p_j^T z_j^r \}. \end{aligned} \tag{9.2}$$

Denoting

$$M_j \equiv r_j^T z_{j-1}^p - \zeta_j^\gamma p_j^T A p_j + p_j^T z_j^r, \tag{9.3}$$

the identity (9.2) is

$$p_j^T r_{j+1} = \delta_j\, p_{j-1}^T r_j + \varepsilon\, M_j. \tag{9.4}$$

Recursive application of (9.4) for $p_{j-1}^T r_j, \ldots, p_1^T r_2$ with $p_0^T r_1 = \|r_0\|^2 - \gamma_0\, p_0^T A p_0 + \varepsilon\, p_0^T z_0^r = \varepsilon\, \{-\zeta_0^\gamma r_0^T A r_0 + p_0^T z_0^r\} \equiv \varepsilon\, M_0$, gives

$$p_j^T r_{j+1} = \varepsilon\, M_j + \varepsilon \sum_{i=1}^{j} \left(\prod_{k=i}^{j} \delta_k \right) M_{i-1}. \tag{9.5}$$

Since

$$\varepsilon \prod_{k=i}^{j} \delta_k = \varepsilon \prod_{k=i}^{j} \frac{\|r_k\|^2}{\|r_{k-1}\|^2} + O(\varepsilon^2) = \varepsilon\, \frac{\|r_j\|^2}{\|r_{i-1}\|^2} + O(\varepsilon^2),$$

we can express (9.5) as

$$p_j^T r_{j+1} = \varepsilon\, \|r_j\|^2 \sum_{i=0}^{j} \frac{M_i}{\|r_i\|^2} + O(\varepsilon^2). \tag{9.6}$$

Using (9.3),

$$\frac{|M_i|}{\|r_i\|^2} \le \frac{\|z_{i-1}^p\|}{\|r_i\|} + |\zeta_i^\gamma| \frac{p_i^T A p_i}{\|r_i\|^2} + \frac{\|p_i\| \|z_i^r\|}{\|r_i\|^2}. \tag{9.7}$$

¿From (7.11) it follows

$$\varepsilon\, \frac{\|z_{i-1}^p\|}{\|r_i\|} \le \varepsilon \left\{ 3 + 2 \frac{\|p_i\|}{\|r_i\|} \right\} + O(\varepsilon^2). \tag{9.8}$$

Using (7.14),

$$\varepsilon\, |\zeta_i^\gamma| \frac{p_i^T A p_i}{\|r_i\|^2} \le \varepsilon\, \kappa(A)\, O(n + c) + O(\varepsilon^2). \tag{9.9}$$

The last part of (9.7) is bounded using (7.10) and (7.12)

$$\begin{aligned} \varepsilon\, \frac{\|p_i\| \|z_i^r\|}{\|r_i\|^2} &\le \varepsilon \left\{ \frac{\|p_i\| \|r_i\|}{\|r_i\|^2} + 2\, \gamma_i \frac{\|p_i\| \|A p_i\|}{\|r_i\|^2} + c\, \gamma_i \frac{\|p_i\| \|A\| \|p_i\|}{\|r_i\|^2} \right\} + O(\varepsilon^2) \\ &= \varepsilon \left\{ \frac{\|p_i\|}{\|r_i\|} + 2\, \frac{\|p_i\| \|A p_i\|}{p_i^T A p_i} + c\, \frac{\|A\| \|p_i\|^2}{p_i^T A p_i} \right\} + O(\varepsilon^2) \\ &\le \varepsilon \left\{ \frac{\|p_i\|}{\|r_i\|} + (2 + c)\, \kappa(A) \right\} + O(\varepsilon^2), \end{aligned} \tag{9.10}$$

where

$$\varepsilon\, \frac{\|p_i\|}{\|r_i\|} \le \varepsilon\, \frac{\|r_i\| + \delta_i \|p_{i-1}\|}{\|r_i\|} + O(\varepsilon^2) \le \varepsilon \left\{ 1 + \frac{\|r_i\|}{\|r_{i-1}\|} \frac{\|p_{i-1}\|}{\|r_{i-1}\|} \right\} + O(\varepsilon^2). \tag{9.11}$$

Recursive application of (9.11) for $\|p_{i-1}\|/\|r_{i-1}\|$, $\|p_{i-2}\|/\|r_{i-2}\|$, $\ldots$, $\|p_1\|/\|r_1\|$ with $\|p_0\|/\|r_0\| = 1$ gives

$$\varepsilon\,\frac{\|p_i\|}{\|r_i\|} \le \varepsilon\left\{1 + \frac{\|r_i\|}{\|r_{i-1}\|} + \frac{\|r_i\|}{\|r_{i-2}\|} + \ldots + \frac{\|r_i\|}{\|r_0\|}\right\} + O(\varepsilon^2). \tag{9.12}$$

The size of $\varepsilon\,\|r_i\|/\|r_k\|$, $i \ge k$ is, according to (7.20), less or equal than $\varepsilon\,\kappa(A)^{1/2} + O(\varepsilon^2)$. Consequently,

$$\varepsilon\,\frac{\|p_i\|}{\|r_i\|} \le \varepsilon\,\{1 + i\,\kappa(A)^{1/2}\} + O(\varepsilon^2). \tag{9.13}$$

Summarizing (9.8), (9.9), (9.10) and (9.13), the ratio $\varepsilon\,|M_i|/\|r_i\|^2$ is bounded as

$$\varepsilon\,\frac{|M_i|}{\|r_i\|^2} \le \varepsilon\,\kappa(A)\,O(8 + 2c + n + 3i) + O(\varepsilon^2). \tag{9.14}$$

Combining this result with (9.6) proves the following theorem.

THEOREM 9.1. *Using the previous notation, let $\varepsilon\,(n + c)\,\kappa(A) \ll 1$. Then the local orthogonality between the direction vectors and the iteratively computed residuals is in the finite precision implementation of the conjugate gradient method* (*7.6*)–(*7.8*) *and* (*7.12*) *bounded by*

$$|p_j^T r_{j+1}| \le \varepsilon\,\|r_j\|^2 \kappa(A)\,O((j+1)(8 + 2c + n + 3j)) + O(\varepsilon^2). \tag{9.15}$$

10. Final precision analysis – conclusions.

We now return to (8.6) and finalize our discussion. Using (9.1) and (9.6),

$$\begin{aligned} \|x - x_j\|_A^2 - \|x - x_{j+1}\|_A^2 \;=\; & \gamma_j \|r_j\|^2 \\ & + \varepsilon\,\gamma_j \|r_j\|^2 \left\{ \zeta_j^\gamma \frac{p_j^T A p_j}{\|r_j\|^2} + 2\sum_{i=0}^{j} \frac{M_i}{\|r_i\|^2} \right\} \\ & + 2\varepsilon\,\{(f_{j+1} + A z_j^x)^T (x_{j+1} - x_j) + r_{j+1}^T z_j^x\} \\ & + O(\varepsilon^2). \end{aligned} \tag{10.1}$$

The term

$$E_j^{(1)} \equiv \varepsilon \left\{ \zeta_j^\gamma \frac{p_j^T A p_j}{\|r_j\|^2} + 2\sum_{i=0}^{j} \frac{M_i}{\|r_i\|^2} \right\}$$

is bounded using (7.14) and (9.14),

$$|E_j^{(1)}| \le \varepsilon\,\kappa(A)\,O(n + c + 2(j+1)(8 + 2c + n + 3j)) + O(\varepsilon^2). \tag{10.2}$$

We write the remaining term on the right hand side of (10.1) proportional to ε as

$$2\varepsilon\,\{(f_{j+1} + A z_j^x)^T (x_{j+1} - x_j) + r_{j+1}^T z_j^x\} \equiv \|x - x_j\|_A\, E_j^{(2)}, \tag{10.3}$$

where

$$\begin{aligned} |E_j^{(2)}| &= 2\varepsilon \left| (f_{j+1} + A z_j^x)^T \left(\frac{x_{j+1} - x + x - x_j}{\|x - x_j\|_A} \right) + \frac{r_{j+1}^T}{\|x - x_j\|_A} z_j^x \right| \\ &\le 2\varepsilon\,\{2\,(\|f_{j+1}\|\lambda_1^{-1/2} + \|A\|^{1/2}\|z_j^x\|) + \|A\|^{1/2}\|z_j^x\|\}. \end{aligned} \tag{10.4}$$

With (7.16) and (7.9),

$$\begin{aligned}
|E_j^{(2)}| &\le 4\varepsilon\|A\|^{1/2}\kappa(A)^{1/2}(\|x\| + \max_{0\le i\le j+1}\|x_i\|)\,O(jc)\\
&\quad + 5\|A\|^{1/2}\varepsilon(3\|x_j\| + 2\|x_{j+1}\|) + O(\varepsilon^2)\\
&\le \varepsilon\|A\|^{1/2}\kappa(A)^{1/2}(\|x\| + \max_{0\le i\le j+1}\|x_i\|)\,O(4jc+25) + O(\varepsilon^2). \qquad (10.5)
\end{aligned}$$

Finally, using the fact that the monotonicity of the A-norm and the Euclidean norm of the error is preserved also in finite precision CG computations (with small additional inaccuracy, see [19], [22]), we obtain the finite precision analogy of (4.4), which is formulated as a theorem.

THEOREM 10.1. *With the notation defined above, let* $\varepsilon\,(n+c)\,\kappa(A) \ll 1$. *Then the CG approximate solutions computed in finite precision arithmetic satisfy*

$$\|x - x_j\|_A^2 - \|x - x_{j+d}\|_A^2 \;=\; \nu_{j,d} + \nu_{j,d}\,E_{j,d}^{(1)} \;+\; \|x - x_j\|_A\,E_{j,d}^{(2)} + O(\varepsilon^2), \qquad (10.6)$$

where

$$\nu_{j,d} = \sum_{i=j}^{j+d-1} \gamma_i\|r_i\|^2 \qquad (10.7)$$

and the terms due to rounding errors are bounded by

$$\begin{aligned}
&|E_{j,d}^{(1)}| \le O(d)\max_{j\le i\le j+d-1}|E_i^{(1)}|, \qquad (10.8)\\
&\qquad |E_i^{(1)}| \le \varepsilon\,\kappa(A)\,O(t^{(1)}(n)) + O(\varepsilon^2),\\
&|E_{j,d}^{(2)}| \le O(d)\max_{j\le i\le j+d-1}|E_i^{(2)}|, \qquad (10.9)\\
&\qquad |E_i^{(2)}| \le \varepsilon\,\|A\|^{1/2}\kappa(A)^{1/2}(\|x\| + \max_{0\le i\le j+1}\|x_i\|)\,O(t^{(2)}(n)) + O(\varepsilon^2).
\end{aligned}$$

$O(t^{(1)}(n))$ *and* $O(t^{(2)}(n))$ *represent terms bounded by a small degree polynomial in* n *independent of any other variables.*

Please note that the value $\nu_{j,d}$ is in Theorem 10.1 computed *exactly* using (10.7). Errors in computing $\nu_{j,d}$ *numerically* (i.e. in computing $\mathrm{fl}(\sum_{i=j}^{j+d-1}\gamma_i\|r_i\|^2)$) are negligible in comparison to $\nu_{j,d}$ multiplied by the bound for the term $|E_i^{(1)}|$ and need not be considered here. Theorem 10.1 therefore says that for the numerically computed approximate solutions

$$\|x - x_j\|_A^2 - \|x - x_{j+d}\|_A^2 = \mathrm{fl}(\nu_{j,d}) + \widetilde{\nu}_{j,d}, \qquad (10.10)$$

where the term $\widetilde{\nu}_{j,d}$ "perturbes" the ideal identity (4.4) in the finite precision case. Here $\widetilde{\nu}_{j,d}$ denotes quantity insignificantly different from $\widetilde{\nu}_{j,d}$ in (8.1). Consequently, the numerically computed value $\nu_{j,d}$ can be trusted until it reaches the level of $\widetilde{\nu}_{j,d}$. Based on the assumption $\varepsilon(n+c)\,\kappa(A) \ll 1$ and (10.8) we consider $|E_i^{(1)}| \ll 1$. Then, assuming (4.5), the numerically computed value $\nu_{j,d}$ gives a good estimate for the A-norm of the error $\|x - x_j\|_A^2$ until

$$\|x - x_j\|_A\,|E_{j,d}^{(2)}| \ll \|x - x_j\|_A^2,$$

which is equivalent to

$$\|x - x_j\|_A \gg |E_{j,d}^{(2)}|. \qquad (10.11)$$

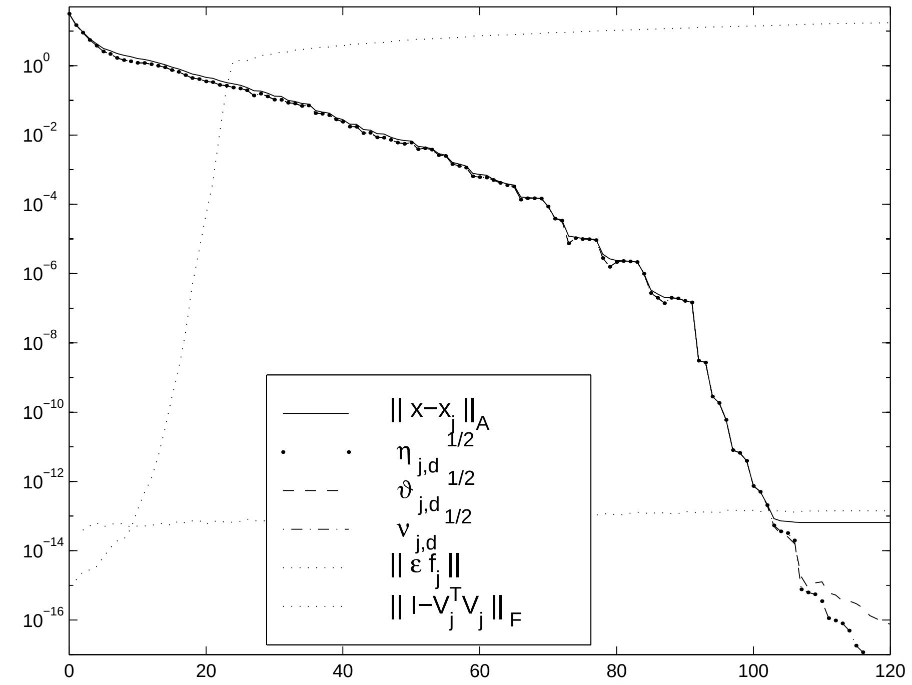


FIG. 11.1. *Error estimates* $\eta_{j,d}^{1/2}$ *(dots),* $\vartheta_{j,d}^{1/2}$ *(dashed-line) and* $\nu_{j,d}^{1/2}$ *(dash-doted line). They essentially coincide until* $\|x - x_j\|_A$ *(solid line) reaches its ultimate attainable accuracy. The loss of orthogonality is plotted by the dotted line. We used* $d = 4$*.*

The value $E_{j,d}^{(2)}$ represents various terms. Its upper bound is, apart from $\kappa(A)^{1/2}$, which comes into play as an effect of the worst-case rounding error analysis, linearly dependent on an upper bound for $\|x - x_0\|_A$. The value of $E_{j,d}^{(2)}$ is (as similar terms or constants in any other rounding error analysis) not important. What is important is the following possible interpretation of (10.11): until $\|x - x_j\|_A$ reaches a level close to $\varepsilon\|x - x_0\|_A$, the computed estimate $\nu_{j,d}$ must work.

11. Numerical Experiments. We present illustrative experimental results for the system $Ax = b$ described in Section 5. We set $d = 4$.

Fig. 11.1 demonstrates, that the estimates $\eta_{j,d}^{1/2}$ (computed by the algorithm CGQL [16], dotted line), $\vartheta_{j,d}^{1/2}$ (dashed line) and $\nu_{j,d}^{1/2}$ (dash-dotted line) give in the presence of rounding errors similar results; all the lines essentially coincide until $\|x - x_j\|_A$ (solid line) reaches its ultimate attainable accuracy level. Loss of orthogonality, measured by $\|I - V_j^T V_j\|_F$, is plotted by the strictly increasing dotted line. We see that the orthogonality of the computed Lanczos basis is completely lost at $j \sim 22$. The term $\|\varepsilon f_j\|$ measuring the difference between the directly and iteratively computed residuals (horizontal dotted line) remains close to machine precision $\varepsilon \sim 10^{-16}$ throughout the whole computation.

Fig. 11.2 shows, in addition to the loss of orthogonality (dotted line) and the Euclidean norm of the error $\|x - x_j\|$, the bound for the last one derived in the following way from (1.7). Using the identity

$$\|x - x_j\|^2 = \sum_{i=j}^{j+d-1} \frac{\|p_i\|^2}{(p_i, Ap_i)} \left(\|x - x_i\|_A^2 + \|x - x_{i+1}\|_A^2\right) + \|x - x_{j+d}\|^2, \tag{11.1}$$

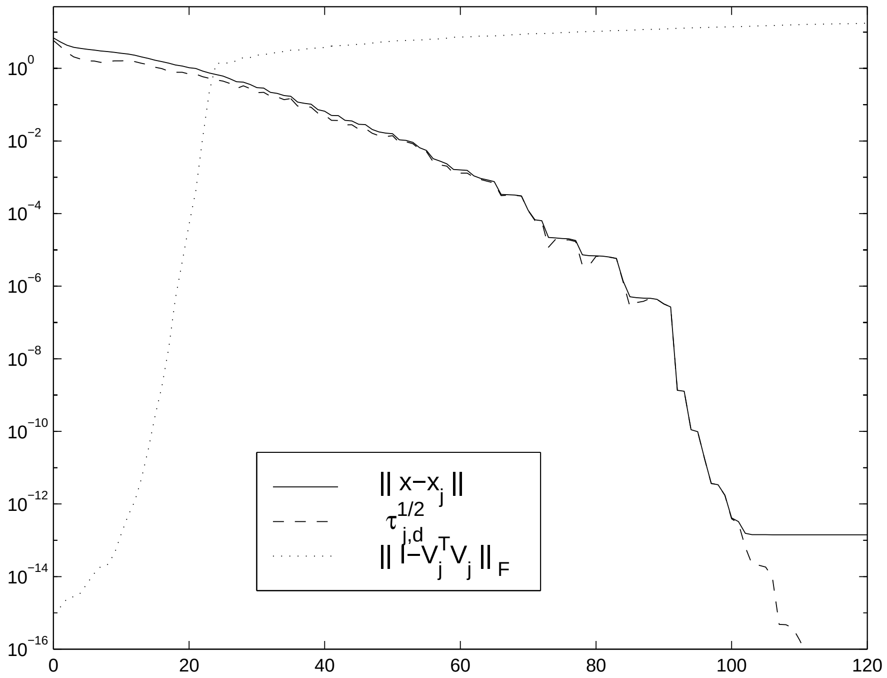


FIG. 11.2. *Lower bound* $\tau_{j,d}^{1/2}$ (*dashed line*) *for the Euclidean norm of the error* (*solid line*). *The bound* $\tau_{j,d}^{1/2}$ (*with* $d = 4$) *gives, despite the loss of orthogonality* (*dotted line*), *very good approximation to* $\|x - x_j\|$.

and replacing the unknown squares of the A-norms of the errors

$$\|x - x_j\|_A^2,\ \|x - x_{j+1}\|_A^2,\ \ldots,\ \|x - x_{j+d}\|_A^2$$

by their estimates

$$\sum_{i=j}^{j+2d-1} \gamma_i \|r_i\|^2,\ \sum_{i=j+1}^{j+2d-1} \gamma_i \|r_i\|^2,\ \ldots,\ \sum_{i=j+d}^{j+2d-1} \gamma_i \|r_i\|^2$$

gives ideally

$$\|x - x_j\|^2 \geq \sum_{i=j}^{j+d-1} \frac{\|p_i\|^2}{(p_i, Ap_i)} \left(\gamma_i \|r_i\|^2 + 2 \sum_{k=i+1}^{j+2d-1} \gamma_k \|r_k\|^2 \right) + \|x - x_{j+d}\|^2. \tag{11.2}$$

Similarly as above, if d is chosen such that

$$\|x - x_j\|^2 \gg \|x - x_{j+d}\|^2 \quad \text{and} \quad \|x - x_{j+d}\|_A^2 \gg \|x - x_{j+2d}\|_A^2,$$

then

$$\tau_{j,d} \equiv \sum_{i=j}^{j+d-1} \frac{\|p_i\|^2}{(p_i, Ap_i)} \left(\gamma_i \|r_i\|^2 + 2 \sum_{k=i+1}^{j+2d-1} \gamma_k \|r_k\|^2 \right) \tag{11.3}$$

represents ideally a tight lower bound for the squared Euclidean norm of the CG error $\|x - x_j\|^2$. Please note that evaluating (11.3) requires $2d$ extra steps.

In experiments shown in Fig. 11.1 and Fig. 6.1 we used a fixed value $d = 4$. It would be interesting to design an adaptive error estimator, which would use some heuristics for adjusting d according to the desired accuracy of the estimate and the convergence behaviour. A similar approach can be used for eliminating the disadvantage of $2d$ extra steps related to (11.3). We hope to report results of our work on that subject elsewhere.

12. Conclusions. Based on the results presented above we believe that the estimate for the A-norm of the error $\nu_{j,d}^{1/2}$ should be incorporated into any software realization of the CG method. It is simple and numerically stable. It is worth to consider the estimate $\tau_{j,d}^{1/2}$ for the Euclidean norm of the error, and compare it (including complexity and numerical stability) with other existing approaches not discussed here (e.g. [6], [31]). The choice of d remains a subject of further work.

By this paper we wish to pay a tribute to the truly seminal paper of Hestenes and Stiefel [24] and to the work of Golub who shaped the whole field.

Acknowledgments. Many people have contributed to the presentation of this paper by their advice, helpful objections, and remarks. We wish to especially thank Martin Gutknecht, Gerard Meurant, Chris Paige, Beresford Parlett, Lothar Reichel, Miroslav Rozložník, and the anonymous referee for their help in revising the text.

REFERENCES

[1] M. Arioli, *Stopping criterion for the Conjugate Gradient algorithm in a Finite Element method framework*, submitted to Numer. Math., (2001).

[2] M. Arioli and L. Baldini, *Backward error analysis of a null space algorithm in sparse quadratic programming*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 425–442.

[3] O. Axelsson and I. Kaporin, *Error norm estimation and stopping criteria in preconditioned Conjugate Gradient iterations*, Numer. Linear Algebra Appl., 8 (2001), pp. 265–286.

[4] D. Calvetti, G. H. Golub, and L. Reichel, *Estimation of the L-curve via Lanczos bidiagonalization*, BIT, 39 (1999), pp. 603–609.

[5] D. Calvetti, S. Morigi, L. Reichel, and F. Sgallari, *Computable error bounds and estimates for the Conjugate Gradient Method*, Numer. Algorithms, 25 (2000), pp. 79–88.

[6] ———, *An iterative method with error estimators*, J. Comput. Appl. Math., 127 (2001), pp. 93–119.

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

[8] G. Dahlquist, G. H. Golub, and S. G. Nash, *Bounds for the error in linear systems*, in Proc. Workshop on Semi-Infinite Programming, R. Hettich, ed., Springer, Berlin, 1978, pp. 154–172.

[9] P. Deuflhard, *Cascadic conjugate gradient methods for elliptic partial differential equations I: Algorithm and numerical results*, preprint SC 93-23, Konrad-Zuse-Zentrum für Informationstechnik Berlin, Heilbronnen Str., D-10711 Derlin, October 1993.

[10] ———, *Cascadic conjugate gradient methods for elliptic partial differential equations: Algorithm and numerical results*, Contemp. Math., 180 (1994), pp. 29–42.

[11] B. Fischer, *Polynomial Based Iteration Methods for Symmetric Linear Systems*, Wiley Teubner Advances in Numerical Mathematics, Wiley Teubner, 1996.

[12] B. Fischer and G. H. Golub, *On the error computation for polynomial based iteration methods*, in Recent Advances in Iterative Methods, G. H. Golub, A. Greenbaum, and M. Luskin, eds., Springer, N.Y., 1994, pp. 59–67.

[13] A. Frommer and A. Weinberg, *Verified error bounds for linear systems through the Lanczos process*, Reliable Computing, 5 (1999), pp. 255–267.

[14] W. Gautschi, *A survey of Gauss-Christoffel quadrature formulae*, in E.B. Christoffel. The Influence of His Work on Mathematics and the Physical Sciences, P. Bultzer and F. Fehér, eds., Birkhauser, Boston, 1981, pp. 73–157.

[15] G. H. Golub and G. Meurant, *Matrices, moments and quadrature*, in Numerical Analysis 1993, vol 303, Pitman research notes in mathematics series, D. Griffiths and G. Watson, eds., Longman Sci. Tech. Publ., 1994, pp. 105–156.

[16] ———, *Matrices, moments and quadrature II: How to compute the norm of the error in iterative methods*, BIT, 37 (1997), pp. 687–705.

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

[18] G. H. Golub and C. van Loan, *Matrix Computation*, The Johns Hopkins University Press, Baltimore MD, third ed., 1996.

[19] A. Greenbaum, *Behavior of slightly perturbed Lanczos and Conjugate Gradient recurrences*, Linear Algebra Appl., 113 (1989), pp. 7–63.

[20] ———, *Estimating the attainable accuracy of recursively computed Residual methods*, SIAM J. Matrix Anal. Appl., 18 (3) (1997), pp. 535–551.

[21] ———, *Iterative Methods for Solving Linear Systems*, SIAM, Philadelphia, 1997.

[22] A. Greenbaum and Z. Strakoš, *Predicting the behavior of finite precision Lanczos and Conjugate Gradient computations*, SIAM J. Matrix Anal. Appl., 18 (1992), pp. 121–137.

[23] M. H. Gutknecht and Z. Strakoš, *Accuracy of two three-term and three two-term recurrences for Krylov space solvers*, SIAM J. Matrix Anal. Appl., 22 (2001), pp. 213–229.

[24] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), pp. 409–435.

[25] N. J. Higham, *Accuracy and stability of numerical algorithms*, SIAM, Philadelphia, PA, 1996.

[26] S. Karlin and L. S. Shapley, *Geometry of moment spaces*, Memoirs of the Americam Mathematical Society 12, Providence, (1953).

[27] C. Lanczos, *Solution of systems of linear equations by minimized iterations*, J. Research Nat. Bur. Standards, 49 (1952), pp. 33–53.

[28] G. Meurant, *The computation of bounds for the norm of the error in the Conjugate Gradient algorithm*, Numer. Algorithms, 16 (1997), pp. 77–87.

[29] ———, *Computer Solution of Large Linear Systems*, vol. 28 of Studies in Mathematics and Its Applications, Elsevier, 1999.

[30] ———, *Numerical experiments in computing bounds for the norm of the error in the preconditioned Conjugate Gradient algorithm*, Numer. Algorithms 22, 3-4 (1999), pp. 353–365.

[31] ———, *Towards a reliable implementation of the conjugate gradient method*. Invited plenary lecture at the Latsis Symposium: Iterative Solvers for Large Linear Systems, Zurich, February 2002.

[32] Y. Notay, *On the convergence rate of the Conjugate Gradients in the presence of rounding errors*, Numer. Math., 65 (1993), pp. 301–317.

[33] C. C. Paige, *The Computation of Eigenvalues and Eigenvectors of Very Large Sparse Matrices*, PhD thesis, Intitute of Computer Science, University of London, London, U.K., 1971.

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

[35] ———, *Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix*, J. Inst. Math. Appl., 18 (1976), pp. 341–349.

[36] ———, *Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem*, Linear Algebra Appl., 34 (1980), pp. 235–258.

[37] C. C. Paige and Z. Strakoš, *Correspondence between exact arithmetic and finite precision behaviour of Krylov space methods*, in XIV. Householder Symposium, J. Varah, ed., University of British Columbia, 1999, pp. 250–253.

[38] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice Hall, Englewood Cliffs, 1980.

[39] ———, *Do we fully understand the symmetric Lanczos algorithm yet?*, in Proceedins of the Lanczos Centenary Conference, Philadelphia, 1994, SIAM, pp. 93–107.

[40] P. E. Saylor and D. C. Smolarski, *Addendum to: Why Gaussian quadrature in the complex plane?*, Numer. Algorithms, 27 (2001), pp. 215–217.

[41] ———, *Why Gaussian quadrature in the complex plane?*, Numer. Algorithms, 26 (2001), pp. 251–280.

[42] G. L. Sleijpen, H. A. van der Vorst, and D. R. Fokkema, *BiCGstab(l) and other hybrid BiCG methods*, Numer. Algorithms, 7 (1994), pp. 75–109.

[43] Z. Strakoš, *On the real convergence rate of the Conjugate Gradient method*, Linear Algebra Appl., 154-156 (1991), pp. 535–549.

[44] ———, *Convergence and numerical behaviour of the Krylov space methods*, in Algorithms for Large Sparse Linear Algebraic Systems: The State of the Art and Applications in Science and Engineering, G. W. Althaus and E. Spedicato, eds., NATO ASI Institute, Kluwer Academic, 1998, pp. 175–197.

[45] Z. Strakoš and A. Greenbaum, *Open questions in the convergence analysis of the Lanczos process for the real symmetric eigenvalue problem*, IMA preprint series 934, University of Minnesota, 1992.

[46] G. Szegö, *Orthogonal Polynomials*, AMS Colloq. Publ. 23, AMS, Providence, 1939.

[47] K. F. Warnick, *Nonincreasing error bound for the Biconjugate Gradient method*, report, University of Illinois, 2000.

BIT Numerical Mathematics (2005) 45: 789–817
DOI: 10.1007/s10543-005-0032-1

ERROR ESTIMATION IN PRECONDITIONED CONJUGATE GRADIENTS*

ZDENĚK STRAKOŠ[1,**] and PETR TICHÝ[1,***]

[1] *Institute of Computer Science, Academy of Sciences of the Czech Republic,*
Pod Vodárenskou věží 2, 182 07 Praha 8, Czech Republic.
email: {strakos,tichy}@cs.cas.cz

Abstract.

In practical problems, iterative methods can hardly be used without some acceleration of convergence, commonly called preconditioning, which is typically achieved by incorporation of some (incomplete or modified) direct algorithm as a part of the iteration. Effectiveness of preconditioned iterative methods increases with possibility of stopping the iteration when the desired accuracy is reached. This requires, however, incorporating a proper measure of achieved accuracy as a part of computation.

The goal of this paper is to describe a simple and numerically reliable estimation of the size of the error in the preconditioned conjugate gradient method. In this way this paper extends results from [Z. Strakoš and P. Tichý, ETNA, 13 (2002), pp. 56–80] and communicates them to practical users of the preconditioned conjugate gradient method.

AMS subject classification (2000): 15A06, 65F10, 65F25, 65G50.

Key words: preconditioned conjugate gradient method, error bounds, stopping criteria, evaluation of convergence, numerical stability, finite precision arithmetic, rounding errors.

1 Introduction.

Discretization of mathematical models of real-world problems often leads to large and sparse (possibly structured) systems of linear algebraic equations. All steps of mathematical modeling (mathematical description of reality in the form of a mathematical model, its discretization and numerical solution of the discretized problem) are subject to errors (errors of the model, discretization errors

* Received September 2004. Accepted in final form August 2005. Communicated by Iain Duff.

** This work was supported by the Program Information Society, project 1ET400300415, and by the Grant Agency of Academy of Sciences of the Czech Republic under grant No. KJB1030306.

*** This work was performed during the academic year 2003/2004 while this coauthor was on leave at the Institute of Mathematics, TU Berlin, Germany, sponsored by the Emmy Noether Program of the Deutsche Forschungsgemeinschaft.

and computational errors, the last being often composed of two parts – truncation errors and errors due to roundoff). An output of the solution process must therefore be confronted with its possible errors through *verification and validation.* While verification addresses the question – whether and how accurately the obtained (approximate) solution conforms to the mathematical model, validation deals with the more general question – to which extent the whole modeling process represents the modeled reality (for a recent discussion of these fundamental topics we refer to [7]). It is desirable that the errors of the model, discretization errors and computational errors are in some balance. They do not need to be of the same order; the discretization and computational errors should not significantly contribute to the total error and affect negatively the validation process [7].

When the linear algebraic systems arising from mathematical modeling are very large (of orders of hundreds of thousands or millions of unknowns), preconditioned iterative methods are taking ground over the purely direct methods. Iterative methods can in very large scale computations exploit a fundamental advantage – they can increase effectiveness of the whole solution process by stopping the iteration when the desired accuracy (as compared to the discretization error) is reached (cf. [1, 4]). This requires, however, a cheap and reliable evaluation of convergence, which is the essential ingredience for choosing proper stopping criteria.

In this paper we consider a system of linear algebraic equations

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

where A is a symmetric positive definite n by n matrix and b is n-dimensional vector (for simplicity of notation we consider A, b real; all results presented here can trivially be extended to the complex case). For such systems the preconditioned conjugate gradient method [22, 26, 34, 40] represents in most large scale cases a good choice. A goal of this paper is to summarize and discuss evaluation of convergence in the preconditioned conjugate gradient method. In particular, we will focus on estimating the A-norm of the error.

Estimating the A-norm of the error in the conjugate gradient method was subject of many papers, reports and subsections in the books. History and various aspects of estimating the A-norm of the error in the unpreconditioned conjugate gradient method were thoroughly described in [38]. The formulas presented in [38] were published (in some form) previously, e.g. in [22, 12] and [6]. The original contribution of [38] consists, to our opinion, in providing *theoretical justification for practical use of the error estimates* and in putting different estimates in the proper context. Our present paper extends the results from [38] to the preconditioned conjugate gradient method. A need for such paper can be seen from [6, Section 6] or [1, Section 3], which thoroughly and extensively examine estimating error norms in the preconditioned conjugate gradients. Both papers [6, 1] present interesting original results and offer new insight into the error estimation in the preconditioned conjugate gradients. They do not consider, however, an influence of rounding errors. All derivations in [6, Section 6] or [1, Section 3] assume exact arithmetic. Consequently, they unrealistically assume preserving orthogonality, and the results are based on exploiting the finite termination property, i.e., on getting the *exact solution* in a finite number of steps (which does not exceed the dimension of the problem). These assumptions are clearly drastically violated in most practical computations. In order to be widely used, practical error estimators need a proper justification including a thorough analysis of rounding error effects (for a related discussion, see [38] and also [16]).

Section 2 summarizes fundamentals of the conjugate gradient method and briefly recalls several possible ways of convergence evaluation. Section 3 presents a simple estimate for the A-norm of the error in the preconditioned conjugate gradient method. Section 4 deals with numerical stability of the proposed estimate and Section 5 contains numerical experiments which demonstrate its effectivity and possible drawbacks. The paper ends with concluding remarks.

2 Fundamentals of convergence evaluation.

The conjugate gradient method (CG) [22] belongs to the class of the so-called Krylov subspace methods. Starting with an initial approximation x_0, it constructs the subsequent approximations x_j, $j = 1, 2, \dots$ to the solution x on the linear manifolds

$$x_j \in x_0 + \mathcal{K}_j(A, r_0) \tag{2.1}$$

where

$$\mathcal{K}_j(A, r_0) = \operatorname{span}\left\{r_0, Ar_0, \dots, A^{j-1}r_0\right\}$$

represents the jth Krylov subspace, $r_0 = b - Ax_0$. CG determines its approximations by orthogonal projections, i.e., the residual $r_j = b - Ax_j$ of the jth approximate solution is orthogonal to the jth Krylov subspace $\mathcal{K}_j(A, r_0)$. This means that $x_j = x_0 + y_j$ can be obtained from the solution y_j of the j-dimensional problem

$$P_j\{r_0 - Ay\} = 0\,, \tag{2.2}$$

where P_j stands for the orthogonal projection onto $\mathcal{K}_j(A, r_0)$, and $y \in \mathcal{K}_j(A, r_0)$ (the operator A is in (2.2) restricted to $\mathcal{K}_j(A, r_0)$). It is well known [22] that, until x_j converges to the exact solution x (which must in the absence of roundoff happen in at most n steps), x_j is uniquely determined by (2.2).

In practical problems we hope that the acceptable approximate solution is attained for j much smaller than the dimension of the problem n. Thus, CG represents a typical model-reduction approach, in which the original problem (represented by the large discretized model) is reduced (here by restriction and orthogonal projection onto the Krylov subspace) to the problem of much smaller dimension. The resulting reduced problem determines the approximate solution. Quality of the approximate solution depends on the amount of significant information about the original problem passed to the reduced problem.

The condition (2.2) is equivalent to the minimization of the A-norm of the error over the manifold (2.1). The jth CG approximation is therefore uniquely determined by the minimizing condition

$$\|x - x_j\|_A \ = \min_{u \in x_0 + \mathcal{K}_j(A, r_0)} \|x - u\|_A \,, \tag{2.3}$$

where

$$\|x - u\|_A \ = \ (x - u, A(x - u))^{\frac{1}{2}} \ . \tag{2.4}$$

The A-norm of the error on the algebraic level (2.4) typically has a counterpart in the original real-world problem. In some applications it can be interpreted as the discretized measure of energy which is to be minimized see, e.g. [1, 4]. Then CG with stopping criterion based on the A-norm of the error consistently reduces large discretized models to small ones. In other applications (such as in image processing) the Euclidean norm of the error $\|x - x_j\|$ plays an important role. In this paper we focus in particular on estimating the A-norm of the error.

Hestenes and Stiefel [22] considered the A-norm of the error a possible candidate for measuring the "goodness" of x_j as an estimate of x. They showed that though it was impossible to compute the A-norm of the jth error without knowing the solution x, it was possible to estimate it. Later, and independently of [22], the idea of estimating errors in CG was promoted by Golub in relation to the problem of moments, Gauss quadrature and its modifications [10, 11]. A comprehensive summary of this approach was given in the papers coauthored with Meurant [14, 15].

In [38] it was shown that the lower bound for the A-norm of the error based on the Gauss quadrature is mathematically equivalent to the lower bound derived from the identity given by Hestenes and Stiefel in [22]. The estimate by Hestenes and Stiefel can be computed at a negligible cost of several floating point operations per iteration. Until the A-norm of the error reaches its ultimate level of accuracy, this estimate is numerically stable.

In [32, 3], backward error perturbation theory (see e.g. [30, 35, 2]) was used to derive a family of stopping criteria for iterative methods. In particular, given x_j, the relative norms $\|\Delta A\|/\|A\| = \|\Delta b\|/\|b\|$ of the smallest perturbations ΔA and Δb such that the *approximate solution* x_j represents the *exact solution* of the perturbed system

$$(A + \Delta A)\, x_j = b + \Delta b$$

can be computed by the normwise backward error

$$\frac{\|r_j\|}{\|A\| \|x_j\| + \|b\|} \,. \tag{2.5}$$

This approach can be generalized in order to quantify levels of confidence in A and b, see [32, 3]. Normwise backward error is, as a base for stopping criteria, frequently recommended in the numerical analysis literature, see, e.g. [8, 23], and it is used and popularized by numerical analysts [29, 13]. Despite this effort, evaluating convergence is in most of scientific computations still based on the relative residual norm

$$\frac{\|r_j\|}{\|r_0\|}\,. \tag{2.6}$$

With $x_0 = 0$, it measures the relative norm $\|\Delta b\|/\|b\|$ of the smallest perturbation Δb in the right-hand side b only (A is considered unperturbed) such that x_j is the exact solution of the perturbed system $A\,x_j = b + \Delta b$. For $x_0 \neq 0$ (2.6) strongly depends on the initial approximation x_0 and can give a misleading information about convergence, see, e.g. [33]. For some additional information see also [5, 20].

We do not argue that the relative residual norm can not be useful. In some cases it is a proper quantity to be checked. Sometimes it is a part of more complex convergence considerations, e.g. in solving nonlinear systems or in numerical optimization. We do argue, however, that in many other cases, and in particular in numerical solving of partial differential equations, the relative residual norm is often uncritically used as the only measure of convergence.

Mathematically (ignoring effects of rounding errors), extension of the approaches mentioned above to preconditioned methods does not represent a problem, see, e.g., [29, 13]. Extension of the Gauss quadrature-based formulas for estimating the A-norm of the error in CG (algorithm CGQL [15]) to the preconditioned conjugate gradient method (PCG) was published in [27, 28] (algorithm PCGQL). In the following section we deal with the extension of error estimates based on the Hestenes and Stiefel formula [22, 38].

3 PCG error estimates.

In the standard view of preconditioning, the CG method is thought of as being applied to a "preconditioned" system

$$\hat{A}\hat{x} = \hat{b}, \tag{3.1}$$

$$\hat{A} = L^{-1}AL^{-T}, \quad \hat{b} = L^{-1}b, \tag{3.2}$$

where L represents a proper nonsingular (lower triangular) matrix, giving

Algorithm 1. *CG for* $\hat{A}\hat{x} = \hat{b}$

```
given $\hat{x}_0$, $\hat{r}_0 = \hat{b} - \hat{A}\hat{x}_0$,
for $j = 0, 1, \ldots$
    $\gamma_j = \frac{(\hat{r}_j, \hat{r}_j)}{(\hat{p}_j, \hat{A}\hat{p}_j)}$
    $\hat{x}_{j+1} = \hat{x}_j + \gamma_j\,\hat{p}_j$
    $\hat{r}_{j+1} = \hat{r}_j - \gamma_j\,\hat{A}\hat{p}_j$
    $\hat{\delta}_{j+1} = \frac{(\hat{r}_{j+1}, \hat{r}_{j+1})}{(\hat{r}_j, \hat{r}_j)}$
    $\hat{p}_{j+1} = \hat{r}_{j+1} + \hat{\delta}_{j+1}\,\hat{p}_j$
end for.
```

Defining

$$\begin{aligned}(3.3)\quad &\gamma_j \equiv \hat{\gamma}_j, \quad \delta_j \equiv \hat{\delta}_j,\\ &x_j \equiv L^{-T}\hat{x}_j, \quad r_j \equiv L\,\hat{r}_j, \quad p_j \equiv L^{-T}\hat{p}_j, \quad s_j \equiv L^{-T}L^{-1}r_j \equiv M^{-1}r_j,\end{aligned}$$

(here x_j and r_j represent the approximate solution and residual for the original problem $Ax = b$), we obtain the standard version of the PCG method

Algorithm 2. *PCG for $Ax = b$*

```
given $x_0$, $r_0 = b - Ax_0$, $s_0 = M^{-1}r_0$, $p_0 = s_0$,
for $j = 0, 1, \ldots$
        $\gamma_j = \frac{(r_j, s_j)}{(p_j, Ap_j)}$
        $x_{j+1} = x_j + \gamma_j\, p_j$
        $r_{j+1} = r_j - \gamma_j\, Ap_j$
        $s_{j+1} = M^{-1}r_{j+1}$
        $\delta_{j+1} = \frac{(r_{j+1}, s_{j+1})}{(r_j, s_j)}$
        $p_{j+1} = s_{j+1} + \delta_{j+1}\, p_j$
end for.
```

The preconditioner

$$M = LL^T \tag{3.4}$$

is chosen so that the linear system with the matrix M is easy to solve, while the matrix $L^{-1}AL^{-T}$ should ensure fast convergence of CG. The last goal is fulfilled, e.g., when $L^{-1}AL^{-T}$ is well conditioned (approximates the identity matrix) or has properly clustered eigenvalues. Here we emphasize that *location* as well as *diameter* of the clusters are important; improperly located clusters of very small diameter do not necessarily ensure fast convergence, see [21, 37]. Location of the clusters is sometimes omitted from consideration, and this leads to inaccurate or even false statements, which can be found in widespread literature.

3.1 Estimating the A-norm of the error.

In PCG, the A-norm of the error can be estimated similarly as in ordinary CG. For a given d, the approximate solutions $\hat{x}_j$ of the system (3.1) satisfy

$$\|\hat{x} - \hat{x}_j\|_{\hat{A}}^2 = \sum_{i=j}^{j+d-1} \hat{\gamma}_i \|\hat{r}_i\|^2 + \|\hat{x} - \hat{x}_{j+d}\|_{\hat{A}}^2, \tag{3.5}$$

see [38, (4.4)]. Using (3.3),

$$\|\hat{r}_j\|^2 = r_j^T L^{-T} L^{-1} r_j = r_j^T M^{-1} r_j = (r_j, s_j)\,,$$

and

$$\|\hat{x} - \hat{x}_j\|_{\hat{A}}^2 = (L^T x - L^T x_j)^T L^{-1} A L^{-T} (L^T x - L^T x_j) = \|x - x_j\|_A^2.$$

The identity (3.5) can therefore be written in the form

$$\|x - x_j\|_A^2 = \sum_{i=j}^{j+d-1} \gamma_i \, (r_i, s_i) + \|x - x_{j+d}\|_A^2. \tag{3.6}$$

Assuming a reasonable decrease of the A-norm of the error in the steps $j+1$ through $j+d$, the square root of the quantity

$$\nu_{j,d} \equiv \sum_{i=j}^{j+d-1} \gamma_i \, (r_i, s_i) \tag{3.7}$$

gives a tight lower bound for the A-norm of the jth error of PCG applied to the system $Ax = b$. Please notice that (similarly as in the ordinary CG) the quantities γ_i and (r_i, s_i) are at our disposal during the PCG iterations. For earlier publications of these identities please see [39, 1].

3.2 Estimating the relative A-norm of the error.

Consider PCG applied to linear algebraic systems arising from a finite element discretization of self-adjoint elliptic partial differential equations. Then it is natural to use the stopping criterion that compares the relative A-norm of the error

$$\frac{\|x - x_j\|_A}{\|x\|_A} \tag{3.8}$$

with the discretization error, see [1].

In [1], however, the A-norm of the jth error is estimated using (3.7), while the estimate of the A-norm of the solution $\|x\|_A$ is based on the formula

$$\|x\|_A^2 = r_0^T x_j + b^T x_0 + \|x - x_j\|_A^2 \tag{3.9}$$

which gives the lower bound

$$\|x\|_A^2 \geq \tilde{\xi}_j \equiv r_0^T x_j + b^T x_0 \, . \tag{3.10}$$

Estimating the A-norm of the solution using the value $\tilde{\xi}_j^{1/2}$ has, besides computing an unnecessary scalar product $r_0^T x_j$, a possible disadvantage. Derivation of the identity (3.9) assumes preserving of global orthogonality during the PCG computations, cf. [1, p. 9]. In particular, it can be shown that in finite precision arithmetic it holds (up to some small inaccuracy)

$$\|x\|_A^2 \approx r_0^T x_j + b^T x_0 + r_j^T (x_j - x_0) + \|x - x_j\|_A^2. \tag{3.11}$$

In exact arithmetic, the term $r_j^T(x_j - x_0)$ is equal to zero. In finite precision arithmetic, however, its size can be close to $\|r_j\| \, \|x_j - x_0\|$. Consequently, the estimate $\tilde{\xi}_j^{1/2}$ can for large $r_j^T(x_j - x_0) + \|x - x_j\|_A^2$ (in comparison to $\|x\|_A^2$) provide misleading information about the size of $\|x\|_A$.

A mathematically equivalent identity to (3.9) that overcomes previous difficulties can be obtained in the following way. Subtracting

$$\|x - x_0\|_A^2 = \nu_{0,j+d} + \|x - x_{j+d}\|_A^2,$$

$$\|x - x_0\|_A^2 = \|x\|_A^2 - 2b^T x_0 + \|x_0\|_A^2 = \|x\|_A^2 - b^T x_0 - r_0^T x_0, \tag{3.12}$$

the identity

$$\|x\|_A^2 = \nu_{0,j+d} + b^T x_0 + r_0^T x_0 + \|x - x_{j+d}\|_A^2 \tag{3.13}$$

gives the corresponding lower bound

$$\|x\|_A^2 \;\geq\; \xi_{j+d} \equiv \nu_{0,j+d} + b^T x_0 + r_0^T x_0\,. \tag{3.14}$$

With $d = 0$, the identities (3.13) and (3.9), as well as the estimates ξ_j and $\tilde{\xi}_j$ are *mathematically equivalent*. However, the evaluation of ξ_j is cheaper than the evaluation of $\tilde{\xi}_j$ and, more substantially, (3.13) holds with a small inaccuracy also in finite precision PCG computations independently on the loss of global orthogonality, cf. Section 4.

Replacing the squared A-norm of the solution $\|x\|_A^2$ by the lower bound ξ_{j+d} and the squared jth A-norm of the error $\|x - x_j\|_A^2$ by the lower bound $\nu_{j,d}$, we obtain the estimate $\varrho_{j,d}$ for the squared relative A-norm of the error

$$\varrho_{j,d} \equiv \frac{\nu_{j,d}}{\xi_{j+d}}. \tag{3.15}$$

It should be noted that an improper choice of x_0 can give $\xi_{j+d} \leq 0$ which makes the estimate $\varrho_{j,d}$ in such case useless. We will, however, explain that $\xi_{j+d} \leq 0$ means a meaningless choice of x_0. First, a nonzero x_0 should not be used in an application of the CG method (and of any other Krylov subspace method) unless there is a good reason for using it. In CG, the very natural condition

$$\|x - x_0\|_A^2 \;\leq\; \|x\|_A^2 \tag{3.16}$$

should always be imposed. Though we can not compute the individual values $\|x\|_A^2$, $\|x - x_0\|_A^2$, its difference can easily be checked using (3.12). Second, if $\xi_{j+d} \leq 0$, then from (3.14)

$$b^T x_0 + r_0^T x_0 = \|x\|_A^2 - \|x - x_0\|_A^2 < 0 \tag{3.17}$$

and x_0 violates the condition (3.16). In such case, x_0 should be discarded or properly scaled in order to satisfy (3.16). In particular, x_0 can be scaled such that $\|x - \alpha x_0\|_A^2$ is minimal using

$$\alpha = \frac{b^T x_0}{x_0^T A x_0},$$

(for another application of the same little trick see [33, p. 1903]). With (3.16) $\xi_{j+d} > 0$ and, using (3.13),

$$0 < \varrho_{j,d} \;=\; \frac{\|x - x_j\|_A^2 - \|x - x_{j+d}\|_A^2}{\|x\|_A^2 - \|x - x_{j+d}\|_A^2} \;\leq\; \frac{\|x - x_j\|_A^2}{\|x\|_A^2},$$

i.e. $\varrho_{j,d}^{1/2}$ is a lower bound on the jth relative A-norm of the error. Please note that $\varrho_{j,d}^{1/2}$ can be close to the relative A-norm of the error even when $\nu_{j,d}^{1/2}$ is far from $\|x - x_j\|_A$.

3.3 Estimating the M-norm of the error.

In our paper [38] we described an estimate of the Euclidean norm of the error in CG. For CG applied to $\hat{A}\hat{x} = \hat{b}$, Algorithm 1, the estimate is based on the identity

$$\begin{aligned}(3.18)\qquad \|\hat{x}-\hat{x}_j\|^2 = \sum_{i=j}^{j+d-1} \frac{\|\hat{p}_i\|^2}{(\hat{p}_i, \hat{A}\hat{p}_i)} \left(\|\hat{x}-\hat{x}_i\|^2_{\hat{A}} + \|\hat{x}-\hat{x}_{i+1}\|^2_{\hat{A}}\right) \\ + \|\hat{x}-\hat{x}_{j+d}\|^2.\end{aligned}$$

Using (3.3), (3.18) can be rewritten as

$$\begin{aligned}(3.19)\qquad \|x-x_j\|^2_M = \sum_{i=j}^{j+d-1} \frac{\|p_i\|^2_M}{(p_i, Ap_i)} \left(\|x-x_i\|^2_A + \|x-x_{i+1}\|^2_A\right) \\ + \|x-x_{j+d}\|^2_M\end{aligned}$$

where x_j represents the PCG approximate solution for the original problem $Ax = b$. Replacing the unknown $\|x-x_i\|^2_A$ for $i = j, \ldots, j+d$ by the estimates $\nu_{i,2d-i+j}$ (see [38]) we obtain

$$(3.20)\qquad \|x-x_j\|^2_M \;\geq\; \tau_{j,d} + \|x-x_{j+d}\|^2_M$$

where the square root of the quantity

$$(3.21)\qquad \tau_{j,d} \;\equiv\; \sum_{i=j}^{j+d-1} \frac{\|p_i\|^2_M}{(p_i, Ap_i)} \left(\gamma_i\,(r_i, s_i) + 2 \sum_{k=i+1}^{j+2d-1} \gamma_k\,(r_k, s_k)\right)$$

represents a lower bound for the M-norm of the error.

4 Numerical stability analysis.

In [38] we showed that the Hestenes and Stiefel estimate is numerically stable (i.e. it is in finite precision CG computations not substantially affected by rounding errors) until the A-norm of the error approaches its ultimate level of accuracy. A similar result can be shown for the estimate (3.7) of the A-norm of the error in PCG.

PCG computes at each step an additional vector s_{j+1} as a solution of the linear system

$$(4.1)\qquad Ms_{j+1} = r_{j+1}\,,$$

and uses

$$(4.2)\qquad (r_{j+1}, s_{j+1})$$

for computation of the coefficients γ_{j+1} and δ_{j+1} needed for determining of the new direction vector p_{j+1}. This is the difference which must be addressed in extension of the results from CG [38] to PCG.

From now on x_{j+1}, x_j, γ_j, p_j, r_{j+1}, r_j, s_{j+1}, δ_{j+1} and p_{j+1} will represent numerically computed quantities. Numerical stability analysis of the estimate (3.7) must answer a question to which extent the identity (3.6) holds for quantities computed in finite precision arithmetic. Please note that this question is fundamentally different from its trivial part examining the error in *computing* $\nu_{j,d}$ from γ_i and $\mathrm{fl}[(r_i, s_i)]$, where $\mathrm{fl}[\cdot]$ denotes the result of the operation performed in finite precision arithmetic, using (3.7). In order to justify the estimate (3.7), we have to derive the identity for the computed quantities analogous to (3.6) without using any assumption which does not hold in finite precision computations. In particular, we can not use any assumption about orthogonality or finite termination.

The key step considers the *exact* identity for *numerically computed* quantities

$$\begin{aligned} \|x - x_j\|_A^2 &= \|x - x_{j+1} + x_{j+1} - x_j\|_A^2 \\ &= \|x - x_{j+1}\|_A^2 + 2(x - x_{j+1})^T A(x_{j+1} - x_j) + \|x_j - x_{j+1}\|_A^2 \end{aligned}$$

which gives the desired one-step difference

$$\begin{aligned} \|x - x_j\|_A^2 - \|x - x_{j+1}\|_A^2 &= \|x_j - x_{j+1}\|_A^2 \\ &\quad + 2(x - x_{j+1})^T A(x_{j+1} - x_j). \end{aligned} \tag{4.3}$$

The technically complicated and quite tedious analysis which must follow can be summarized in several logically simple steps:

- First, the difference $x_{j+1} - x_j$ is equal to $\gamma_j p_j$ perturbed by inaccuracies due to rounding errors. Consequently, $\|x_{j+1} - x_j\|_A^2$ can be expressed as $\gamma_j(r_j, s_j)$ plus some additional terms depending on machine precision ε characterizing the finite precision arithmetic. These additional terms are small (this is not obvious; the proof requires a careful analysis).
- Second, considering the *approximation* of $A(x - x_{j+1})$ by the residual vector r_{j+1} computed in the $(j+1)$th iteration, the term $2(x - x_{j+1})^T A(x_{j+1} - x_j)$ can be seen as $2\gamma_j(r_{j+1}, p_j)$ plus additional small terms depending on ε (again, bounding the size of these terms needs nontrivial work).

The whole problem of justification of the estimate (3.7) in finite precision arithmetic is in this way reduced to proving that local orthogonality between the computed $(j + 1)$th residual r_{j+1} and the computed jth direction vector p_j is in PCG maintained proportionally to machine precision. This represents the technically most complicated part of the whole analysis.

In following four subsections we present a detailed rounding error analysis of the identity (3.6). Subsection 4.1 describes the rounding errors arising in PCG iterates due to finite precision arithmetic. In Subsection 4.2 we develop a finite precision analogue of the identity (3.6) for $d = 1$. Subsection 4.3 shows that the local orthogonality between the vectors r_{j+1} and p_j is preserved, up to a term proportional to machine precision, in finite precision PCG computation. We finalize the rounding error analysis in Subsection 4.4.

Readers who wish to skip the details of our rounding error analysis may proceed immediately to Subsection 4.4 or even to numerical experiments in Section 5.

4.1 Finite precision PCG computations.

In the analysis we assume the standard model of floating point arithmetic with machine precision ε, see, e.g. [23, (2.4)],

$$\mathrm{fl}[a \circ b] = (a \circ b)(1+\delta), \quad |\delta| \leq \varepsilon, \tag{4.4}$$

where a and b stands for floating-point numbers and the symbol $\circ$ stands for the operations addition, subtraction, multiplication and division. We assume that this model holds also for the square root operation. Under this model, we have for operations involving vectors v, w, a scalar α and the matrix A the following standard results [17], see also [19], [31]

$$\|\alpha\, v - \mathrm{fl}[\alpha\, v]\| \leq \varepsilon\, \|\alpha\, v\|, \tag{4.5}$$

$$\|v + w - \mathrm{fl}[v+w]\| \leq \varepsilon\, (\|v\| + \|w\|), \tag{4.6}$$

$$|(v,w) - \mathrm{fl}[(v,w)]| \leq \varepsilon\, n\, (1 + \mathcal{O}(\varepsilon))\, \|v\|\, \|w\|, \tag{4.7}$$

$$\|Av - \mathrm{fl}[Av]\| \leq \varepsilon\, c\, \|A\| \|v\|. \tag{4.8}$$

When A is a matrix with at most h nonzeros in any row and if the matrix-vector product is computed in the standard way, $c = hn^{1/2}$. In the following analysis we count only for the terms linear in the machine precision ε and express the higher order terms as $\mathcal{O}(\varepsilon^2)$. By $\mathcal{O}(const)$ where *const* is different from ε^2 we denote *const* multiplied by a bounded positive term of an insignificant size which is independent of the *const* and of any other variables present in the bounds.

Numerically, the PCG iterates satisfy

$$x_{j+1} = x_j + \gamma_j p_j + \varepsilon z_j^x, \tag{4.9}$$

$$r_{j+1} = r_j - \gamma_j A p_j + \varepsilon z_j^r, \tag{4.10}$$

$$p_{j+1} = s_{j+1} + \delta_{j+1} p_j + \varepsilon z_j^p, \tag{4.11}$$

where εz_j^x, εz_j^r and εz_j^p account for the local roundoff ($r_0 = b - Ax_0 - \varepsilon f_0$, $\varepsilon \|f_0\| \leq \varepsilon\{\|b\| + \|Ax_0\| + c\|A\|\|x_0\|\} + \mathcal{O}(\varepsilon^2)$). The local roundoff can be bounded according to the standard results (4.5)–(4.8) in the following way

$$\begin{aligned} \varepsilon\, \|z_j^x\| &\leq \varepsilon\, \{\|x_j\| + 2\, \|\gamma_j p_j\|\} + \mathcal{O}(\varepsilon^2) \\ &\leq \varepsilon\, \{3\|x_j\| + 2\|x_{j+1}\|\} + \mathcal{O}(\varepsilon^2), \end{aligned} \tag{4.12}$$

$$\varepsilon\, \|z_j^r\| \leq \varepsilon\, \{\|r_j\| + 2\, \|\gamma_j A p_j\| + c\, \|A\| \|\gamma_j p_j\|\} + \mathcal{O}(\varepsilon^2), \tag{4.13}$$

$$\begin{aligned} \varepsilon\, \|z_j^p\| &\leq \varepsilon\, \{\|s_{j+1}\| + 2\, \|\delta_{j+1} p_j\|\} + \mathcal{O}(\varepsilon^2) \\ &\leq \varepsilon\, \{3\|s_{j+1}\| + 2\|p_{j+1}\|\} + \mathcal{O}(\varepsilon^2). \end{aligned} \tag{4.14}$$

Similarly, the computed coefficients γ_j and δ_j satisfy

$$\gamma_j = \frac{(r_j, s_j)}{(p_j, Ap_j)} + \varepsilon \zeta_j^\gamma, \quad \delta_j = \frac{(r_j, s_j)}{(r_{j-1}, s_{j-1})} + \varepsilon \zeta_j^\delta. \tag{4.15}$$

In order to bound the local terms $|\varepsilon \zeta_j^\gamma|$ and $|\varepsilon \zeta_j^\delta|$ we need following two lemmas.

Lemma 4.1. *Consider the standard model of floating point arithmetic with machine precision* ε [23, 38], $\varepsilon\, n \ll 1$. *Let* L *be a nonsingular lower triangular matrix and* $M = LL^T$. *Then the numerically computed vector* s_{j+1} *is the exact solution of the perturbed system*

$$(M + \Delta M)\, s_{j+1} = r_{j+1}, \quad \|\Delta M\| \le \frac{\varepsilon\, n^2}{1 - \varepsilon\, n}\, \|M\|. \tag{4.16}$$

Proof. To prove (4.16) we use standard results of backward error analysis [23]. Using the Theorem 9.4 [23, p. 175] and the fact that we have exact Cholesky factorization of the matrix $M = LL^T$ we obtain

$$(M + \Delta M)\, s_{j+1} = r_{j+1}, \quad |\Delta M| \le \frac{\varepsilon\, n}{1 - \varepsilon\, n}\, |L||L^T|$$

where $|L|$ denotes the matrix L with elements in absolute value. As shown in the proof of the Theorem 10.4 in [23, p. 206],

$$\|\, |L||L^T|\, \| \le n\, \|M\|.$$

Summarizing,

$$\|\Delta M\| \le \|\, |\Delta M|\, \| \le \frac{\varepsilon\, n}{1 - \varepsilon\, n}\, \|\, |L||L^T|\, \| \le n\, \frac{\varepsilon\, n}{1 - \varepsilon\, n}\, \|M\|$$

which completes the proof. □

Remark. The assumption $M = LL^T$ is not substantial. The result similar to (4.16) and the following analysis, will remain valid also if the Cholesky decomposition of M is computed numerically, see e.g. [17].

Lemma 4.2. *Consider the standard model of floating point arithmetic with machine precision* ε [23, 38], *let* $\varepsilon\, n^2\, \kappa(M) \ll 1$. *The numerically computed inner product* $\mathrm{fl}[(r_j, s_j)]$ *satisfies*

$$\begin{aligned} \mathrm{fl}[(r_j, s_j)] &= (r_j, s_j) + \varepsilon\, \zeta_j^{rs}, \\ \varepsilon|\zeta_j^{rs}| &\le \varepsilon\, \kappa(M)^{1/2} (r_j, s_j)\, \mathcal{O}(n) + \mathcal{O}(\varepsilon^2)\,, \end{aligned} \tag{4.17}$$

where $\kappa(M)$ *denotes the condition number of the matrix* M. *Moreover,* (r_j, s_j) *is bounded from below by*

$$(r_j, s_j) \ge \frac{\|r_j\|\, \|s_j\|}{\kappa(M)^{1/2}}\, \mathcal{O}(1)\,. \tag{4.18}$$

Proof. Using (4.7), $\varepsilon|\zeta_j^{rs}|$ can be bounded as

$$\varepsilon|\zeta_j^{rs}| \le \varepsilon\, n\, \|r_j\|\, \|s_j\| + \mathcal{O}(\varepsilon^2). \tag{4.19}$$

To prove (4.17), we have to relate $\|r_j\|\, \|s_j\|$ to (r_j, s_j). From (4.16) it follows

$$\begin{aligned} \|r_j\|\, \|s_j\| &\le \|r_j\|\, \big\|(M + \Delta M)^{-1} r_j\big\| \\ &= \|r_j\|\, \big\|(I + M^{-1}\Delta M)^{-1} M^{-1} r_j\big\| \\ &\le \|r_j\|\, \big\|M^{-1} r_j\big\|\, \|(I + M^{-1}\Delta M)^{-1}\|\,. \end{aligned} \tag{4.20}$$

Assuming $\varepsilon\, n^2\, \kappa(M) \ll 1$, it holds $\|M^{-1}\Delta M\| \ll 1$ and the matrix inverse $(I+M^{-1}\Delta M)^{-1}$ can be approximated by two terms of the Neumann expansion. Then, (4.20) changes to

$$\|r_j\|\, \|s_j\| \;\leq\; \|r_j\|\, \big\|M^{-1} r_j\big\|\, C_M \,(\, 1 + \mathcal{O}(\|M^{-1}\Delta M\|^2))\,, \tag{4.21}$$

where

$$C_M \;\equiv\; \|\, I - M^{-1}\Delta M\,\|$$

is a constant close to one. It remains to bound the product $\|r_j\|\, \|M^{-1} r_j\|$. A simple manipulation gives

$$\begin{aligned} \|r_j\|\, \big\|M^{-1} r_j\big\| \;&=\; \frac{\|r_j\|\, \big\|M^{-1/2} M^{-1/2} r_j\big\|}{\big(M^{-1/2} r_j, M^{-1/2} r_j\big)}\, \big(r_j, M^{-1} r_j\big) \\ &\leq\; \|M^{-1/2}\| \frac{\|r_j\|}{\big\|M^{-1/2} r_j\big\|}\, (r_j, M^{-1} r_j). \end{aligned} \tag{4.22}$$

Using $M s_j + \Delta M s_j = r_j$ we get

$$\begin{aligned} \big(r_j, M^{-1} r_j\big) \;&=\; (r_j, s_j) + \big(r_j, M^{-1}\Delta M s_j\big) \\ &=\; (r_j, s_j) + \big(M^{-1/2} r_j, M^{-1/2}\Delta M s_j\big) \end{aligned}$$

and $\|r_j\|\, \|M^{-1} r_j\|$ can be bounded by

$$\begin{aligned} \|r_j\|\, \big\|M^{-1} r_j\big\| \;&\leq\; \frac{\|M^{-1/2}\|\, \|r_j\|}{\big\|M^{-1/2} r_j\big\|}\, (r_j, s_j) \\ &\quad + \frac{\|M^{-1/2}\|\, \|r_j\|}{\big\|M^{-1/2} r_j\big\|}\, \big(M^{-1/2} r_j, M^{-1/2}\Delta M s_j\big) \\ &\leq\; \kappa(M)^{1/2} (r_j, s_j) + \frac{\varepsilon\, n^2}{1-\varepsilon\, n}\, \kappa(M)\, \|r_j\|\, \|s_j\|\,. \end{aligned} \tag{4.23}$$

From (4.21) and (4.23) it follows

$$\begin{aligned} \|r_j\|\, \|s_j\| \;&\leq\; \varepsilon\, \kappa(M)^{1/2} (r_j, s_j)\, C_M \\ &\quad + \frac{\varepsilon\, n^2}{1-\varepsilon\, n}\, \kappa(M)\, \|r_j\|\, \|s_j\|\, C_M + \mathcal{O}(\|M^{-1}\Delta M\|^2)\,. \end{aligned} \tag{4.24}$$

Defining

$$D_M \;\equiv\; C_M \left(1 - \frac{\varepsilon\, n^2}{1-\varepsilon\, n}\, \kappa(M)\, C_M \right)^{-1},$$

(4.24) can be written in the form

$$\|r_j\|\, \|s_j\| \leq \kappa(M)^{1/2} (r_j, s_j)\, D_M \;+\; \mathcal{O}(\|M^{-1}\Delta M\|^2)\,. \tag{4.25}$$

Since $\varepsilon\, n^2\, \kappa(M) \ll 1$ and C_M is close to one, the definition of D_M implies that D_M is close to one also. The term $\mathcal{O}(\|M^{-1}\Delta M\|^2)$ is under our assumption unimportant and will not be further explicitly considered. Finally, (4.25) gives

$$\|r_j\|\, \|s_j\| \leq \kappa(M)^{1/2} (r_j, s_j)\, \mathcal{O}(1)\,, \tag{4.26}$$

where $\mathcal{O}(1)$ stands for a number close to one. (4.17) follows immediately from (4.26) and (4.19). Dividing (4.26) by $\kappa(M)^{1/2}$ gives (4.18), which finishes the proof. □

Assuming $\varepsilon\, n^2\, \kappa(M) \ll 1$, the local term $\varepsilon\zeta_j^\delta$ is bounded, according to (4.4), (4.7) and (4.17), by

$$\varepsilon\big|\zeta_j^\delta\big| \leq \varepsilon\,\frac{(r_j, s_j)}{(r_{j-1}, s_{j-1})}\,\kappa(M)^{1/2}\,\mathcal{O}(n) + \mathcal{O}(\varepsilon^2). \tag{4.27}$$

Using (4.5)–(4.8) and $\|A\|\|p_j\|^2/(p_j, Ap_j) \leq \kappa(A)$,

$$\begin{aligned}\mathrm{fl}[(p_j, Ap_j)] &= (p_j, Ap_j) + \varepsilon\,\|Ap_j\|\|p_j\|\mathcal{O}(n) + \varepsilon\,\|A\|\|p_j\|^2\mathcal{O}(c) + \mathcal{O}(\varepsilon^2)\\ &= (p_j, Ap_j)\big(1 + \varepsilon\,\kappa(A)\mathcal{O}(n+c)\big) + \mathcal{O}(\varepsilon^2).\end{aligned}$$

Assuming $\varepsilon(n+c)\,\kappa(A) \ll 1$, the local roundoff $\varepsilon\zeta_j^\gamma$ is bounded by

$$\varepsilon\big|\zeta_j^\gamma\big| \leq \varepsilon\,(\kappa(A) + \kappa(M)^{1/2})\frac{(r_j, s_j)}{(p_j, Ap_j)}\,\mathcal{O}(n+c) + \mathcal{O}(\varepsilon^2). \tag{4.28}$$

It is well known that in finite precision arithmetic the true residual $b - Ax_j$ differs from the recursively updated residual vector r_j,

$$r_j = b - Ax_j - \varepsilon f_j. \tag{4.29}$$

This topic was studied in [36] and [19]. The results can be written in the following form

$$\|\varepsilon f_j\| \leq \varepsilon\,\|A\|\,(\|x\| + \max_{0\leq i\leq j}\|x_i\|)\,\mathcal{O}(jc), \tag{4.30}$$

$$\|r_j\| = \|b - Ax_j\|\,(1 + \varepsilon F_j), \tag{4.31}$$

where εF_j is bounded by

$$|\varepsilon F_j| = \frac{|\|r_j\| - \|b - Ax_j\||}{\|b - Ax_j\|} \leq \frac{\|r_j - (b - Ax_j)\|}{\|b - Ax_j\|} = \frac{\varepsilon\|f_j\|}{\|b - Ax_j\|}. \tag{4.32}$$

Rounding errors affect results of PCG computations in two main ways: they delay convergence and limit the ultimate attainable accuracy. Here we are primarily interested in estimating the convergence rate. We therefore assume that the final accuracy level has not been reached yet and εf_j is, in comparison to the size of the true and iterative residuals, small. In the subsequent text we will relate the numerical inaccuracies to the A-norm of the error $\|x - x_j\|_A$. The following inequalities derived from (4.32) will prove useful,

$$\lambda_1^{1/2}\,\|x - x_j\|_A\,(1 + \varepsilon\,F_j) \leq \|r_j\| \leq \lambda_n^{1/2}\,\|x - x_j\|_A\,(1 + \varepsilon\,F_j). \tag{4.33}$$

Similarly as in the ordinary CG (see [18], [21]) we can argue that the monotonicity of the A-norm is in PCG preserved (with small additional inaccuracy) also in finite precision computations. Using this fact we get for $j \geq i$

$$\varepsilon\,\frac{\|r_j\|}{\|r_i\|} \leq \varepsilon\,\frac{\lambda_n^{1/2}}{\lambda_1^{1/2}}\cdot\frac{\|x - x_j\|_A}{\|x - x_i\|_A}\cdot\frac{(1 + \varepsilon\,F_j)}{(1 + \varepsilon\,F_i)} \leq \varepsilon\,\kappa(A)^{1/2} + \mathcal{O}(\varepsilon^2). \tag{4.34}$$

This bound will be used later.

4.2 Finite precision analysis – basic identity.

We show that the ideal (exact precision) identity (3.6) changes numerically to

$$\|x - x_j\|_A^2 = \nu_{j,d} + \|x - x_{j+d}\|_A^2 + \widetilde{\nu}_{j,d} \tag{4.35}$$

where $\widetilde{\nu}_{j,d}$ is as small as it can be. We once more emphasize that the difference between (3.6) and (4.35) *is not trivial.* The ideal and numerical counterparts of each individual term in these identities may be orders of magnitude different! Due to the facts that rounding errors in computing $\nu_{j,d}$ numerically from the quantities γ_i and $\mathrm{fl}[(r_i, s_i)]$ are negligible and that $\widetilde{\nu}_{j,d}$ will be related to $\varepsilon\,\|x - x_j\|_A$, (4.35) will justify the estimate $\nu_{j,d}$ in finite precision computations.

In order to get the desired form leading to (4.35), we will develop the right hand side of (4.3). In this derivation we will rely on local properties (4.9)–(4.11) and (4.15)–(4.16) of the finite precision PCG recurrences.

Using (4.9), the first term on the right hand side of (4.3) can be written as

$$\begin{aligned}\|x_{j+1} - x_j\|_A^2 &= (\gamma_j p_j + \varepsilon\, z_j^x)^T A(\gamma_j p_j + \varepsilon\, z_j^x)\\ &= \gamma_j^2\,(p_j, Ap_j) + 2\varepsilon\,\gamma_j (p_j, Az_j^x) + \mathcal{O}(\varepsilon^2)\\ &= \gamma_j\,(p_j, Ap_j) + 2\varepsilon\,(x_{j+1} - x_j)^T Az_j^x + \mathcal{O}(\varepsilon^2).\end{aligned} \tag{4.36}$$

Similarly, the second term on the right hand side of (4.3) transforms, using (4.29), to the form

$$\begin{aligned}2\,(x - x_{j+1})^T A(x_{j+1} - x_j) &= 2\,(r_{j+1} + \varepsilon\, f_{j+1})^T (x_{j+1} - x_j)\\ &= 2\, r_{j+1}^T (x_{j+1} - x_j) + 2\varepsilon\, f_{j+1}^T (x_{j+1} - x_j).\end{aligned} \tag{4.37}$$

Combining (4.3), (4.36) and (4.37),

$$\begin{aligned}\|x - x_j\|_A^2 - \|x - x_{j+1}\|_A^2 = \gamma_j^2\,(p_j, Ap_j) &+ 2\, r_{j+1}^T (x_{j+1} - x_j)\\ &+ 2\varepsilon\,(f_{j+1} + Az_j^x)^T (x_{j+1} - x_j) + \mathcal{O}(\varepsilon^2).\end{aligned} \tag{4.38}$$

Substituting for γ_j from (4.15), the first term in (4.38) can be written as

$$\begin{aligned}\gamma_j^2\,(p_j, Ap_j) &= \gamma_j (r_j, s_j) + \varepsilon\,\gamma_j\,(p_j, Ap_j)\,\zeta_j^\gamma\\ &= \gamma_j (r_j, s_j) + \varepsilon\,\gamma_j (r_j, s_j) \left\{ \zeta_j^\gamma \frac{(p_j, Ap_j)}{(r_j, s_j)} \right\}.\end{aligned}$$

Consequently, the difference between the squared A-norms of the error in the consecutive steps can be written in the form convenient for the further analysis

$$\begin{aligned}\|x - x_j\|_A^2 - \|x - x_{j+1}\|_A^2 \;=\; \gamma_j (r_j, s_j) + \varepsilon\,\gamma_j (r_j, s_j) \left\{ \zeta_j^\gamma \frac{(p_j, Ap_j)}{(r_j, s_j)} \right\}\\ + 2\, r_{j+1}^T (x_{j+1} - x_j) + 2\varepsilon\,(f_{j+1} + Az_j^x)^T (x_{j+1} - x_j) + \mathcal{O}(\varepsilon^2).\end{aligned} \tag{4.39}$$

The goal of the following analysis is to show that until $\|x - x_j\|_A$ reaches its ultimate attainable accuracy level, the terms on the right hand side of (4.39) are, except for $\gamma_j(r_j, s_j)$ insignificant. Bounding the second term will not represent a problem. The norm of the difference $x_{j+1} - x_j = (x - x_j) - (x - x_{j+1})$ is bounded by $2\|x - x_j\|_A/\lambda_1^{1/2}$, and therefore the size of the fourth term is proportional to $\varepsilon\,\|x - x_j\|_A$. The third term is related to the line-search principle. Ideally (in exact arithmetic), the $(j+1)$-th residual $\hat{r}_{j+1}$ is orthogonal to the difference between the $(j+1)$-th and j-th approximation $\hat{x}_{j+1} - \hat{x}_j$ (which is a multiple of the j-th direction vector $\hat{p}_j$). This is equivalent to the line-search: ideally, in terms of the transformed quantities used in Algorithm 2, the $(j+1)$-th PCG approximation minimizes the A-norm of the error along the line determined by the j-th approximation and the j-th direction vector. Here the term $r_{j+1}^T(x_{j+1} - x_j)$, with r_{j+1}, x_j and x_{j+1} computed numerically, examines how closely the line-search holds in finite precision arithmetic. In fact, bounding the local orthogonality $r_{j+1}^T(x_{j+1} - x_j)$ represents the technically most difficult part of the remaining analysis.

4.3 Local orthogonality.

Since the classical work of Paige it is well known that in the three-term Lanczos recurrence local orthogonality is preserved close to the machine epsilon (see [31]). We will derive an analogue of this for the PCG algorithm, and state it as an independent result.

The local orthogonality term $r_{j+1}^T(x_{j+1} - x_j)$ can be written in the form

$$r_{j+1}^T(x_{j+1} - x_j) = r_{j+1}^T(\gamma_j p_j + \varepsilon\, z_j^x) = \gamma_j (r_{j+1}, p_j) + \varepsilon\,(r_{j+1}, z_j^x). \tag{4.40}$$

Using the bound

$$\|r_{j+1}\| \le \lambda_n^{1/2}\|x - x_{j+1}\|_A(1 + \varepsilon\, F_{j+1}) \le \lambda_n^{1/2}\|x - x_j\|_A(1 + \varepsilon\, F_{j+1})\,,$$

see (4.33), the size of the second term in (4.40) is proportional to $\varepsilon\,\|x - x_j\|_A$. The main step consist of showing that the term (r_{j+1}, p_j) is sufficiently small. Scalar multiplying the recurrence (4.10) for r_{j+1} by the vector p_j gives (using (4.11) and (4.15))

$$\begin{aligned}(p_j, r_{j+1}) &= (p_j, r_j) - \gamma_j (p_j, Ap_j) + \varepsilon\,(p_j, z_j^r)\\ &= (s_j + \delta_j p_{j-1} + \varepsilon\, z_{j-1}^p)^T r_j\\ &\quad - \left(\frac{(r_j, s_j)}{(p_j, Ap_j)} + \varepsilon\,\zeta_j^\gamma\right)(p_j, Ap_j) + \varepsilon\,(p_j, z_j^r)\\ &= \delta_j\,(p_{j-1}, r_j) + \varepsilon\left\{(r_j, z_{j-1}^p) - \zeta_j^\gamma (p_j, Ap_j) + (p_j, z_j^r)\right\}.\end{aligned} \tag{4.41}$$

Denoting

$$G_j \equiv (r_j, z_{j-1}^p) - \zeta_j^\gamma (p_j, Ap_j) + (p_j, z_j^r), \tag{4.42}$$

the identity (4.41) is

$$(p_j, r_{j+1}) = \delta_j\,(p_{j-1}, r_j) + \varepsilon\, G_j. \tag{4.43}$$

Recursive application of (4.43) for $(p_{j-1}, r_j), \ldots, (p_1, r_2)$ with $(p_0, r_1) = (p_0, r_0) - \gamma_0\,(p_0, Ap_0) + \varepsilon\,(p_0, z_0^r) = \varepsilon \left\{ - \zeta_0^\gamma\,(s_0, As_0) + (s_0, z_0^r) \right\} \equiv \varepsilon\, G_0$, gives

$$(p_j, r_{j+1}) = \varepsilon\, G_j + \varepsilon \sum_{i=1}^{j} \left(\prod_{k=i}^{j} \delta_k \right) G_{i-1}. \tag{4.44}$$

Since

$$\varepsilon \prod_{k=i}^{j} \delta_k = \varepsilon \prod_{k=i}^{j} \frac{(r_k, s_k)}{(r_{k-1}, s_{k-1})} + \mathcal{O}(\varepsilon^2) = \varepsilon\, \frac{(r_j, s_j)}{(r_{i-1}, s_{i-1})} + \mathcal{O}(\varepsilon^2),$$

we can express (4.44) as

$$(p_j, r_{j+1}) = \varepsilon\,(r_j, s_j) \sum_{i=0}^{j} \frac{G_i}{(r_i, s_i)} + \mathcal{O}(\varepsilon^2). \tag{4.45}$$

Using (4.42),

$$\frac{|G_i|}{(r_i, s_i)} \le \frac{\|r_i\| \|z_{i-1}^p\|}{(r_i, s_i)} + |\zeta_i^\gamma| \frac{(p_i, Ap_i)}{(r_i, s_i)} + \frac{\|p_i\| \|z_i^r\|}{(r_i, s_i)}. \tag{4.46}$$

When bounding the first and the last terms on the right hand side of (4.46), we will use the inequality (4.18) proved in Lemma 4.2. From (4.14) it follows

$$\varepsilon\, \frac{\|r_i\| \|z_{i-1}^p\|}{(r_i, s_i)} \le \varepsilon\, \kappa(M)^{1/2} \left\{ 3 + 2 \frac{\|p_i\|}{\|s_i\|} \right\} \mathcal{O}(1) \;+\; \mathcal{O}(\varepsilon^2). \tag{4.47}$$

Using (4.28),

$$\varepsilon\, |\zeta_i^\gamma| \frac{(p_i, Ap_i)}{(r_i, s_i)} \le \varepsilon\, (\kappa(A) + \kappa(M)^{1/2})\, \mathcal{O}(n + c) + \mathcal{O}(\varepsilon^2). \tag{4.48}$$

The last part of (4.46) is bounded using (4.13) and (4.18)

$$\begin{aligned} \varepsilon\, \frac{\|p_i\| \|z_i^r\|}{(r_i, s_i)} &\le \varepsilon \left\{ \kappa(M)^{1/2} \frac{\|p_i\| \|r_i\|}{\|s_i\| \|r_i\|}\, \mathcal{O}(1) \right\} \\ &\quad + \varepsilon \left\{ 2\, \gamma_i \frac{\|p_i\| \|Ap_i\|}{(r_i, s_i)} + c\, \gamma_i \frac{\|p_i\| \|A\| \|p_i\|}{(r_i, s_i)} \right\} + \mathcal{O}(\varepsilon^2) \\ &= \varepsilon \left\{ \kappa(M)^{1/2} \frac{\|p_i\|}{\|s_i\|}\, \mathcal{O}(1) \right\} \\ &\quad + \varepsilon \left\{ 2\, \frac{\|p_i\| \|Ap_i\|}{(p_i, Ap_i)} + c\, \frac{\|A\| \|p_i\|^2}{(p_i, Ap_i)} \right\} + \mathcal{O}(\varepsilon^2) \\ &\le \varepsilon \left\{ \kappa(M)^{1/2} \frac{\|p_i\|}{\|s_i\|}\, \mathcal{O}(1) + (2 + c)\, \kappa(A) \right\} + \mathcal{O}(\varepsilon^2), \end{aligned} \tag{4.49}$$

where

$$\begin{aligned} \varepsilon\, \frac{\|p_i\|}{\|s_i\|} &\le \varepsilon\, \frac{\|s_i\| + \delta_i \|p_{i-1}\|}{\|s_i\|} + \mathcal{O}(\varepsilon^2) \\ &\le \varepsilon \left\{ 1 + \delta_i \frac{\|s_{i-1}\|}{\|s_i\|} \frac{\|p_{i-1}\|}{\|s_{i-1}\|} \right\} + \mathcal{O}(\varepsilon^2). \end{aligned} \tag{4.50}$$

Recursive application of (4.50) for $\|p_{i-1}\|/\|s_{i-1}\|$, $\|p_{i-2}\|/\|s_{i-2}\|$, …, $\|p_1\|/\|s_1\|$ with $\|p_0\|/\|s_0\| = 1$ gives

$$\begin{aligned}
\varepsilon\frac{\|p_i\|}{\|s_i\|} &\leq \varepsilon\left\{1 + \frac{(s_i, r_i)}{(s_{i-1}, r_{i-1})}\frac{\|s_{i-1}\|}{\|s_i\|} + \cdots + \frac{(s_i, r_i)}{(s_0, r_0)}\frac{\|s_0\|}{\|s_i\|}\right\} + \mathcal{O}(\varepsilon^2)\\
&\leq \varepsilon\left\{1 + \frac{\|r_i\|\|s_{i-1}\|}{(s_{i-1}, r_{i-1})} + \cdots + \frac{\|r_i\|\|s_0\|}{(s_0, r_0)}\right\} + \mathcal{O}(\varepsilon^2)\\
&\leq \varepsilon\left\{1 + \kappa(M)^{1/2}\frac{\|r_i\|}{\|r_{i-1}\|} + \cdots + \kappa(M)^{1/2}\frac{\|r_i\|}{\|r_0\|}\right\}\mathcal{O}(1) + \mathcal{O}(\varepsilon^2).
\end{aligned}$$

The size of $\varepsilon\,\|r_i\|/\|r_k\|$, $i \geq k$ is, according to (4.34), less or equal than the value $\varepsilon\,\kappa(A)^{1/2} + \mathcal{O}(\varepsilon^2)$. Consequently,

$$\varepsilon\,\frac{\|p_i\|}{\|s_i\|} \leq \varepsilon\,\{1 + i\,\kappa(A)^{1/2}\kappa(M)^{1/2}\}\,\mathcal{O}(1) + \mathcal{O}(\varepsilon^2). \tag{4.51}$$

Denote

$$\kappa(A, M) \equiv \max(\kappa(A), \kappa(M)\kappa(A)^{1/2}).$$

Summarizing (4.47), (4.48), (4.49) and (4.51), the ratio $\varepsilon\,|G_i|/(r_i, s_i)$ is bounded as

$$\varepsilon\,\frac{|G_i|}{(r_i, s_i)} \leq \varepsilon\,\kappa(A, M)\,\mathcal{O}(8 + 3c + 2n + 3i) + \mathcal{O}(\varepsilon^2). \tag{4.52}$$

Combining this result with (4.45) proves the following theorem.

Theorem 4.3. *Let $\varepsilon\,(n + c)\,\kappa(A) \ll 1$, $\varepsilon\,n^2\,\kappa(M) \ll 1$. Then the local orthogonality between the direction vectors and the iteratively computed residuals is in the finite precision implementation of the preconditioned conjugate gradient method* (4.9)–(4.11) *and* (4.15)–(4.16) *bounded by*

$$|(p_j, r_{j+1})| \leq \varepsilon\,(r_j, s_j)\,\kappa(A, M)\,\mathcal{O}((j + 1)(8 + 3c + 2n + 3j)) + \mathcal{O}(\varepsilon^2) \tag{4.53}$$

where

$$\kappa(A, M) \equiv \max\left(\kappa(A), \kappa(M)\kappa(A)^{1/2}\right).$$

4.4 Finite precision analysis – conclusions.

We now return to (4.39) and finalize our discussion. Using (4.40) and (4.45),

$$\begin{aligned}
\|x - x_j\|_A^2 - \|x - x_{j+1}\|_A^2 &= \gamma_j(r_j, s_j)\\
&+ \varepsilon\,\gamma_j(r_j, s_j)\left\{\zeta_j^\gamma\frac{(p_j, Ap_j)}{(r_j, s_j)} + 2\sum_{i=0}^{j}\frac{G_i}{(r_j, s_j)}\right\}\\
&+ 2\varepsilon\left\{(f_{j+1} + Az_j^x)^T(x_{j+1} - x_j) + (r_{j+1}, z_j^x)\right\} + \mathcal{O}(\varepsilon^2).
\end{aligned} \tag{4.54}$$

The term

$$E_j^{(1)} \equiv \varepsilon\left\{\zeta_j^\gamma\frac{(p_j, Ap_j)}{(r_j, s_j)} + 2\sum_{i=0}^{j}\frac{G_i}{(r_j, s_j)}\right\}$$

is bounded using (4.28) and (4.52),

$$|E_j^{(1)}| \;\leq\; \varepsilon\kappa(A,M)\,\mathcal{O}\big(2n+2c+2(j+1)(8+3c+2n+3j)\big) \;+\; \mathcal{O}(\varepsilon^2). \tag{4.55}$$

We write the remaining term on the right hand side of (4.54) proportional to ε

$$2\varepsilon\left\{(f_{j+1}+Az_j^x)^T(x_{j+1}-x_j)+(r_{j+1},z_j^x)\right\} \equiv \|x-x_j\|_A\,E_j^{(2)} \tag{4.56}$$

where

$$\begin{aligned} |E_j^{(2)}| &= 2\varepsilon\left|(f_{j+1}+Az_j^x)^T\left(\frac{x_{j+1}-x+x-x_j}{\|x-x_j\|_A}\right)+\frac{(r_{j+1},z_j^x)}{\|x-x_j\|_A}\right| \\ &\leq 2\varepsilon\left\{2\,(\|f_{j+1}\|\lambda_1^{-1/2}+\|A\|^{1/2}\|z_j^x\|)+\|A\|^{1/2}\|z_j^x\|\right\}. \end{aligned} \tag{4.57}$$

With (4.30) and (4.12),

$$\begin{aligned} |E_j^{(2)}| &\leq 4\varepsilon\|A\|^{1/2}\kappa(A)^{1/2}(\|x\|+\max_{0\leq i\leq j+1}\|x_i\|)\,\mathcal{O}(jc) \\ &\quad+5\|A\|^{1/2}\varepsilon(3\|x_j\|+2\|x_{j+1}\|)+\mathcal{O}(\varepsilon^2) \\ &\leq \varepsilon\|A\|^{1/2}\kappa(A)^{1/2}(\|x\|+\max_{0\leq i\leq j+1}\|x_i\|)\,\mathcal{O}(4jc+25)+\mathcal{O}(\varepsilon^2). \end{aligned} \tag{4.58}$$

Finally, using the fact that the monotonicity of the A-norm is with small additional inaccuracy preserved also in finite precision PCG computations (see also the discussion following (4.33)), we obtain the finite precision analogue of (3.6), which is formulated as a theorem.

THEOREM 4.4. *Let $\varepsilon\,(n+c)\,\kappa(A) \ll 1$, $\varepsilon\,n^2\,\kappa(M) \ll 1$. Then the PCG approximate solutions computed in finite precision arithmetic satisfy*

$$\|x-x_j\|_A^2-\|x-x_{j+d}\|_A^2=\nu_{j,d}\;+\;\nu_{j,d}\,E_{j,d}^{(1)}\;+\;\|x-x_j\|_A\,E_{j,d}^{(2)}+O(\varepsilon^2), \tag{4.59}$$

where

$$\nu_{j,d}\;=\;\sum_{i=j}^{j+d-1}\gamma_i\,(r_i,s_i). \tag{4.60}$$

The terms due to rounding errors are bounded by

$$\begin{aligned} &|E_{j,d}^{(1)}| \leq \varepsilon\,\kappa(A,M)\,p^{(1)}(n,d)+O(\varepsilon^2), \\ &|E_{j,d}^{(2)}| \leq \varepsilon\,\|A\|^{1/2}\kappa(A)^{1/2}\,(\|x\|+\max_{0\leq i\leq j+1}\|x_i\|)\,p^{(2)}(n,d)+O(\varepsilon^2), \end{aligned} \tag{4.61}$$

where

$$\kappa(A,M)\equiv\max\,(\kappa(A),\kappa(M)\kappa(A)^{1/2}),$$

$p^{(1)}(n,d)$ and $p^{(2)}(n,d)$ represent small degree polynomials in n and d independent of any other variables.

Based on the assumptions we consider $|E_{j,d}^{(1)}| \ll 1$. Then, assuming that the A-norm of the error reasonably decreases, the numerically computed value $\nu_{j,d}$

gives a good estimate for the A-norm of the error $\|x - x_j\|_A^2$ until

$$\|x - x_j\|_A \, |E_{j,d}^{(2)}| \ll \|x - x_j\|_A^2,$$

which is equivalent to

$$\|x - x_j\|_A \gg |E_{j,d}^{(2)}|. \tag{4.62}$$

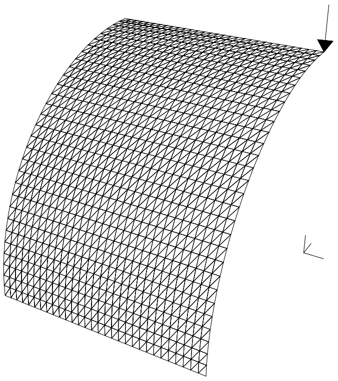
The quantity $E_{j,d}^{(2)}$ represents various terms. Its upper bound is, apart from $\kappa(A)^{1/2}$, which comes into play as an effect of the worst-case rounding error analysis, linearly dependent on an upper bound for $\|x - x_0\|_A$. The value of $E_{j,d}^{(2)}$ is (similar to terms or constants in any other rounding error analysis) not important. What is important is the following possible interpretation of (4.62): until $\|x - x_j\|_A$ reaches a level close to $\varepsilon \|x - x_0\|_A$, the computed estimate $\nu_{j,d}^{1/2}$ must work.

Please note that $\nu_{j,d}$ represents here the exact value determined from the computed inputs γ_i, r_i and s_i. In fact, we should consider the computed value $\mathrm{fl}[\nu_{j,d}]$. Additional rounding errors in evaluating the formula (4.60) are, however, negligible in comparison to the other rounding error terms in (4.59), and need not be considered here.

5 Numerical experiments.

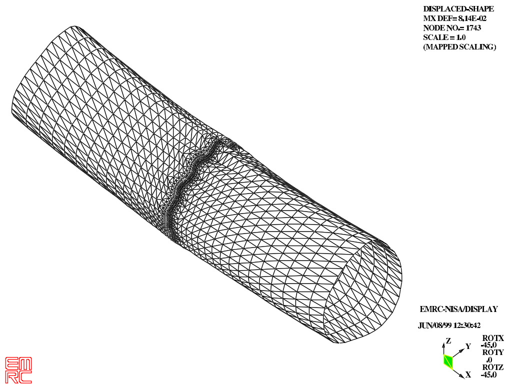
We test our theoretical results on three linear systems with a symmetric positive definite matrix A. The first two systems (by R. Kouhia) arise from cylindrical shell modeling. The matrices are large and sparse, and PCG represents a natural choice for solving the systems in practical computations. The third system (by P. Benner) appears in large-scale control problems. PCG is not used for practical solution of the last (rather small) system. We use it here for illustration of how the estimate of the A-norm of the error works for this type of problem. We describe the problems in more detail.

The system `s3dkt3m2`. The collection Cylshell (by R. Kouhia) from the electronic library Matrix Market [25] contains matrices that represent low order finite element discretization of a shell element test, the pinched cylinder. An illustration of the mesh for this problem provided by R. Kouhia is given below.



In our experiments we use the matrix `s3dkt3m2` of the order $n = 90449$. The matrix has $\mathrm{nnz}(A) = 1921955$ nonzero elements, and the condition number $\kappa(A) = 3.62\mathrm{e}{+}11$. Only the last element of the right-hand side vector b is nonzero, which corresponds to the given physical problem (for more details see [24] and the references in [24]). The preconditioner was determined by incomplete Cholesky decomposition with no fill-in.

The system `tube`. The second system is given at the R. Kouhia's homepage `http://www.hut.fi/~kouhia/` (the system `tube1-2`). The tube is a cylindrical shell with the constant wall thickness, loaded with an axial stress distribution at both ends. The mesh is refined at the center, and it is almost uniform towards the ends.



The order of the matrix A is $n = 21498$, $\mathrm{nnz}(A) = 894490$. The factor L of the preconditioner M is determined by the incomplete Cholesky decomposition with the drop tolerance 1e–5, $\mathrm{nnz}(L) = 4384369$.

The system `stahl`. We consider the problem of optimal cooling of steel profile, that arises, e.g. in a rolling mill when different steps in the production process require different temperatures of the raw material. The problem is modeled using a boundary control (given by the temperature of the cooling fluid) for a heat-diffusion process described by the linearized heat equations. This leads to the Lyapunov equations that are solved by the ADI iterations. For more detail about this problem see [9]. We test the proposed estimates on the system from the initial step of the ADI iteration. The matrix is of the order $n = 5177$, $\kappa(A) = 1.56\mathrm{e}{+}05$, $\mathrm{nnz}(A) = 35241$. The system is preconditioned by incomplete Cholesky decomposition with no fill-in.

In all experiments we use the initial approximation $x_0 = 0$. We do not tune the preconditioner for the best performance; our aim is to demonstrate the behaviour of the estimate of the A-norm of the error in practical computations. The substitutes for the exact solutions x used in the figures are for each system computed in two steps: 1. We apply PCG to the system and iterate until ultimate level of accuracy is reached (the norm of true and recursive residuals start to differ). 2. We apply PCG to the system for the second time, with the initial approximation given by the approximate solution computed in the first step. In this way, we obtain approximate solutions that represent for our purpose sufficiently accurate approximations to the exact solutions x. Our numerical experiments showed that even for the first step the obtained residual norms were comparable with that ones obtained by the direct Cholesky decomposition solver. After the second step the residual norms further decreased, but less than by a factor of 10.

In experiments with the system `s3dkt3m2` we use a Fortran program CG6 provided us by M. Tůma. The other two systems are solved using our implementation of PCG in Matlab 6.5; we use the Matlab-function `cholinc` to determine the incomplete Cholesky decomposition of the matrix A. All experiments were performed on a AMD Athlon XP 2100+ personal computer with machine precision $\varepsilon \sim 10^{-16}$.

5.1 Estimates for the A-norm of the error.

In the first numerical experiment we test the estimate $\nu_{j,d}^{1/2}$ of the A-norm of the error and the estimate $\varrho_{j,d}^{1/2}$ of the relative A-norm of the error in PCG applied to the three systems described above. The results are presented in the figures Figure 5.1 (`s3dkt3m2`), Figure 5.2 (`tube`) and Figure 5.3 (`stahl`). All three figures consist of two parts. The left part includes various convergence characteristics: the A-norm of the error $\|x - x_j\|_A$ (dashed line), its estimate $\nu_{j,d}^{1/2}$ for some particular value of the parameter d (bold solid line), the residual norm $\|b - Ax_j\|$ (dash-dotted line) and the normwise backward error $\|b - Ax_j\|/(\|A\|\,\|x_j\| + \|b\|)$ (dotted line). In the right part of the figure we plot the relative A-norm of the error $\|x - x_j\|_A/\|x\|_A$ (dashed line) and its estimates $\varrho_{j,d}^{1/2}$ for different values of d (solid lines). The bold line corresponds to the same value of d as the bold line in the left part of the figure.

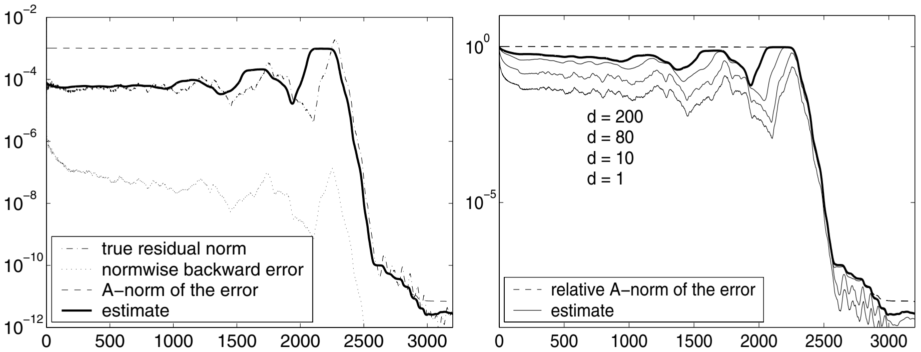


Figure 5.1: The system `s3dkt3m2`. In an extremal case of very slow PCG convergence the estimate $\nu_{j,d}^{1/2}$ can significantly underestimate the actual A-norm of the error (left part). The estimate $\varrho_{j,d}^{1/2}$ of the relative A-norm of the error (right part) is in general much tighter than the estimate of the A-norm of the error.

Figure 5.1 (`s3dkt3m2`), *left part.* We start with the most difficult situation when the A-norm of the error (dashed line) almost stagnates for many steps (here up to the iteration $\sim$ 2400). Then the estimate $\nu_{j,d}^{1/2}$ (bold solid line) can give a poor information about the actual A-norm of the error. The values of $\|x - x_j\|_A$ and $\nu_{j,d}^{1/2}$, can significantly differ even for a considerably large value of the parameter d (here $d = 200$). Please notice that the situation just described is not frequent in practical computations. It corresponds to an extremely slow convergence of PCG, i.e. to the case of very difficult problem which is hard to precondition. We have chosen such problem on purpose to show the possible drawback of the proposed error estimator. We emphasize that this situation represents an extremal case. Typical situation is demonstrated below on Figure 5.2 (`tube`) and Figure 5.3 (`stahl`). As soon as the convergence takes place (around the iteration 2400), we get a tight lower bound for the A-norm of the error.

In CG, we often observe a close correlation between the behaviour of the residual norm and the estimate $\nu_{j,d}^{1/2}$ for small values of d. This is a consequence of the fact that in ordinary CG the coefficients γ_j usually oscillate around some value and, apart from this oscillations, the behaviour of $\|r_j\|$ determines the behaviour of $\nu_{j,d}^{1/2}$. Similar phenomenon appears also in the PCG iterations. Here $\nu_{j,d}$ and $(r_j, M^{-1}r_j)$ (the squared M^{-1}-norm of the residual r_j) are correlated for small values of d. The M^{-1}-norm of the residual r_j frequently behaves in practical computation similarly as a constant multiple of the Euclidean norm of the residual. Then the correlation between $\|r_j\|$ and $\nu_{j,d}^{1/2}$ is observed also in the PCG iterations. For larger values of d, however, there is, in general, no correlation between the behaviour of $\|r_j\|$ and $\nu_{j,d}^{1/2}$. In the left part of Figure 5.1 (where $d = 200$) we clearly see periods of decrease of $\|r_j\|$ with simultaneous increase of $\nu_{j,d}^{1/2}$, and vice versa.

By the dotted line we plot the normwise backward error. After the convergence becomes steady, the values of $\|x_j\|$ typically stabilize. The residual norm and the normwise backward error are then in a strong correlation. Until then, however, both characteristics can behave differently. This fact is demonstrated by the convergence curves in the first 500 iterations; the backward error decreases while the residual norm stagnates.

Figure 5.1 (`s3dkt3m2`), *right part.* In the right part of the Figure 5.1 we plot the relative A-norm of the error (3.8) (dashed line) and its estimate $\varrho_{j,d}^{1/2}$ for $d = 1$, $d = 10$, $d = 80$ (solid lines) and $d = 200$ (bold solid line). The estimate $\varrho_{j,1}^{1/2}$, and sometimes even $\varrho_{j,10}^{1/2}$, $\varrho_{j,80}^{1/2}$ and $\varrho_{j,200}^{1/2}$, are not tight when the A-norm of the error almost stagnates. In the other cases $\varrho_{j,1}^{1/2}$ as well as the bounds for the larger d are close to the considered convergence curve. By the bold solid line we plot the estimate for $d = 200$. In comparison to the left part of the Figure 5.1, the estimate of the relative A-norm of the error gives better results (it is closer to the approximated curve) than the estimate of the absolute A-norm of the error.

Figure 5.2 (`tube`), *left part.* When the A-norm of the error (dashed line) decreases rapidly (iterations $350 - 400$), we can not visually distinguish this quantity from its estimate $\nu_{j,d}^{1/2}$ (bold solid line). On the other hand, when the convergence is slow (iterations $1-350$), the difference between the actual A-norm of the error and its estimate is observable but insignificant. The normwise backward

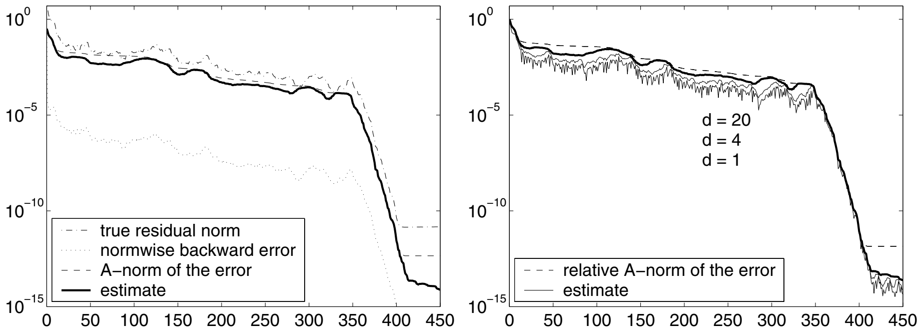


Figure 5.2: The system `tube`. Even a slow decrease of the A-norm of the error is sufficient for obtaining a satisfactory value of the estimate $\nu_{j,d}^{1/2}$ of the A-norm of the error. The erratic behaviour for $d = 1$ is caused by the oscillations of the coefficients γ_j (right part). By increasing the value of d, the curves are more smooth and closer to the relative A-norm of the error.

error (dotted line) behaves similarly, apart from the difference in magnitude, as the residual norm (dash dotted line).

Figure 5.2 (`tube`), *right part.* The right part of the Figure 5.2 contains the curve of the relative A-norm of the error (dashed line) and its estimates for $d = 1$, $d = 4$ (solid lines) and $d = 20$ (bold solid line). For $d = 1$, the curve of the estimate is erratic. The irregularity of the curve is due to the oscillations of the coefficients γ_j. The estimate $\varrho_{j,1}^{1/2}$ does not differ from the actual relative A-norm of the error for more than a single order of magnitude, although the convergence is in iterations 1–350 slow. Increasing d provides a very good estimate throughout the whole computation.

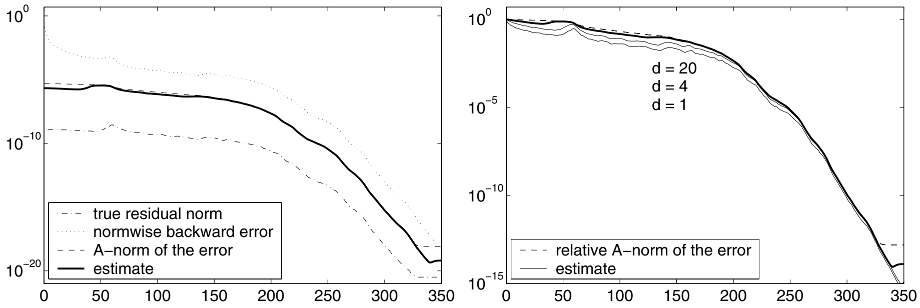


Figure 5.3: The system `stahl`. The estimates for the absolute and relative A-norm of the error are tight throughout the whole computation.

Figure 5.3 (`stahl`), *left part.* The preconditioning by incomplete Cholesky decomposition represents here a very good choice; the convergence of the A-norm of the error (dashed line) is fast during the whole computation and the estimate (bold solid line) for the parameter $d = 20$ describes very well the convergence curve.

Figure 5.3 (`stahl`), *right part.* The estimates of the relative A-norm of the error give a satisfactory information about the convergence also for small values of d ($d = 1,\ d = 4$).

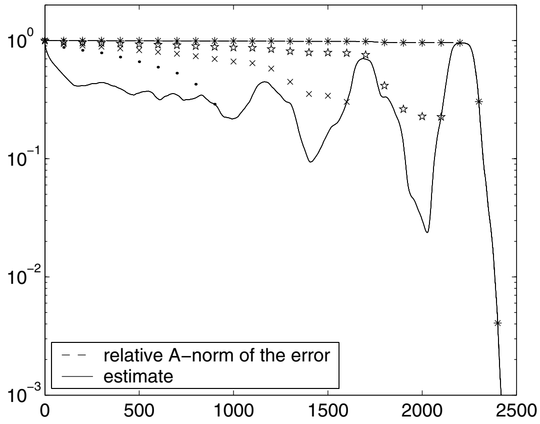


Figure 5.4: The system `s3dkt3m2`. The relative A-norm of the error (dashed line), the estimate of the relative A-norm of the error with $d = 100$ (solid line) and the curves reconstructed at iterations 1000 (dots), 1700 (x-marks), 2200 (pentagrams) and 2500 (stars).

5.2 Reconstruction of the convergence curve.

Up to now we estimated the A-norm of the error at the iteration step j at the price of running d extra steps, and we considered d to be fixed. The simple form of the estimate $\nu_{j,d}$, see (3.6), (3.7) enables at the given iteration step j updating of the estimates of the A-norm of the error at the steps $j-d, j-2d, \ldots$ at a negligible cost. Indeed, assuming, for simplicity of exposition, that j is a multiple of the chosen d ($j \bmod d = 0$), the identity (3.6) gives

$$\|x - x_{j-id}\|_A^2 = \sum_{l=0}^{i} \nu_{j-ld,d} + \|x - x_{j+d}\|_A^2, \quad i = 0, 1, \ldots . \tag{5.1}$$

In this way,

$$\nu_{j-id,(i+1)d}^{1/2} = \left(\sum_{l=0}^{i} \nu_{j-ld,d} \right)^{1/2} \tag{5.2}$$

approximates $\|x - x_{j-id}\|_A$ with the inaccuracy at most $\|x - x_{j+d}\|_A$. In practical computations we can simply store the values of $\nu_{0,d},\ \nu_{d,d},\ \nu_{2d,d}, \ldots,\ \nu_{j-d,d}$, and with the additional d steps update the estimates for the A-norm of the error in the steps $0, d, 2d, \ldots, j-d$ to

$$\nu_{0,j+d}^{1/2},\ \nu_{d,j}^{1/2},\ \nu_{2d,j-d}^{1/2}, \ldots,\ \nu_{j-d,2d}^{1/2}.$$

Dividing by $\nu_{0,j+d}^{1/2}$ we get the corresponding values of the estimates $\varrho_{d,j}^{1/2}$, $\varrho_{2d,j}^{1/2}$, $\dots$, $\varrho_{j-d,2d}^{1/2}$ for the relative A-norm of the error. We illustrate this "reconstruction" of the convergence curve in Figure 5.4, computed for the problem `s3dkt3m2` with $d = 100$, where we plot the relative A-norm of the error (dashed line), its estimate $\varrho_{j,d}^{1/2}$ (solid line) and the updated estimates of the relative A-norm of the error computed for $j = 1000$ (dots), $j = 1700$ (x-marks), $j = 2200$ (pentagrams) and $j = 2500$ (stars). Please notice that when $\|x - x_j\|_A$ almost stagnates, the updated estimates can significantly differ from the original ones represented by the solid line.

We point out that in this paper we deal with evaluation of convergence, and we left heuristics for proper stopping criteria to further investigation. The problem `s3dkt3m2` illustrates that the last question is not trivial. Though, e.g., the computed estimates (even those updated at the iteration $j = 2200$) significantly decrease in the iterations 1800–2000, the actual value of the A-norm of the error still almost stagnates. We emphasize that neither the residual norm nor the normwise backward error reliably indicate the convergence of the A-norm of the error (cf. Figure 5.1, iterations 1800–2000).

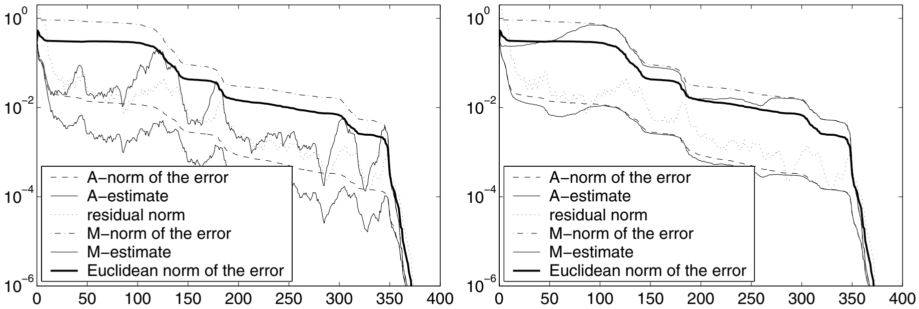


Figure 5.5: The system `tube`. The norms of the errors show similar behaviour. For $d = 4$, the estimates of $\|x - x_j\|_M$ and of $\|x - x_j\|_A$ behave erratically, similarly to the residual norm (left part). For $d = 40$, the estimates are smoother and closer to the approximated curves (right part).

5.3 Comparison of the convergence characteristics.

In Figure 5.5 we plot various convergence characteristics and error estimates for the system `tube`. We have used $d = 4$ (left part) and $d = 40$ (right part). The M-norm of the error $\|x - x_j\|_M$ (dash-dotted line), the Euclidean norm of the error $\|x - x_j\|$ (bold solid line) and the A-norm of the error $\|x - x_j\|_A$ (dashed line) show, except for a few initial iterations, similar behaviour. The estimates both of $\|x - x_j\|_M$ and $\|x - x_j\|_A$ are plotted by the solid lines (no confusion is possible; the line that is always under the dashed curve is the estimate of the A-norm of the error). The A-norm of the error is estimated more accurately than the M-norm of the error; while the estimate $\nu_{j,d}^{1/2}$ differs for no more that one order of magnitude from $\|x - x_j\|_A$, $\tau_{j,d}^{1/2}$ differs often for about two orders of magnitude. The behaviour of both estimates is similar, but the peaks on the line representing $\tau_{j,d}^{1/2}$ are higher than the peaks on the line representing $\nu_{j,d}^{1/2}$. For $d = 4$ both estimates behave erratically, similarly to the residual norm (dotted line). By increasing the value of d, the estimates are smoother and closer to the approximated curves (see right part). The estimate of the M-norm of the error is in our example more sensitive to a slow decrease of error norms.

6 Conclusions.

We propose to incorporate the estimate for the A-norm of the error $\nu_{j,d}^{1/2}$ (see (3.7)) and the estimate for the relative A-norm of the error $\varrho_{j,d}^{1/2}$ (see (3.15)) into software realizations of the PCG method. They are simple and numerically stable, and can complement with a great benefit the quantities commonly used for evaluating convergence. The estimates are tight if the A-norm of the error reasonably decreases. With a good preconditioner ensuring fast convergence we get an authentic information about convergence in terms of the A-norm of the error. Similarly, the estimate $\tau_{j,d}^{1/2}$ (see (3.21)) for the M-norm of the error should be used whenever appropriate.

The proposed estimates can be combined with the standard quantities, such as residual norm or normwise backward error, for constructing a proper stopping criteria. The last topic as well as the (variable) choice of the parameter d in the estimates still needs further work.

Acknowledgment.

The authors wish to thank to R. Kouhia, M. Tůma and P. Benner for their invaluable help, advice and sharing results of their work, which made possible to prepare the experimental part of our paper, and to J. Liesen for his comments which improved the presentation. The comments of M. Arioli were very helpful in revising Section 3.2 of the original manuscript. The authors are also indebted to an anonymous referee for careful reading of the original manuscript and for valuable suggestions which clarified numerous formulations.

REFERENCES

1. M. Arioli, *A stopping criterion for the conjugate gradient algorithms in a finite element method framework*, Numer. Math., 97 (2004), pp. 1–24.
2. M. Arioli, J. W. Demmel and I. S. Duff, *Solving sparse linear systems with sparse backward error*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 165–190.
3. M. Arioli, I. Duff and D. Ruiz, *Stopping criteria for iterative solvers*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 138–144.
4. M. Arioli, E. Noulard and A. Russo, *Stopping criteria for iterative methods: applications to PDE's*, Calcolo, 38 (2001), pp. 97–112.
5. O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, 1994.
6. O. Axelsson and I. Kaporin, *Error norm estimation and stopping criteria in preconditioned conjugate gradient iterations*, Numer. Linear Algebra Appl., 8 (2001), pp. 265–286.

7. I. Babuška, *Mathematics of the verification and validation in computational engineering*, in Mathematical and Computer Modelling in Science and Engineering, M. Kočandrlová and V. Kelar, eds., pp. 5–12, Union of Czech Mathematicians and Physicists, Prague, 2003.

8. R. Barrett, M. Berry, T. F. Chan et al., *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*, SIAM, Philadelphia, PA, 1994.

9. P. Benner, *Solving large-scale control problems*, to appear in IEEE Control Syst. Magazine, 24 (2004), pp. 44–59.

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

11. G. Dahlquist, G. H. Golub and S. G. Nash, *Bounds for the error in linear systems*, in Proc. Workshop on Semi-Infinite Programming, R. Hettich, ed., pp. 154–172, Springer, Berlin, 1978.

12. P. Deuflhard, *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), Contemp. Math., vol. 180, pp. 29–42, Am. Math. Soc., Providence, RI, 1994.

13. V. Frayssé, L. Giraud, S. Gratton and J. Langou, *A set of GMRES routines for real and complex arithmetics on on high performance computers*, TR/PA/03/3, CERFACS, Toulouse Cedex, France, 2003.

14. G. H. Golub and G. Meurant, *Matrices, moments and quadrature*, in Proc. 15-th Dundee Conf., June 1993, D. Sciffeths and G. Watson, eds., pp. 105–156, Longman Sci. Tech. Publ., 1994.

15. G. H. Golub and G. Meurant, *Matrices, moments and quadrature. II. How to compute the norm of the error in iterative methods*, BIT, 37 (1997), pp. 687–705.

16. G. H. Golub and Z. Strakoš, *Estimates in quadratic formulas*, Numer. Algorithms, 8 (1994), pp. 241–268.

17. G. H. Golub and C. van Loan, *Matrix Computation*, The Johns Hopkins University Press, Baltimore MD, third edn., 1996.

18. A. Greenbaum, *Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences*, Linear Algebra Appl., 113 (1989), pp. 7–63.

19. A. Greenbaum, *Estimating the attainable accuracy of recursively computed residual methods*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 535–551.

20. A. Greenbaum, *Iterative methods for solving linear systems*, Frontiers in Applied Mathematics, vol. 17, SIAM, Philadelphia, PA., 1997.

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

22. M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bureau Stand., 49 (1952), pp. 409–435.

23. N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996.

24. R. Kouhia, *Description of the CYLSHELL set*, Laboratory of Structural Mechanics, Finland, May 1998. Matrix Market.

25. Matrix Market, `http://math.nist.gov/MatrixMarket/`. The Matrix Market is a service of the Mathematical and Computational Sciences Division of the Information Technology Laboratory of the National Institute of Standards and Technology.

26. G. Meurant, *Computer solution of large linear systems*, Studies in Mathematics and its Applications, vol. 28, North-Holland Publishing Co., Amsterdam, 1999.

27. G. Meurant, *Numerical experiments in computing bounds for the norm of the error in the preconditioned conjugate gradient algorithm*, Numer. Algorithms 22, 3–4 (1999), pp. 353–365.

28. G. Meurant, *Towards a reliable implementation of the conjugate gradient method*, Invited plenary lecture at the Latsis Symposium: Iterative Solvers for Large Linear Systems, Zurich, February 2002.

29. E. Noulard and M. Arioli, *Vector stopping criteria for iterative methods: Theoretical tools*, pubblicazioni n. 956, Instituto di Analisi Numerica, Pavia, Italy, 1995.

30. W. Oettli and W. Prager, *Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides*, Numer. Math., 6 (1964), pp. 405–409.

31. C. C. Paige, *Error analysis of the lanczos algorithm for tridiagonalizing a symmetric matrix*, J. Inst. Math. Appl, 18 (1976), pp. 341–349.

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

33. C. C. Paige and Z. Strakoš, *Residual and backward error bounds in minimum residual Krylov subspace methods*, SIAM J. Sci. Comput., 23 (2002), pp. 1898–1923 (electronic).

34. Y. Saad, *Iterative Methods for Sparse Linear Systems*, SIAM, Philadelphia, PA, second edn., 2003.

35. R. D. Skeel, *Iterative refinement implies numerical stability for Gaussian elimination*, Math. Comp., 35 (1980), pp. 817–832.

36. G. L. G. Sleijpen, H. A. van der Vorst and D. R. Fokkema, BiCGstab(l) *and other hybrid Bi-CG methods*, Numer. Algorithms, 7 (1994), pp. 75–109.

37. Z. Strakoš, *Theory of Convergence and Effects of Finite Precision Arithmetic in Krylov Subspace Methods*, thesis for the degree doctor of science, Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, February 2001.

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

39. Z. Strakoš and P. Tichý, *On estimation of the A-norm of the error in CG and PCG*, PAMM, 3 (2003), pp. 553–554 (published online).

40. H. A. van der Vorst, *Iterative Krylov methods for large linear systems*, Cambridge Monographs on Applied and Computational Mathematics, vol. 13, Cambridge University Press, Cambridge, 2003.

SIAM J. Sci. Comput.
Vol. 33, No. 2, pp. 565–587

ON EFFICIENT NUMERICAL APPROXIMATION OF THE BILINEAR FORM $c^*A^{-1}b$[†]

ZDENĚK STRAKOŠ[‡] AND PETR TICHÝ[§]

Abstract. Let $A \in \mathbb{C}^{N\times N}$ be a nonsingular complex matrix and b and c be complex vectors of length N. The goal of this paper is to investigate approaches for efficient approximations of the bilinear form $c^*A^{-1}b$. Equivalently, we wish to approximate the scalar value c^*x, where x solves the linear system $Ax = b$. Here the matrix A can be very large or its elements can be too costly to compute so that A is not explicitly available and it is used only in the form of the matrix-vector product. Therefore a direct method is not an option. For A Hermitian positive definite, $b^*A^{-1}b$ can be efficiently approximated as a by-product of the conjugate-gradient iterations, which is mathematically equivalent to the matching moment approximations computed via the Gauss–Christoffel quadrature. In this paper we propose a new method using the biconjugate gradient iterations which is applicable to the general complex case. The proposed approach will be compared with existing ones using analytic arguments and numerical experiments.

Key words. bilinear forms, scattering amplitude, method of moments, Krylov subspace methods, conjugate gradient method, biconjugate gradient method, Lanczos algorithm, Arnoldi algorithm, Gauss–Christoffel quadrature, model reduction

AMS subject classifications. 15A06, 65F10, 65F25, 65G50

DOI. 10.1137/090753723

1. Introduction.

Given a nonsingular square matrix $A \in \mathbb{C}^{N\times N}$ and vectors b and c of compatible dimensions, many applications require approximation of the quantity

$$c^*A^{-1}b\,. \tag{1.1}$$

They arise in signal processing under the name scattering amplitude, as well as in nuclear physics, quantum mechanics, and computational fluid dynamics; see [44, 20] and the references therein. In numerical linear algebra they arise naturally in computing error bounds for iterative methods, in solving inverse problems, in least and total least squares problems, etc.; see [19]. This paper presents an approach for approximating $c^*A^{-1}b$ that is designed to be computationally efficient. For context, we also briefly summarize existing techniques for approximating $c^*A^{-1}b$, notably in the special cases when A, b, and c are real or when A is Hermitian positive definite (HPD).

Given the solution x of the linear algebraic system $Ax = b$, (1.1) can be reformulated as

$$c^*A^{-1}b = c^*x\,.$$

[†] Submitted to the journal's Methods and Algorithms for Scientific Computing section March 24, 2009; accepted for publication (in revised form) November 29, 2010; published electronically March 1, 2011. This work is a part of the research project MSM0021620839 financed by the Ministry of Education of the Czech Republic (Z. Strakoš) and of the Institutional Research Plan AV0Z10300504 (P. Tichý). This work was also partially supported by GAAS grant IAA100300802, project M100300901 of the institutional support of ASCR, and GACR grant 201/09/0917.

http://www.siam.org/journals/sisc/33-2/75372.html

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

[§] Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 18207 Prague, Czech Republic (tichy@cs.cas.cz).

In most applications, $c^*A^{-1}b$ need not be computed to a high accuracy; an approximation correct to very few digits of accuracy is sufficient. Therefore the direct solution of $Ax = b$ is inefficient even for problems of moderate size. If A is sufficiently large or the elements of A are too costly to compute, then the direct solution is not possible. A strategy used by several authors is to generate a sequence $\{x_k\}$ of approximate solutions to $Ax = b$ using a Krylov subspace method and to approximate $c^*A^{-1}b$ by c^*x_n for sufficiently large n. However, even when A is HPD, this approximation may require a large number of iterations as a result of rounding errors affecting x_n; see [52, 53]. A variety of approaches for approximating $c^*A^{-1}b$ have been developed based on quadrature and moments; see, for example, [17]. The extensive literature about connections between moments, iterative methods, and model reduction is too large to summarize here; we mention, as five examples among hundreds, [24, 13, 4, 2, 11]. The same is true for related literature in the area of physical chemistry and solid state physics computations; for reviews of early papers see [23, 40, 45]. The mathematical roots can be found in the work on orthogonal polynomials and continued fractions by Chebyshev [7][1] and Stieltjes [49].[2]

The ideas in this paper for the general complex case (which also includes the real nonsymmetric case) are based on non-Hermitian generalizations of Vorobyev moment problems [55] (to be defined in section 2). Algorithmically, this paper extends the results presented in [52, 53] from the HPD case and the conjugate gradient (CG) method to the general complex case and the biconjugate gradient (BiCG) method.

2. Matching moments in Krylov subspace methods and the Vorobyev moment problem.

To motivate our approach, sections 2.1–2.2 summarize some of the well-known connections between two Krylov subspace methods, model reduction, and moments. In Krylov subspace methods it might be convenient to consider nonzero initial approximations x_0 and y_0 to the solutions of $Ax = b$ and $A^*y = c$, respectively. That is equivalent to applications of the same methods, with the zero initial approximations, to $A\mathbf{x} = \mathbf{b}$ respectively to $A^*\mathbf{y} = \mathbf{c}$, where $\mathbf{b} = b - Ax_0$ respectively $\mathbf{c} = c - A^*y_0$ are the initial residuals and $\mathbf{x} = x - x_0$, $\mathbf{y} = y - y_0$ are unknown. Using

$$c^*A^{-1}b = c^*x_0 + y_0^*\mathbf{b} + \mathbf{c}^*A^{-1}\mathbf{b}\,,$$

$c^*A^{-1}b$ can always be approximated via $\mathbf{c}^*A^{-1}\mathbf{b}$ using zero initial approximations of $\mathbf{x}$, $\mathbf{y}$. Throughout this paper we will therefore consider, with no loss of generality, zero initial approximations.

2.1. Lanczos algorithm as model reduction.

Let $A \in \mathbb{C}^{N\times N}$ be a nonsingular matrix, and let the vectors v_1 and w_1 of length N satisfy $\|v_1\| = 1$, $w_1^*v_1 = 1$. The nth step of the non-Hermitian Lanczos algorithm applied to A with the starting vectors v_1 and w_1 is associated with the following relations:

$$\begin{aligned} AV_n &= V_nT_n + \delta_{n+1}v_{n+1}e_n^T\,, \\ A^*W_n &= W_nT_n^* + \beta_{n+1}^*w_{n+1}e_n^T\,, \end{aligned} \tag{2.1}$$

where $W_n^*V_n = I$, $T_n = W_n^*AV_n$, $\|v_{n+1}\| = 1$, $w_{n+1}^*v_{n+1} = 1$, and the main diagonal, the first subdiagonal, and the first superdiagonal of T_n are given by $\gamma_1, \ldots, \gamma_n$,

[1] This article was reprinted in Oeuvres I, Vol. 11, Chelsea, New York, 1962, pp. 203–230.

[2] This article was reprinted in Oeuvres II (P. Noordhoff, Groningen, 1918), pp. 402–566. The English translation *Investigations on continued fractions* is in Thomas Jan Stieltjes, Collected Papers, Vol. II (Springer-Verlag, Berlin, 1993), pp. 609–745.

$\delta_2, \dots, \delta_n$, and $\beta_2, \dots, \beta_n$, respectively, where $\delta_\ell > 0$, $\beta_\ell \neq 0$, $\ell = 2, \dots, n$; see, e.g., [41, section 7.1]. Here it is assumed that the algorithm does not break down in steps 1 through n. The columns of V_n form a basis of $\mathcal{K}_n(A, v_1)$,

$$\mathcal{K}_n(A, v_1) \equiv \operatorname{span}\{v_1, Av_1, \dots, A^{n-1}v_1\} = \operatorname{span}\{v_1, \dots, v_n\},$$

while the columns of W_n form a basis of $\mathcal{K}_n(A^*, w_1)$. Under the given assumption on existence of steps 1 through n, the non-Hermitian Lanczos algorithm represents the reduction of the *original model* which consists of the matrix A and two vectors v_1 and w_1 to the *reduced model* which consists of the matrix T_n and two identical vectors e_1 and e_1. The reduced model matches the first $2n$ moments:

$$w_1^* A^k v_1 = e_1^T T_n^k e_1, \quad k = 0, 1, \dots, 2n-1. \tag{2.2}$$

Relation (2.2) can be derived from the Vorobyev moment problem, which is to determine a linear operator A_n on $\mathcal{K}_n(A, v_1)$ such that

$$A_n^j v_1 = A^j v_1, \quad j = 1, \dots, n-1, \quad \text{and} \quad A_n^n v_1 = V_n W_n^* A^n v_1. \tag{2.3}$$

Defining A_n as the restriction of A to $\mathcal{K}_n(A, v_1)$ projected orthogonally to $\mathcal{K}_n(A^*, w_1)$ (which represents an oblique projection to $\mathcal{K}_n(A, v_1)$),

$$A_n = V_n W_n^* A V_n W_n^*, \tag{2.4}$$

it follows from the relation $T_n = W_n^* A V_n$ that

$$A_n = V_n T_n W_n^* \tag{2.5}$$

and

$$w_1^* A^k v_1 = w_1^* A_n^k v_1 = e_1^T T_n^k e_1, \quad k = 0, 1, \dots, 2n-1; \tag{2.6}$$

see [51]. The matching moment property (2.2) of the non-Hermitian Lanczos algorithm will be linked with the new numerical approximation of the bilinear form (1.1) proposed in section 3.1.

If A is Hermitian and $w_1 = v_1$, the non-Hermitian Lanczos algorithm reduces to the Hermitian Lanczos algorithm that is associated with the relation

$$AV_n = V_n T_n + \delta_{n+1} v_{n+1} e_n^T,$$

where T_n is the Jacobi matrix, and $V_n^* V_n = I$. In this case, the linear operator A_n is the restriction of A to $\mathcal{K}_n(A, v_1)$ projected orthogonally to $\mathcal{K}_n(A, v_1)$. For more details see [55, Chapter III, sections 2–4], with the summary given in [51].

2.2. Arnoldi algorithm as model reduction. The model reduction represented by the Lanczos algorithm matches the first $2n$ moments (2.2). In the non-Hermitian case, the matrix T_n in (2.2) is determined by oblique projections. This may affect in a negative way conveying information from the original to the reduced model. We therefore need to compare the new numerical approximation proposed in section 3.1 with the model reduction determined by orthogonal projections. This in the non-Hermitian case leads to long recurrences and the Arnoldi algorithm.

Let $A \in \mathbb{C}^{N \times N}$ be a nonsingular matrix, let v_1 and u_1 be vectors of length N, and let $\|v_1\| = \|u_1\| = 1$. The nth step of the Arnoldi algorithm applied to A with v_1 is associated with the relation

$$AV_n = V_n H_n + h_{n+1,n} v_{n+1} e_n^T, \tag{2.7}$$

where $V_n^* V_n = I_n$, $H_n = V_n^* A V_n$, $V_n^* v_{n+1} = 0$, and H_n is the upper Hessenberg matrix with positive entries on the first subdiagonal; see, e.g., [41, section 6.3]. The matching moment property of the Arnoldi algorithm can be expressed in the form

$$u_1^* A^k v_1 = u_1^* V_n H_n^k e_1 = t_n^* H_n^k e_1\,, \quad k = 0, \ldots, n-1\,, \tag{2.8}$$

where $u_1 \equiv V_n t_n + u_1^{\perp} = V_n(V_n^* u_1) + u_1^{\perp}$, and $u_1^{\perp}$ is the component of u_1 orthogonal to $\mathcal{K}_n(A, v_1)$. With $u_1 = v_1$ we can add one more moment. To derive (2.8), we invoke the Vorobyev moment problem linked with the Arnoldi algorithm, which is to determine a linear operator on $\mathcal{K}_n(A, v_1)$ such that

$$A_n^j v_1 = A^j v_1, \quad j = 1, \ldots, n-1, \quad \text{and} \quad A_n^n v_1 = V_n V_n^* A^n v_1. \tag{2.9}$$

Defining A_n as the restriction of A to $\mathcal{K}_n(A, v_1)$ projected orthogonally to $\mathcal{K}_n(A, v_1)$,

$$A_n = V_n V_n^* A V_n V_n^*\,, \tag{2.10}$$

it follows from the relation $H_n = V_n^* A V_n$ that

$$A_n = V_n H_n V_n^* \tag{2.11}$$

and

$$u_1^* A^k v_1 = u_1^* A_n^k v_1 = t_n^* H_n^k e_1\,, \quad k = 0, 1, \ldots, n-1\,. \tag{2.12}$$

Since A is non-Hermitian, the matching moment property cannot in general be extended beyond n moments; see [51].

3. Numerical approximation of the bilinear form $c^* A^{-1} b$.

The relationship of CG to the Gauss–Christoffel quadrature, continued fractions, and moments was pointed out in the founding paper by Hestenes and Stiefel [29, sections 14–18]; see also [55, Chapter III, section 2, pp. 53 and 59] and the summary in [34, pp. 483–484 and p. 493]. In the framework of the Vorobyev moment problem, CG, and the Hermitian Lanczos algorithm, the non-Hermitian Lanczos algorithm and the Arnoldi algorithm look for a reduced order operator A_n (see (2.5) and (2.11)), with the property of matching the maximal number of moments; see (2.6) and (2.12). An approximation of the bilinear form $c^* A^{-1} b$ can then be expressed as

$$c^* A_n^{-1} b\,, \tag{3.1}$$

where A_n^{-1} is the matrix representation of the inverse of the reduced order operator A_n which is restricted onto $\mathcal{K}_n(A, b)$; see, e.g., [30, p. 79]. As an example,

$$A_n^{-1} = V_n T_n^{-1} W_n^* \tag{3.2}$$

holds for the non-Hermitian Lanczos algorithm (see (2.5)). Considering the starting vectors $v_1 = b/\|b\|$ and $w_1 = c\|b\|/c^* b$, we get

$$c^* A_n^{-1} b = \frac{c^* b}{\|b\|} w_1^* V_n T_n^{-1} W_n v_1 \|b\| = (c^* b)\, e_1^T T_n^{-1} e_1\,. \tag{3.3}$$

To our knowledge, the formula $e_1^T T_n^{-1} e_1$ was used for the symmetric positive definite case for the first time by Golub and coworkers [8, 15, 9]; for a survey see, e.g., [19], [34, section 3.3], [16, part V, with the commentary given by Gautschi]. In this section we propose new ways of computing $c^*A_n^{-1}b$ using the BiCG-related methods and relate them to existing approaches.

Our results presented below can be derived without using (3.3) and even without mentioning the Vorobyev moment problem. In order to get an insight into the problem of approximating the bilinear form $c^*A^{-1}b$ (see, e.g., the brief discussion of the Arnoldi algorithm and BiCG in the last section of this paper), this link is, in our opinion, important, similarly as the link with the Gauss–Christoffel quadrature is important for understanding the behavior of the Lanczos algorithm and CG; see, e.g., [29, 25], [21, section 5 on rounding error analysis].

3.1. Approximation based on the BiCG method. The BiCG method [33, 10] (see Algorithm 1) solves simultaneously the primal and dual systems of linear algebraic equations $Ax = b$ and $A^*y = c$; see [50, 20]. BiCG computes sequences of approximations $\{x_n\}$ and $\{y_n\}$ such that $x_n \in \mathcal{K}_n(A, b)$ and $y_n \in \mathcal{K}_n(A^*, c)$, while

$$r_n \equiv b - Ax_n \perp \mathcal{K}_n(A^*, c), \qquad s_n \equiv c - A^*y_n \perp \mathcal{K}_n(A, b)\,. \tag{3.4}$$

Algorithm 1. Biconjugate Gradient (BiCG) Method

```
input $A$, $A^*$, $b$, $c$, $x_0 = 0$, $y_0 = 0$
$r_0 = p_0 = b\,, \quad s_0 = q_0 = c$
for $n = 0, 1, \dots$
    $\alpha_n = \frac{s_n^* r_n}{q_n^* A p_n}$
    $x_{n+1} = x_n + \alpha_n p_n\,, \quad y_{n+1} = y_n + \alpha_n^* q_n$
    $r_{n+1} = r_n - \alpha_n A p_n\,, \quad s_{n+1} = s_n - \alpha_n^* A^* q_n$
    $\eta_{n+1} = \frac{s_{n+1}^* r_{n+1}}{s_n^* r_n}$
    $p_{n+1} = r_{n+1} + \eta_{n+1} p_n\,, \quad q_{n+1} = s_{n+1} + \eta_{n+1}^* q_n$
end
```

Assuming that there is no breakdown in the first n steps, the sequences of approximate solutions in the BiCG method have the form

$$x_n = V_n f_n \quad \text{and} \quad y_n = W_n g_n \tag{3.5}$$

for some vectors f_n and g_n. Relation (3.2), which gives an expression for A_n^{-1}, suggests using $c^*A_n^{-1}b$ as an approximation of $c^*A^{-1}b$; see (3.3). We now show how this approximation computed from the iterates of the non-Hermitian Lanczos algorithm (described in section 2.1), with starting vectors $v_1 = b/\|b\|$, $w_1 = c\|b\|/c^*b$, is related to the BiCG method. In order to derive a formula for $c^*A_n^{-1}b$, we invoke two kinds of global biorthogonality conditions associated with the BiCG method:

$$W_n^* r_n = 0 \quad \text{and} \quad V_n^* s_n = 0, \tag{3.6}$$

$$W_n^* b = \|b\| W_n^* v_1 = \|b\| e_1. \tag{3.7}$$

The conditions (3.6) lead to linear systems $\|b\|e_1 = T_n f_n$ and $(v_1^* c)\, e_1 = T_n^* g_n$ for the unknown coordinates f_n and g_n. Consequently, $x_n = \|b\|\, V_n T_n^{-1} e_1$, $y_n = (v_1^* c)\, W_n (T_n^*)^{-1} e_1$. Then, using the global orthogonality relations (3.7), we have

$$c^* A_n^{-1} b = c^* V_n T_n^{-1} W_n^* b = c^* x_n\,. \tag{3.8}$$

Analogously, the dual quantity is given by

$$b^* (A_n^{-1})^* c = b^* W_n (T_n^*)^{-1} V_n^* c = b^* y_n. \tag{3.9}$$

The last term in (3.8) gives the well-known *scattering amplitude* approximation to c^*x; see [56, 44, 43]. Please note also that from (3.8)

$$c^* A_n^{-1} b = c^* b \, (T_n^{-1})_{1,1} \tag{3.10}$$

(see (3.3)), where the value $(T_n^{-1})_{1,1}$ can be easily computed at a negligible additional cost using the algorithm in [17, p. 135]. It is worth pointing out that evaluation of (3.10) does not require explicit computation of x_n.

The global biorthogonality conditions (3.6) and (3.7) needed for the derivation of (3.8) and (3.9) are in general not satisfied in finite precision computations. Due to rounding errors, computing sufficiently accurate approximations using (3.8) (or (3.9)) may require a large number of iterations that are (as shown below) not necessary. Therefore we present a new mathematically equivalent approximation which will be derived using only *local biorthogonality.* Using the expressions for s_{j+1}, r_{j+1} and p_j in Algorithm 1, we have for $j = 0, \ldots, n-1$

$$\begin{aligned} & s_j^* A^{-1} r_j - s_{j+1}^* A^{-1} r_{j+1} \\ &= (s_{j+1} + \alpha_j^* A^* q_j)^* A^{-1} (r_{j+1} + \alpha_j A p_j) - s_{j+1}^* A^{-1} r_{j+1} \\ &= \alpha_j^2 q_j^* A p_j + \alpha_j s_{j+1}^* p_j + \alpha_j q_j^* r_{j+1} \\ &= \alpha_j s_j^* r_j + \alpha_j (s_{j+1}^* p_j + q_j^* r_{j+1}) \; = \alpha_j s_j^* r_j \, . \end{aligned} \tag{3.11}$$

For the last equality we used the local biorthogonality between the residuals and the search directions of the primal and dual problem

$$s_{j+1}^* p_j = 0 \quad \text{and} \quad q_j^* r_{j+1} = 0 \, . \tag{3.12}$$

Consequently, using

$$c^* A^{-1} b - s_n^* A^{-1} r_n = \sum_{j=0}^{n-1} \left(s_j^* A^{-1} r_j - s_{j+1}^* A^{-1} r_{j+1} \right),$$

we finally obtain

$$c^* A^{-1} b = \sum_{j=0}^{n-1} \alpha_j s_j^* r_j + s_n^* A^{-1} r_n \, . \tag{3.13}$$

Relation (3.13) is significant because it provides an exact expression for $c^* A^{-1} b$, the first term of which is a summation involving the (available) inner product of the BiCG primal and dual residuals. As well, (3.13) generalizes the result from the HPD case, in which $b^* A^{-1} b$ and $r_n A^{-1} r_n$ equal, respectively, the squared A-norms of the errors at steps 0 and n; see [52].

If the primal and dual residuals in the BiCG method become small, the second term $s_n^* A^{-1} r_n$ on the right-hand side of (3.13) will also become small. This suggests approximating $c^* A^{-1} b$ by the following quantity:

$$\xi_n^{\mathrm{B}} \equiv \sum_{j=0}^{n-1} \alpha_j s_j^* r_j, \tag{3.14}$$

where the superscript “B” means “BiCG.” Although, as we show later, ξ_n^{B} is equal to $c^* x_n$ using exact arithmetic, the summation form of ξ_n^{B} in (3.14) is crucial for computational purposes.

Summarizing, $c^*A^{-1}b$ can be approximated using (3.8), (3.10) and by the new ξ_n^{B} defined in (3.14). It remains to prove that these estimates are mathematically (in exact arithmetic) equivalent. A short algebraic manipulation gives

$$\begin{aligned} c^*A^{-1}b - c^*x_n &= c^*A^{-1}r_n \\ &= c^*A^{-1}r_n - y_n^*r_n + y_n^*r_n \\ &= s_n^*A^{-1}r_n + y_n^*r_n \, . \end{aligned} \tag{3.15}$$

Using the global biorthogonality condition (3.6) and $y_n = W_n g_n$ (see (3.5)), we get $y_n^*r_n = 0$ and, consequently,

$$c^*A^{-1}b = c^*x_n + s_n^*A^{-1}r_n \, . \tag{3.16}$$

Comparing (3.13), (3.16), (3.8), and (3.10), we obtain the (exact arithmetic) equivalence

$$\xi_n^{\mathrm{B}} = \sum_{j=0}^{n-1} \alpha_j s_j^* r_j = c^*x_n = c^*b \, (T_n^{-1})_{1,1} \, . \tag{3.17}$$

Although ξ_n^{B} was derived by simple algebraic manipulations without using (3.1), the equivalence (3.17) shows its connection to matching moment model reduction. This connection is, in our opinion, significant for understanding the proposed estimate ξ_n^{B} representing a numerically efficient way of computing (3.1). It is worth pointing out that analogously to the HPD case (see [52]) in finite precision computations (3.17) does not hold, and, as demonstrated below, the individual (mathematically equivalent) approximations can behave very differently.

Saylor and Smolarski [44] introduced formally orthogonal polynomials and complex Gauss quadrature as a tool for approximating the quantity $c^*A^{-1}b$ (for an earlier introduction of the Gauss quadratures associated with the non-Hermitian Lanczos algorithm see, e.g., [12]). The paper [44] presents an approximation to $c^*A^{-1}b$ mathematically equivalent to c^*x_n. Its derivation assumes that the matrix A is diagonalizable (which is restrictive). Moreover, the result is computationally less convenient than the new ξ_n^{B} defined by (3.14). Therefore we will not consider the approximation from [44] in further detail.

Apart from the existence of the BiCG iterations in steps 1 through n, ξ_n^{B} does not require any further assumptions. It can be computed with negligible additional cost from the quantities α_j and $s_j^*r_j$ available during the BiCG run. Please note that in order to compute ξ_n^{B} the approximate solutions x_n and y_n need not be formed.

3.2. Estimating $c^*A^{-1}b$ using hybrid BiCG methods.

Each step of BiCG requires a matrix-vector product with A and a matrix-vector product with A^*. The idea of Sonneveld [48] was to avoid the multiplication with A^*. The resulting conjugate gradient squared (CGS) algorithm uses two multiplications with A per iteration and it computes approximate solutions only to the primal system. In order to smooth out possible oscillations and to obtain faster convergence, Sonneveld's idea was further developed by Van der Vorst, Gutknecht, their coworkers, and other authors to hybrid BiCG methods like BiCG stabilized (BiCGStab) [54]; see also [27, 47], [3, Chapter 5].

Denoting by $\mathbf{r}_n$ the residual corresponding to the approximate solution $\mathbf{x}_n$ computed by a hybrid BiCG method, we get

$$c^*A^{-1}b = c^*x = c^*\mathbf{x}_n + c^*(x - \mathbf{x}_n) = c^*\mathbf{x}_n + c^*A^{-1}\mathbf{r}_n. \tag{3.18}$$

It is natural to ask whether the inner product $c^*\mathbf{x}_n$ provides a better approximation to $c^*A^{-1}b$ than the BiCG-based c^*x_n. To answer this question, we write the residual vector $\mathbf{r}_n$ in the form

$$\mathbf{r}_n = \psi_n(A)\, r_n\,,$$

where r_n is the BiCG residual and ψ_n is a polynomial of degree n such that $\psi_n(0) = 1$, i.e., $\psi_n(z) = 1 + z\varphi_{n-1}(z)$, where φ_{n-1} is a polynomial of degree $n-1$. The choice of ψ_n determines the particular hybrid BiCG method. From

$$b - A\mathbf{x}_n = \mathbf{r}_n = \psi_n(A)r_n = r_n + A\varphi_{n-1}(A)r_n = b - Ax_n + A\varphi_{n-1}(A)r_n$$

we get

$$\mathbf{x}_n = x_n - \varphi_{n-1}(A)r_n.$$

Since $\varphi_{n-1}(A)^*c \in \mathcal{K}_n(A^*, c)$ and $r_n \perp \mathcal{K}_n(A^*, c)$, we finally get

$$c^*\mathbf{x}_n = c^*x_n - (\varphi_{n-1}(A)^*c)^*r_n = c^*x_n\,. \tag{3.19}$$

In other words, although $\mathbf{x}_n$ can be a better (or worse) approximation to x than the BiCG approximation x_n, both provide the mathematically identical approximations to $c^*A^{-1}b$.

The BiCG coefficients α_j are available in hybrid BiCG methods. The BiCG residuals r_j and s_j are not available, but the inner products $s_j^*r_j$ can be computed as $s_j^*r_j = s_0^*\tilde{\psi}_j(A)r_j \equiv \tau_j$, provided that the leading coefficients in $\tilde{\psi}_j$ and in the polynomial defining s_j are equal. Then

$$\xi_n^{\mathrm{B}} = \sum_{j=0}^{n-1} \alpha_j \tau_j\,. \tag{3.20}$$

Alternatively, we can compute τ_j using the explicitly available coefficients η_j as

$$\tau_0 \equiv c^*b, \quad \tau_j \;\equiv\; \eta_j\tau_{j-1} = \prod_{k=0}^{j-1} \frac{s_{k+1}^*r_{k+1}}{s_k^*r_k} = s_j^*r_j\,, \qquad j = 1, \dots, n-1\,. \tag{3.21}$$

Although ξ_n^{B} computed via (3.20) using hybrid BiCG methods is mathematically the same as ξ_n^{B} computed via (3.14) using BiCG, results of their numerical evaluation may differ substantially; see section 7.

3.3. Estimating $c^*A^{-1}b$ via the Arnoldi algorithm. As with the non-Hermitian Lanczos algorithm and the related BiCG, estimating $c^*A^{-1}b$ via the Arnoldi algorithm uses (3.1), where A_n arises from the associated Vorobyev moment problem; see section 2.2. Taking $u_1 = c$ and $v_1 = b/\|b\|$ and using (2.11), the approximation (3.1) is in the Arnoldi algorithm given by

$$c^*A_n^{-1}b = \|b\|\, t_n^*H_n^{-1}e_1\,,$$

where $t_n \equiv V_n^*c$. We therefore denote

$$\xi_n^{\mathrm{A}} \equiv \|b\|\, t_n^*H_n^{-1}e_1\,, \tag{3.22}$$

where the superscript "A" means "Arnoldi." Note that the same formula can be obtained using the quadrature rules in [5, pp. 776–777].

The significance of ξ_n^{A} (in comparison with ξ_n^{B}) is in the fact that A_n associated with the Arnoldi algorithm is based on orthogonal projections; see section 2.2. Moreover, although the Arnoldi algorithm matches fewer moments than the non-Hermitian Lanczos algorithm, it is worth noting that H_n in (3.22) contains $n(n+1)/2+n-1$ generally nonzero elements, while T_n in (3.17) contains only $3n-2$ generally nonzero elements. The upper Hessenberg matrix H_n may contain more information about the original model represented by A, b, and c than the tridiagonal matrix T_n. Since

$$x_n = A_n^{-1}b = \|b\|\, V_n H_n^{-1} e_1 \tag{3.23}$$

represents the approximate solution of $Ax = b$ in the full orthogonalization method (FOM) (see [41, pp. 159–160]), we can write

$$\xi_n^{\mathrm{A}} = c^* x_n\,, \tag{3.24}$$

where x_n is computed by FOM.

In the HPD case and CG the approximate solution x_n is computed using short recurrences. In finite precision arithmetic computations, short recurrences typically lead to a fast loss of orthogonality due to rounding errors and, consequently, to *delay of convergence*. Similar behavior can be expected with non-Hermitian Lanczos, BiCG, and hybrid BiCG methods due to loss of biorthogonality. Since the Arnoldi algorithm uses long recurrences, the orthogonality among the computed basis vectors is lost in finite precision arithmetic computations only gradually (details of rounding error analysis can be found in [36] and in the earlier literature referenced therein). Therefore, unlike in BiCG or in the hybrid BiCG methods (see (3.8) and (3.19)), in FOM the formula (3.24) can be used in practical computations without delay of convergence due to rounding errors.

4. Transformation to the Hermitian positive definite case.

Numerical approximations of the bilinear form $c^*A^{-1}b$ presented in section 3 used non-Hermitian Krylov subspace methods applied to the nonsingular complex matrix A. Here we write the bilinear form as

$$c^*A^{-1}b = c^*A^*(AA^*)^{-1}b = c^*(A^*A)^{-1}A^*b\,, \tag{4.1}$$

which suggests deriving its approximation by defining $\tilde{c} = Ac$ and approximating $\tilde{c}^*(AA^*)^{-1}b$. A second possibility is to approximate $c^*(A^*A)^{-1}\tilde{b}$, where $\tilde{b} = A^*b$. In either case, the problem of interest is to approximate $u^*B^{-1}v$, where B is HPD; see also [17, section 3.2]. For simplicity we consider only the second choice.

4.1. Using the polarization identity.

If B is real, symmetric, and positive definite, it was suggested in [17, pp. 16 and p. 33] and [21, p. 242] that a polarization identity can be used to approximate $u^*B^{-1}v$, where $u \neq v$. On a complex Hilbert space with the inner product $\langle\cdot,\cdot\rangle$, conjugate linear in the second variable, the polarization identity takes the form (see, e.g., [32], [57, p. 23])

$$\begin{aligned} 2\langle v,u\rangle &= (\|v+u\|^2 - \|v-u\|^2 + \mathbf{i}\|v+\mathbf{i}u\|^2 - \mathbf{i}\|v-\mathbf{i}u\|^2)/2 \\ &= \|v+u\|^2 - (1+\mathbf{i})(\|v\|^2 + \|u\|^2) + \mathbf{i}\|v+\mathbf{i}u\|^2\,. \end{aligned} \tag{4.2}$$

Defining $\langle v,u\rangle \equiv u^*B^{-1}v$, the term $\|u\|^2$ in (4.2) is given by $u^*B^{-1}u$. With $v = A^*b$ and $B = A^*A$, it follows that $v^*B^{-1}v = b^*b$. The remaining three terms that need to be approximated are

$$(v+u)^*B^{-1}(v+u)\,, \quad (v+\mathbf{i}u)^*B^{-1}(v+\mathbf{i}u), \quad \text{and} \quad u^*B^{-1}u, \tag{4.3}$$

all of which have the form $w^*B^{-1}w$ with the HPD matrix B. Then BiCG reduces to the standard CG, with (3.13) giving

$$w^*B^{-1}w = \xi_n^{\mathrm{CG}} + r_n^*B^{-1}r_n\,, \qquad \xi_n^{\mathrm{CG}} \equiv \sum_{j=0}^{n-1} \alpha_j \|r_j\|^2\,; \tag{4.4}$$

see [52, relation (3.8)]. Since $B = A^*A$, the quantities α_j, $\|r_j\|^2$ and thus ξ_n^{CG} can conveniently be computed without forming the matrix B using the algorithms CGNR; see [29, section 10], where "NR" comes from normal equation residual [22, section 10.4]. As an alternative one can consider the HPD analogy of (3.10) with $b = c = w$ and T_n resulting from the n steps of the Hermitian Lanczos algorithm applied to the matrix $B = A^*A$ with the starting vector b. This gives

$$w^*B^{-1}w = \|b\|^*(T_n^{-1})_{1,1} + r_n^*B^{-1}r_n, \tag{4.5}$$

where r_n is as in (4.4). Numerically this can be efficiently computed via the algorithm LSQR proposed by Paige and Saunders [38, 37] which uses the Golub–Kahan bidiagonalization [14] and computes the Cholesky factor of T_n.

The approximation error $r_n^*B^{-1}r_n$ in (4.4)–(4.5) is equal to the squared energy norm of the error in CG, and therefore it is monotonically decreasing with n. This represents a significant difference in comparison with (3.13), where the error term $s_n^*A^{-1}r_n$ typically oscillates. There are methods for computing the upper and lower bounds for $w^*B^{-1}w$; see [17, 21, 18, 6]. Consequently, using (4.2), one can compute (assuming exact arithmetic) upper and lower bounds for the real and imaginary parts of the bilinear form $c^*A^{-1}b$. Moreover, (4.4) holds, up to a small error, also for quantities computed in finite precision arithmetic; see [52]. (It is worth pointing out that ξ_n^{CG} computed in finite precision arithmetic can be much larger than its exact arithmetic counterpart computed at the same step.) The price of transforming the non-Hermitian problem to the Hermitian one using the polarization identity (4.2) is, however, substantial. Approximation of three terms (4.3) requires three CG computations with the *same* matrix B and *different* initial vectors, with a total of *six* matrix-vector multiplications (three with A and three with A^*) per one iteration step. In our experiments, the approach using the polarization identity (4.2) was not competitive with ξ_n^{B}.

4.2. Using the normal equations. Another way to approximate the bilinear form $c^*A^{-1}b$ is to apply CGNR to $A^*Ax = A^*b$. The bilinear form can then be approximated by c^*x_n, where x_n is the nth iterate of CGNR. Unlike in section 4.1, here only two matrix-vector products (one with A and one with A^*) are needed at each iteration. As with (3.8) in section 3.1, in finite precision arithmetic computing a sufficiently accurate approximation using c^*x_n may be delayed due to loss of orthogonality caused by rounding errors.

Rewriting the bilinear form using $c^*(A^*A)^{-1}A^*b$ as in (4.1), one can also consider BiCG applied to $B = A^*A$ with *two different* initial vectors $u = A^*b$ and $v = c$; for an analogous approach using the non-Hermitian Lanczos algorithm see [17, sections 3.2 and 4.2]. BiCG applied to a system with the matrix B and two different initial vectors needs *four* matrix-vector multiplications (two with A and two with A^*) per iteration.

4.3. The GLSQR approach. Saunders, Simon, and Yip suggested in [42] the so-called generalized LSQR (GLSQR) method which is applied to a matrix and two starting vectors. It can be seen as the block-Lanczos algorithm applied to the matrix A^*A with the starting block $[c, A^*b]$; see also [39]. The GLSQR method solves simultaneously the primal and dual systems (similarly to BiCG in Algorithm 1). The nth step is associated with the following relations:

$$\begin{aligned} AV_n &= U_nT_n + \zeta_{n+1}u_{n+1}e_n^T\,, \\ A^*U_n &= V_nT_n^* + \theta_{n+1}v_{n+1}e_n^T\,, \end{aligned}$$

where $u_1 = b/\|b\|$, $v_1 = c/\|c\|$; $V_n = [v_1, \ldots, v_n]$ and $U_n = [u_1, \ldots, u_n]$ are orthonormal matrices, $V_n^*v_{n+1} = 0$, $U_n^*u_{n+1} = 0$, T_n is tridiagonal, and ζ_{n+1} and θ_{n+1} are the normalization coefficients. Using GLSQR and applying the block Gauss quadrature rule from [17, sections 3.3 and 4.3], Golub, Stoll, and Wathen derived the following approximation to $c^*A^{-1}b$:

$$\xi_n^{\mathrm{G}} = \|b\|\,\|c\|\; e_1^TT_n^{-1}e_1, \tag{4.6}$$

where the superscript "G" means GLSQR; see [20, section 3.3]. The GLSQR approach requires *two* matrix-vector multiplications (one by A and one by A^*) per iteration.

5. Preconditioning. Let $P_{\scriptscriptstyle L}$ and $P_{\scriptscriptstyle R}$ be nonsingular matrices such that the systems of linear algebraic equations with the matrices $P_{\scriptscriptstyle L}$ and $P_{\scriptscriptstyle R}$ are easily solvable. Clearly

$$c^*A^{-1}b = (P_{\scriptscriptstyle R}^{-*}c)^*(P_{\scriptscriptstyle L}^{-1}AP_{\scriptscriptstyle R}^{-1})^{-1}(P_{\scriptscriptstyle L}^{-1}b) = \mathbf{c}^*\mathbf{A}^{-1}\mathbf{b},$$

where $\mathbf{A} \equiv P_{\scriptscriptstyle L}^{-1}AP_{\scriptscriptstyle R}^{-1}$, $\mathbf{c} \equiv P_{\scriptscriptstyle R}^{-*}c$, and $\mathbf{b} \equiv P_{\scriptscriptstyle L}^{-1}b$. The approximation techniques described above can be applied to the preconditioned problem $\mathbf{c}^*\mathbf{A}^{-1}\mathbf{b}$. Preconditioning should lead to faster convergence. As a side effect, fast convergence can help prevent significant delays due to rounding errors; see the illustrations in section 7. It is obvious that $\mathbf{A}^{-1}$ need not be formed explicitly.

6. Comments on numerical stability issues. A thorough numerical stability analysis of the approaches for approximating the bilinear form $c^*A^{-1}b$ which are presented in this paper is yet to be done. Here we concentrate on supporting arguments for the claim that the new estimate ξ_n^{B} (see (3.14)) should be preferred to the mathematically equivalent (and commonly used) scattering amplitude estimate c^*x_n; see (3.8).

Using $A^{-1}r_n = x - x_n$, we rewrite for clarity of exposition the formulas which express the errors of the computed approximation (see (3.13)–(3.16)):

$$c^*A^{-1}b = \xi_n^{\mathrm{B}} + s_n^*(x - x_n)\,, \qquad \xi_n^{\mathrm{B}} = \sum_{j=0}^{n-1} \alpha_j s_j^* r_j\,, \tag{6.1}$$

$$c^*A^{-1}b = c^*x_n + c^*(x - x_n) \tag{6.2}$$

$$= c^*x_n + s_n^*(x - x_n) + y_n^*(b - Ax_n)\,. \tag{6.3}$$

Mathematically (in exact arithmetic),

$$y_n^*(b - Ax_n) = y_n^*r_n = g_n^*W_nr_n = 0 \tag{6.4}$$

due to the global biorthogonality condition (3.6). Therefore

$$s_n^*(x - x_n) = c^*(x - x_n)\,. \tag{6.5}$$

In computations using finite precision arithmetic the global biorthogonality (3.6) is in general lost, and, subsequently, (6.5) does not hold. Let the quantities computed using finite precision arithmetic be denoted by " ˆ ". Supposing that the BiCG residual $\hat{s}_n$ for the dual problem $A^*y = c$ is small, we may expect

$$|\hat{s}_n^*(x - \hat{x}_n)| \ll |c^*(x - \hat{x}_n)|\,. \tag{6.6}$$

This corresponds to $\hat{\xi}_n^{\mathrm{B}}$ much closer to $c^*A^{-1}b$ than $c^*\hat{x}_n$. In other words, in finite precision arithmetic computations, the term $\hat{y}_n^*(b-A\hat{x}_n)$ as well as a possible difference between the true and iteratively computed residuals must be taken into account (for the symmetric positive definite analogy see [52, section 6]). Provided that the finite precision analogies of (6.1) and (6.3) hold up to a small inaccuracy, the term $\hat{y}_n^*(b - A\hat{x}_n)$ would explain the numerical behavior of the estimate $c^*\hat{x}_n$.

Analogously to (3.15) one can easily derive for the computed approximations $\hat{x}_n$ and $\hat{y}_n$

$$c^*A^{-1}b = c^*\hat{x}_n + (c - A^*\hat{y}_n)^*(x - \hat{x}_n) + \hat{y}_n^*(b - A\hat{x}_n).$$

Therefore (6.3) holds, up to small inaccuracy, also for results of finite precision computations until the true residual $b - A\hat{y}_n$ does not differ significantly from the iteratively computed residual $\hat{s}_n$. For more details on the difference between the true and the iteratively computed residuals see the analysis in [47, 26].

Concerning the finite precision analogy of (6.1), the situation is much more complicated. Consider first $A \in \mathbb{R}^{N\times N}$ symmetric positive definite and $c = b \in \mathbb{R}^N$. Then BiCG reduces to CG, $r_n = s_n$, and (6.1) can be rewritten as

$$b^TA^{-1}b = \xi_n^{\mathrm{CG}} + r_n^TA^{-1}r_n\,, \qquad \xi_n^{\mathrm{CG}} = \sum_{j=0}^{n-1} \alpha_j\,\|r_j\|^2\,, \tag{6.7}$$

or, considering that $r_n^TA^{-1}r_n = (x - x_n)^TA(x - x_n)$, $b^TA^{-1}b = x^TAx$,

$$\|x\|_A^2 = \xi_n^{\mathrm{CG}} + \|x - x_n\|_A^2\,, \tag{6.8}$$

where the A-norm of a vector z is defined by $\|z\|_A \equiv (z^*Az)^{1/2}$. It was proved in [52] that (6.8) holds also for the results of finite precision arithmetic computations up to a term proportional to $\varepsilon\|x\|_A\|x - \hat{x}_n\|_A$; here ε denotes machine precision unit (we omit some tedious details). Consequently, until $\|x - \hat{x}_n\|_A = (\hat{r}_n^TA^{-1}\hat{r}_n)^{1/2}$ becomes close to $\varepsilon\|x\|_A$, the computed $\hat{\xi}_n^{CG}$ approximates $b^TA^{-1}b = \|x\|_A^2$ with the error of the approximation being close to $\hat{r}_n^TA^{-1}\hat{r}_n = \|x - \hat{x}_n\|_A^2$; see [52, Theorem 10.1]. This result is proved in several steps with two main ingredients. First, it is proved that the iteratively computed residual $\hat{r}_j$ (see Algorithm 1 with $A = A^*$ and $s_0 = r_0 = b$) is sufficiently close to the residual $b - A\hat{x}_n$ computed directly from the approximate solution $\hat{x}_n$. Second, it is proved that the local orthogonality between the residuals and the search vectors $\hat{p}_j^T\hat{r}_{j+1}$ is preserved proportionally to machine precision ε; see [52, section 9].

For BiCG one can hardly expect results of the same strength. In particular, a close preservation of the local biorthogonality conditions (3.12) cannot be proved due to the possible occurrence of the so-called breakdowns, when $\hat{q}_j A\hat{p}_j$ or $\hat{s}_j^*\hat{r}_j$ becomes zero. Note that the breakdowns are not caused by rounding errors; they can occur in exact arithmetic.

Using the technique from [52, 53], one can express the inner product $\hat{q}_j^*\hat{r}_{j+1}$ of the quantities computed in finite precision arithmetic using Algorithm 1 as

$$\hat{q}_j^*\hat{r}_{j+1} = \frac{\hat{s}_j^*\hat{r}_j}{\hat{s}_{j-1}^*\hat{r}_{j-1}}\,\hat{q}_{j-1}^*\hat{r}_j + \varepsilon\,\vartheta_j\,,$$

and the size of ϑ_j can be bounded by the norms of the computed vectors, the norm of A, and the size of the coefficient $\hat{\alpha}_j$. By induction we obtain, after some algebraic manipulations (cf. [52, p. 74] or [53, p. 805]),

$$\hat{q}_n^*\hat{r}_{n+1} = \varepsilon\,\hat{s}_n^*\hat{r}_n \sum_{j=0}^{n} \frac{\vartheta_j}{\hat{s}_j^*\hat{r}_j} + \mathcal{O}(\varepsilon^2)\,. \tag{6.9}$$

Now we can clarify the differences between the CG case and the BiCG case.

In the CG case, $\hat{s}_j = \hat{r}_j$ and $\hat{s}_j^*\hat{r}_j = \|\hat{r}_j\|^2$. As shown in [52], the size of ϑ_j is bounded by $\kappa(A)\,\|\hat{r}_j\|^2$. In summary, the local biorthogonality is bounded by a multiple of $\varepsilon\,\|\hat{r}_j\|^2\,\kappa(A)$; see [52, relations (9.14) and (9.15)]. In the BiCG case, $\hat{q}_j^*A\hat{p}_j$ and $\hat{s}_j^*\hat{r}_j$ can become zero due to breakdowns. In practice, the exact breakdowns are very rare, but near breakdowns can cause the corresponding terms in the sum (6.9) to be large. If near breakdowns appear in BiCG, then preserving the local biorthogonality condition (3.12) up to a small inaccuracy cannot be guaranteed in finite precision arithmetic computations. Therefore we were not able to prove that (6.1) holds, up to a small inaccuracy, also in finite precision arithmetic computations. Nevertheless, for ξ_n^{B}, there is no need of preserving the *global orthogonality* conditions (3.6)–(3.7), and, in particular, of $y_n^*r_n = 0$, as in the scattering amplitude approximations. This represents a strong numerical argument in favor of the proposed estimate ξ_n^{B}.

7. Application and numerical experiments.

We will illustrate the behavior of various approaches for approximation of the bilinear form $c^*A^{-1}b$ in several examples of different origins. In this section we omit for simplicity the " ˆ " notation for the computed quantities.

7.1. Test problems.

This paper was practically motivated by the problem of diffraction of light on periodic structures and the RCWA method for its solution; see the monograph [35] and the references therein. Application of the RCWA method can lead to the system of linear algebraic equations, which for the simplest standard two-dimensional model problem has the form (see [28, section 3.5])

$$A\,x \equiv \begin{bmatrix} -I & I & e^{\mathbf{i}\sqrt{C}\varrho} & 0 \\ Y_I & \sqrt{C} & -\sqrt{C}e^{\mathbf{i}\sqrt{C}\varrho} & 0 \\ 0 & e^{\mathbf{i}\sqrt{C}\varrho} & I & -I \\ 0 & \sqrt{C}e^{\mathbf{i}\sqrt{C}\varrho} & -\sqrt{C} & -Y_{\mathrm{II}} \end{bmatrix} x = b\,, \tag{7.1}$$

where Y_{I}, Y_{II} are $(2M+1)\times(2M+1)$ complex diagonal matrices, C is a $(2M+1)\times(2M+1)$ complex Toeplitz plus diagonal matrix, ϱ is a given real and positive parameter, and M is the discretization parameter representing the number of Fourier modes used for approximation of the electric and magnetic fields as well as the material properties. The block structure of (7.1) corresponds to the geometric structure of the physical problem with one slab, where the individual block rows represent the boundary conditions for the electric and magnetic fields on the interface between the slab and the superstrate and the slab and the substrate. For the geometric structure with S slabs the overall number of interfaces is $S+1$, which gives $2(S+1)$ block equations (for (7.1), $2(1+1)=4$). In three-dimensional problems the size of the individual blocks is proportional to the square of the number of Fourier modes.

In real RCWA applications the blocks of the matrix A cannot be formed by evaluating the matrix functions. Considering time constraints given by technological restrictions, that would be too slow. Moreover, one does not need the whole solution of the linear algebraic system. For (7.1) one typically needs only the dominant $(M+1)$st component (here e_{M+1} denotes the vector of the compatible dimension with the $(M+1)$st element equal to one and all other elements equal to zero):

$$e_{M+1}^* A^{-1} b\,; \tag{7.2}$$

see [28, section 3.5, relation (3.45)]. Therefore the problem seems to be well suited for an iterative approximation of the bilinear form (1.1) with $c = e_{M+1}$. In our experiments we use $M = 20$, $S = 1$ and $M = 20$, $S = 20$, leading to the resulting RCWA-motivated matrices:

- `TE2001` (RCWA, 20 Fourier modes and 1 slab): the matrix $A \in \mathbb{C}^{164\times 164}$ is complex nonsymmetric, $\kappa(A) \approx 112$, starting vectors b and c arise from the problem formulation;
- `TE2020` (RCWA, 20 Fourier modes and 20 slabs): the matrix $A \in \mathbb{C}^{1722\times 1722}$ is complex nonsymmetric, $\kappa(A) \approx 2.9e+03$, starting vectors b and c arise from the problem formulation.

In addition, we use in our illustrations four publicly available matrices from different sources:

- `young1c` (ACOUST, HB Collection): the matrix $A \in \mathbb{C}^{841\times 841}$ is complex symmetric, $\kappa(A) \approx 415$;
- `orsirr1` (OILGEN, HB Collection): the matrix $A \in \mathbb{R}^{1030\times 1030}$ is real nonsymmetric, $\kappa(A) \approx 7.7e+04$;
- `pde2961` (MATPDE, NEP Collection): the matrix $A \in \mathbb{R}^{2961\times 2961}$ is real nonsymmetric, $\kappa(A) \approx 642.5$;
- `af23560` (AIRFOI, NEP Collection): the matrix $A \in \mathbb{R}^{23560\times 23560}$ is real nonsymmetric, the condition number estimate computed via the MATLAB command `condest`(A) gives $\kappa(A) \approx 3.5e+05$.

Except for `TE2001` and `TE2020` we choose b and c normalized random vectors.

7.2. An overview of compared methods and their implementations. In this paper we presented three approaches for approximating the bilinear form $c^*A^{-1}b$: the non-Hermitian Lanczos approach, the Arnoldi approach, and the approach based on transformation to the HPD case. In our numerical experiments we use the standard versions of BiCG [10], CGS [48], BiCGStab(4) [47], modified Gram–Schmidt Arnoldi [41], and GLSQR [20]. For illustration of the behavior of BiCG in exact precision arithmetic we run in some experiments BiCGreo with the rebiorthogonalized basis vectors at each step (at step n, r_n is reorthogonalized against the previously computed $s_0, s_1, \ldots, s_{n-1}$, and s_n is reorthogonalized against the previously computed $r_0, r_1, \ldots, r_{n-1}$). We use a special version of the BiCGStab [54] algorithm with the technique suggested in [46] (we choose the free parameter $\Omega = 0.7$) to improve the accuracy of the computed BiCG coefficients. We compare the approximations ξ_n^{B} (see (3.14) and (3.20)) and c^*x_n computed via BiCG, BiCGreo, and the hybrid BiCG

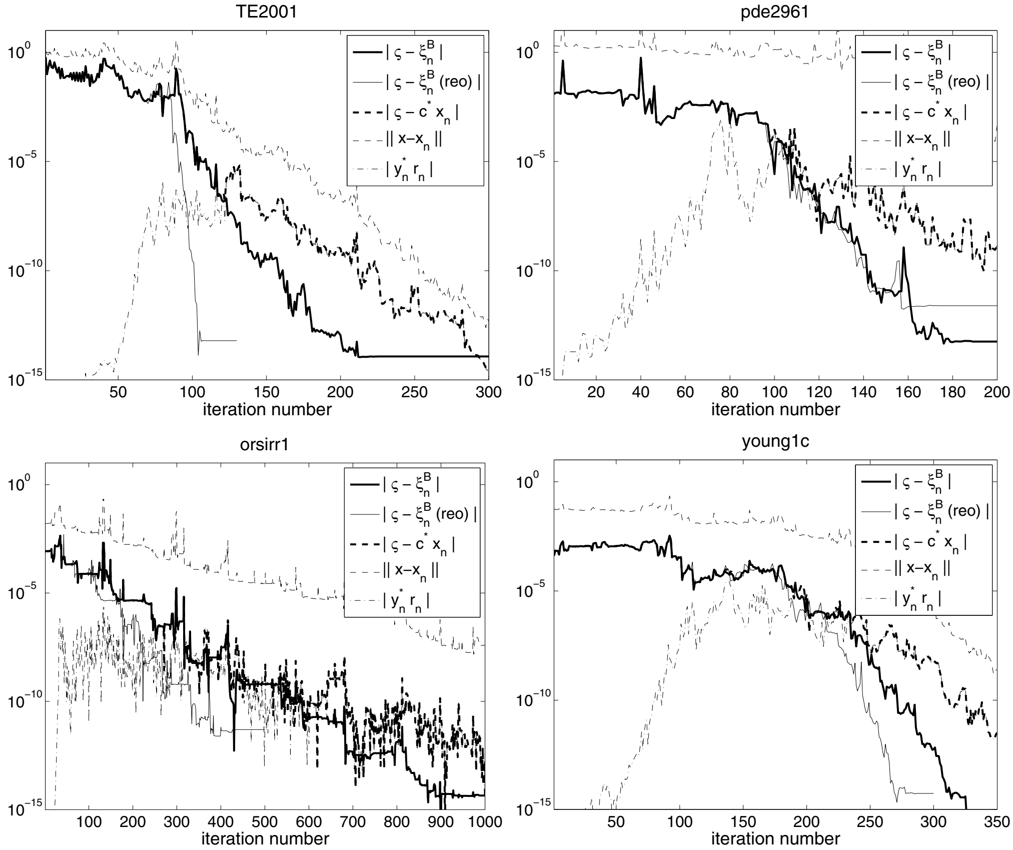


FIG. 7.1. *Comparison of the errors* $|\varsigma - \xi^{\mathrm{B}}_n|$ *(bold solid line) and* $|\varsigma - c^*x_n|$ *(bold dashed line) for the mathematically equivalent approximations computed via BiCG. Both approximations are close to each other until the size of* $|y^*_n r_n|$ *(dash-dotted line) is negligible in comparison to the size of* $|c^*x_n|$*. To simulate the behavior of* ξ^{B}_n *in exact arithmetic, we also plot* $|\varsigma - \xi^{\mathrm{B}}_n \text{ (reo)}|$ *with* ξ^{B}_n (reo) *computed via BiCGreo (solid line).*

methods, ξ^{A}_n (see (3.22)) computed via the Arnoldi algorithm, and ξ^{G}_n (see (4.6)) computed via GLSQR. We do not include in our experiments the approximation (3.10) computed via the non-Hermitian Lanczos algorithm. It gives results very similar to those of ξ^{B}_n computed via BiCG. We also do not present results for the approximations introduced in section 4.1. In our set of problems they do not seem to be competitive with other approximations; see the comment in section 7.5.

Denote for simplicity of further presentation

$$\varsigma(A, b, c) \equiv \varsigma = c^*A^{-1}b\,.$$

The value ς used for determining the approximation error in all subsequent experiments was computed using the MATLAB command `c'(A\b)`.

7.3. Comparison of the approximations ξ^{B}_n and c^*x_n. In Figure 7.1 we compare the error $|\varsigma - \xi^{\mathrm{B}}_n|$ of the new approximation ξ^{B}_n (see (3.14)) (bold solid line) with the error $|\varsigma - c^*x_n|$ of the scattering amplitude approximation c^*x_n (see (3.8)), where x_n is computed by Algorithm 1 (dashed line). In order to illustrate the effects of rounding errors to the BiCG algorithm we also plot $|\varsigma - \xi^{\mathrm{B}}_n \text{ (reo)}|$ for ξ^{B}_n (reo) computed via BiCGreo. The comparison is complemented by the upper bound $\|x - x_n\| \geq$

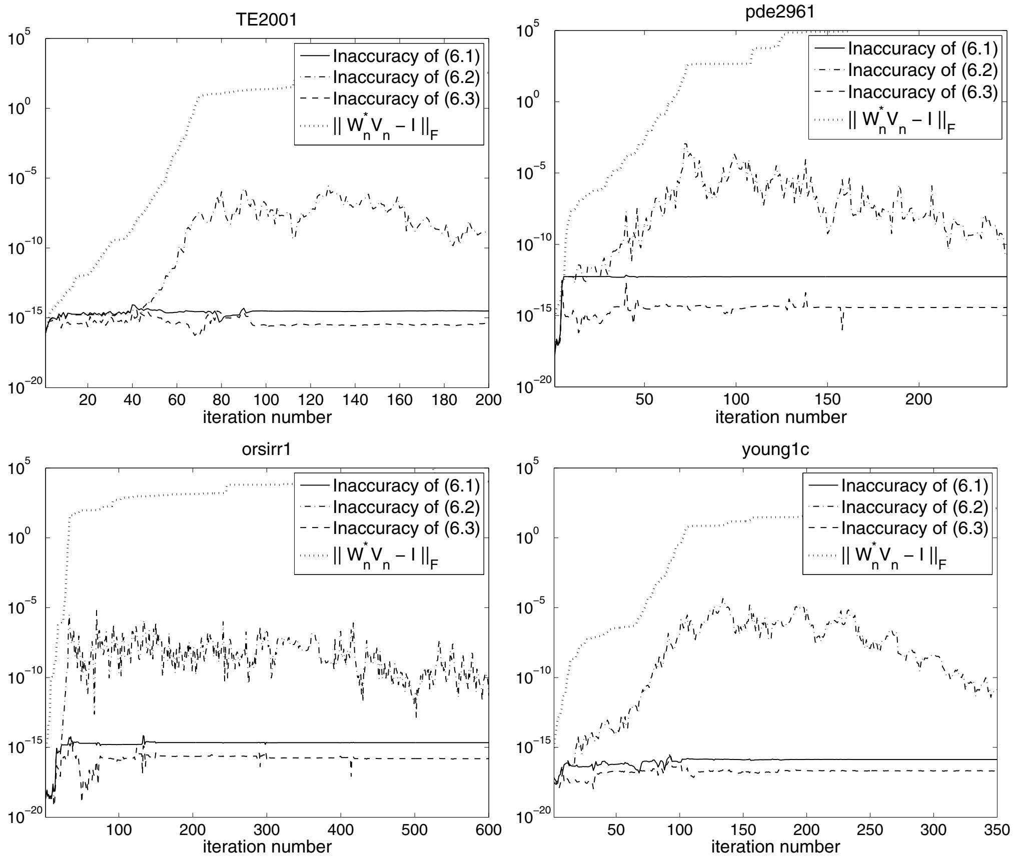


FIG. 7.2. *Inaccuracy in* (6.1)*,* (6.2)*, and* (6.3) *for the quantities computed in finite precision arithmetic. For each equation we plot the absolute value of the difference of the terms on the left- and right-hand sides.*

$|c^*(x - x_n)|$ (here $\|c\| = 1$) and by the value $|y_n^* r_n|$ (dash-dotted line) which in finite precision arithmetic computations determines the difference between ξ_n^{B} and $c^* x_n$; see (6.1) and (6.3). The dashed line coincides in all figures with the bold solid line until the bold solid line is crossed by the dash-dotted line. It is interesting that for the matrix `pde2961` the approximations ξ_n^{B} (reo) and ξ_n^{B} almost coincide except for the fact that ξ_n^{B} (reo) exhibits larger maximal attainable accuracy (that can be attributed to additional accumulation of roundoff due to rebiorthogonalization). All our experiments confirm that the newly proposed approximation ξ_n^{B} should be preferred to computation of the scattering amplitude $c^* x_n$.

Figure 7.2 shows the inaccuracy of (6.1), (6.2), and (6.3) for the quantities computed in finite precision arithmetic as well as the loss of global biorthogonality in BiCG. While (6.1) and (6.3) are for all experiments using the matrices `TE2001`, `pde2961`, `orsirr1`, and `young1c` satisfied up to the inaccuracy remarkably close to machine precision, (6.2) is considerably violated due to the loss of biorthogonality.

7.4. BiCG and hybrid BiCG methods in approximation of $c^* A^{-1} b$. As explained in section 3.2, ξ_n^{B} can be computed using hybrid BiCG methods. It is, however, well known that computing the BiCG coefficients accurately may represent a problem in hybrid BiCG methods. As stated in [46, p. 220], "In order to maintain the convergence properties of the BiCG component in hybrid BiCG methods, it is

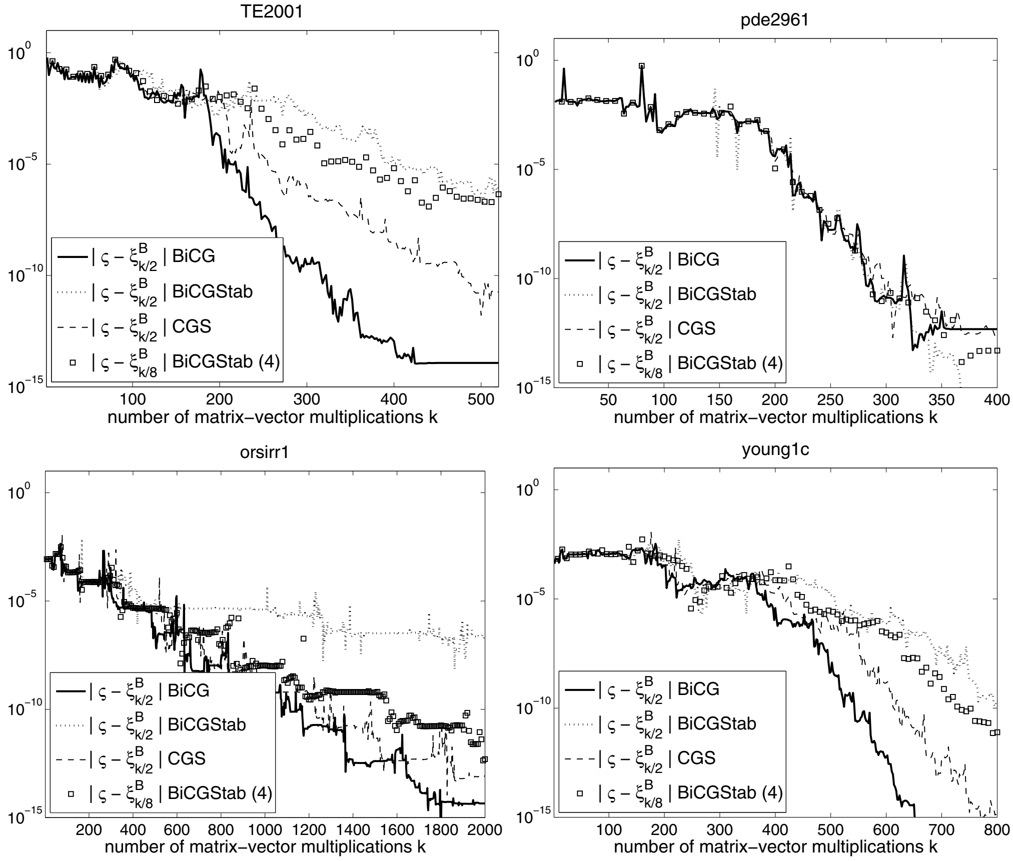


FIG. 7.3. *Comparison of errors* $|\varsigma - \xi^{\mathrm{B}}_{k/2}|$ *for the approximation* $\xi^{\mathrm{B}}_{k/2}$ *computed via BiCG (bold solid line), BiCGStab (dashed line), and CGS (dotted line) and the error* $|\varsigma - \xi^{\mathrm{B}}_{k/8}|$ *of the approximation* $\xi^{\mathrm{B}}_{k/8}$ *computed via BiCGStab*(4) *(squares). The approximations obtained using the hybrid BiCG methods are often significantly more affected by rounding errors than* $\xi^{\mathrm{B}}_{k/2}$ *computed via BiCG. Here k denotes the number of matrix-vector multiplications. For BiCG, BiCGStab, and CGS we have* $k = 2n$ *(two matrix-vector multiplications per iteration). The value* $|\varsigma - \xi^{\mathrm{B}}_{k/2}|$ *is plotted every second value of k. For BiCGStab*(4) *the value* $|\varsigma - \xi^{\mathrm{B}}_{k/8}|$ *is plotted every eight values of k.*

necessary to select polynomial methods for the hybrid part that permit to compute the BiCG coefficients as accurately as possible." The difficulty in using hybrid BiCG methods for approximating the bilinear form $c^*A^{-1}b$ is illustrated in Figure 7.3 for BiCGStab, CGS, and BiCGStab(4). On the x-axis is the number of matrix-vector multiplications, which we denote by k. In all our computations we observed that for the hybrid BiCG methods the computed value ξ^{B}_n (see (3.20)) was always very close to the computed scattering amplitude $c^*\mathbf{x}_n$. This suggests that in hybrid BiCG methods both quantities are affected by rounding errors in a similar way. We observe that none of the hybrid BiCG methods performs in approximating the bilinear form $c^*A^{-1}b$ better than ξ^{B}_n computed via BiCG. On the contrary, in most cases they perform significantly worse. Techniques suggested in [46] applied to BiCGStab did not lead to a substantial improvement of the computed BiCGStab approximations.

In order to get an insight into this observation, we plot (as an example) in the upper part of Figure 7.4 the norm of the error $\|x - x_n\|$ (where x is determined via the MATLAB command `A\b`). Note that the approximations to the solution x lie

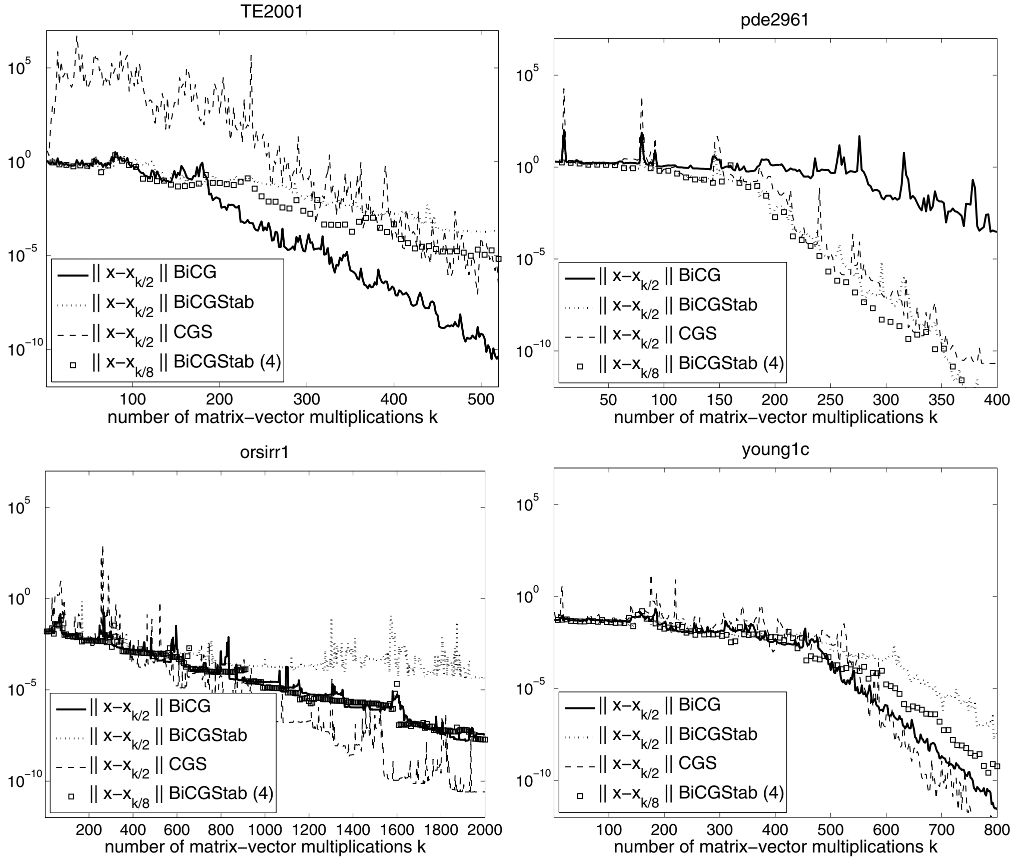


FIG. 7.4. *Euclidean norm of the error of the approximation to the solution of $Ax = b$ computed via BiCG and various hybrid BiCG methods. While BiCG seems to be a winner in approximating the bilinear form $c^*A^{-1}b$ (see Figure* 7.3*), hybrid BiCG methods are often more efficient in solving the system $Ax = b$.*

for various methods in Krylov subspaces of various dimensions. In particular, the BiCG approximation x_n lies in $\mathcal{K}_n(A, b)$, the CGS and BiCGStab approximations $\mathbf{x}_n$ lie in $\mathcal{K}_{2n}(A, b)$, and the BiCGStab(4) approximation $\mathbf{x}_n$ lies in $\mathcal{K}_{8n}(A, b)$. For the matrix `TE2001`, BiCG outperforms the other methods even in computing the approximate solution to $Ax = b$, while for the matrix `pde2961` it performs much worse than the hybrid BiCG methods, with BiCGStab(4) the winner. For `orsirr1` and `young1c`, there is no clear winner (a more detailed comparison of BiCG and hybrid BiCG methods as linear algebraic solvers is out of the scope of this paper). Despite the fact that $\|x - x_n\|$ is for `pde2961` worst for the BiCG algorithm, the behavior of $|s_n^*(x - x_n)|$ still causes ξ_n^{B} to behave even in this case about as well as the approximations computed via the hybrid BiCG methods.

In conclusion, in our experiments (this paper gives a small sample of them) the ξ_n^{B} computed via BiCG was not outperformed by the approximations computed via the hybrid BiCG methods. In most examples ξ_n^{B} computed via BiCG performed significantly better.

7.5. Transformation to the Hermitian positive definite case. From the approaches described in section 4, GLSQR performed in our experiments best both in terms of iteration count and in the number of matrix-vector multiplications. However,

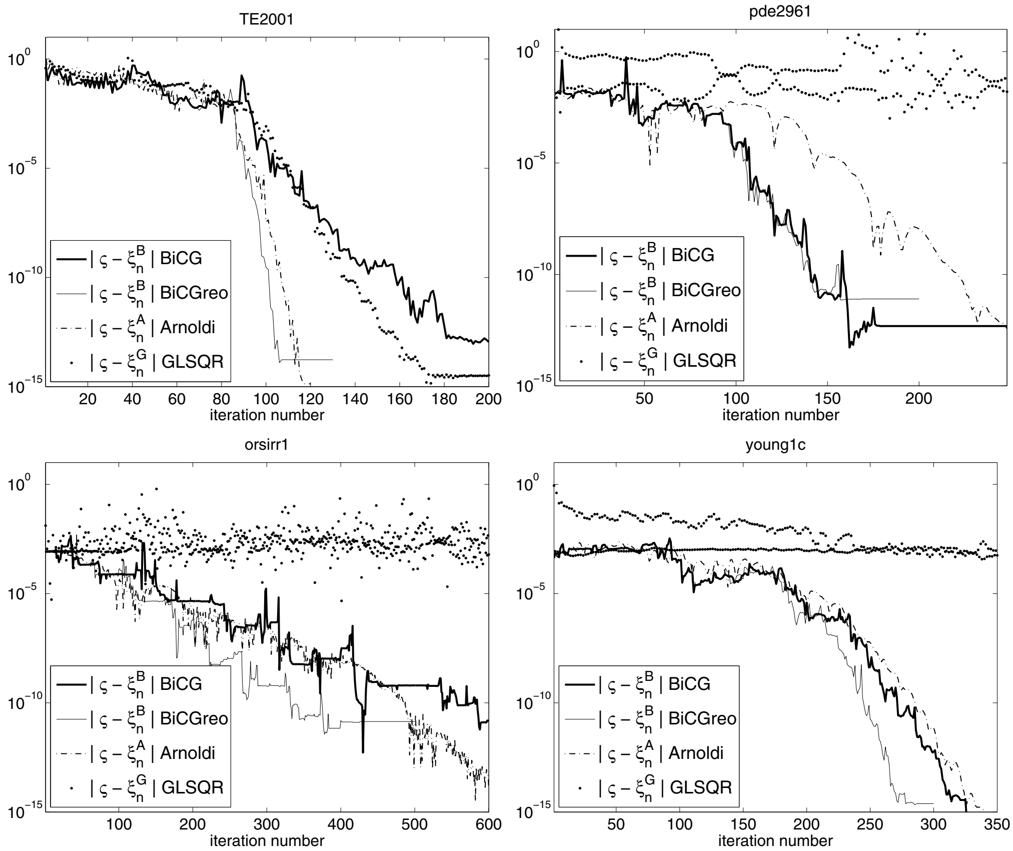


FIG. 7.5. *Comparison of errors for different approaches:* $|\varsigma - \xi_n^{\rm B}|$ *(bold solid line) from section* 3.1 *with* $\xi_n^{\rm B}$ *computed via BiCG,* $|\varsigma - \xi_n^{\rm B}(reo)|$ *(solid line) with* $\xi_n^{\rm B}(reo)$ *computed via BiCGreo,* $|\varsigma - \xi_n^{\rm A}|$ *(dash-dotted line) with* $\xi_n^{\rm A}$ *computed via the modified Gram–Schmidt Arnoldi algorithm from section* 3.3*, and* $|\varsigma - \xi_n^{\rm G}|$ *(dots) with* $\xi_n^{\rm G}$ *computed via GLSQR from section* 4.3.

even GLSQR was in most cases rather slow, as documented below. This observation cannot be explained by an effect of ill-conditioning of the matrix A^*A (in most of our experiments we used matrices with a moderate condition number). Results of further investigation of this topic will be reported elsewhere.

7.6. Comparison of approaches using different Krylov subspace methods. Figure 7.5 compares $|\varsigma - \xi_n^{\rm B}|$ with $\xi_n^{\rm B}$ (see (3.14)) from section 3.1 computed via BiCG (bold solid line), $|\varsigma - \xi_n^{\rm B}$ (reo)$|$ with $\xi_n^{\rm B}$ (reo) computed via BiCGreo (solid line), the error $|\varsigma - \xi_n^{\rm A}|$ with $\xi_n^{\rm A}$ (see (3.24)) computed via the modified Gram–Schmidt Arnoldi algorithm from section 3.3 (dash-dotted line), and the error $|\varsigma - \xi_n^{\rm G}|$ of the GLSQR approximation $\xi_n^{\rm G}$ (see (4.6)) from section 4.3 (dotted line).

We observe that the methods behave differently for different problems. Among the methods using short recurrences, the newly proposed approximation $\xi_n^{\rm B}$ wins except for `TE2001`, where $\xi_n^{\rm G}$ performs slightly better ($\xi_n^{\rm B}$ (reo) is not considered a practical alternative). For other problems GLSQR approximation $\xi_n^{\rm G}$ performs rather poorly (please notice the "double lines" for the problems `pde2961` and `young1c`). The approximation $\xi_n^{\rm A}$ computed via the modified Gram–Schmidt Arnoldi algorithm converges faster than the approximations based on short recurrences (except for `young1c`) but slower than $\xi_n^{\rm B}$ (reo). We emphasize that the cost of the Arnoldi iteration increases

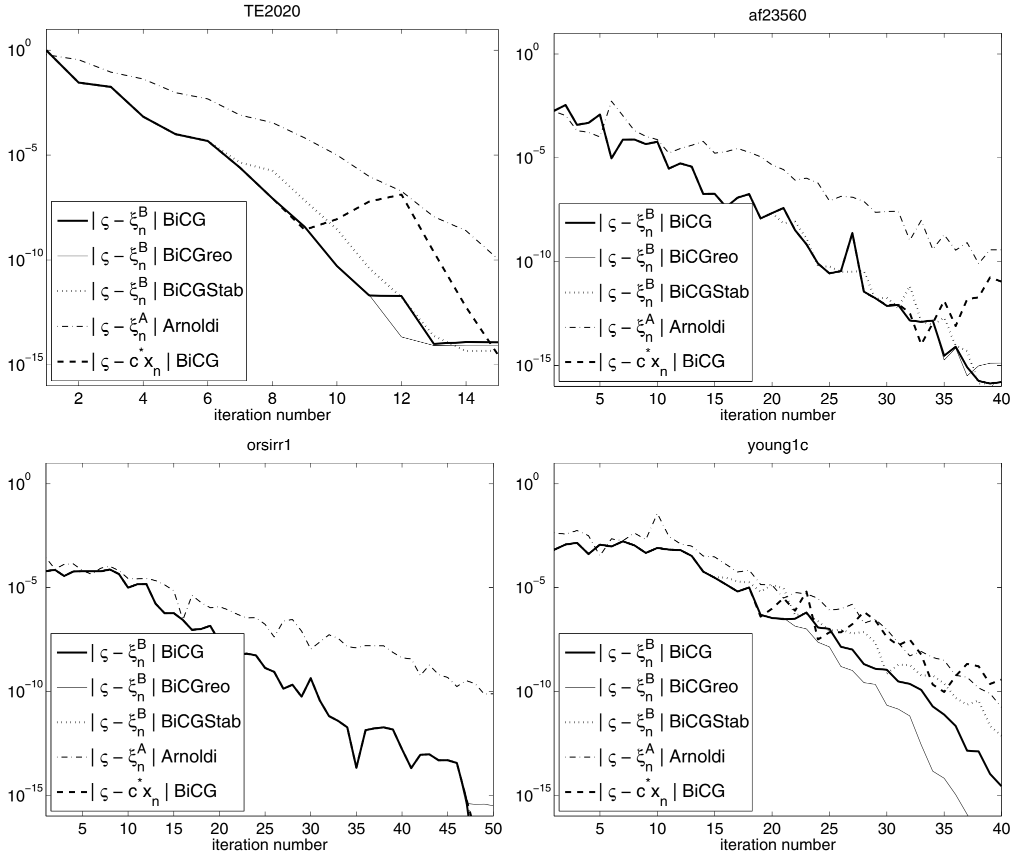


Fig. 7.6. *Comparison of errors for various approaches with preconditioning. The errors of approximation computed via BiCG (bold solid line), BiCGStab (dashed line), and BiCGreo (solid line) are for some problems very close to each other.*

with the iteration number n. The Arnoldi algorithm matches n moments, while the BiCG method matches $2n$ moments; see sections 2.1, 2.2, and 3.1. Since the Arnoldi algorithm uses orthogonal projections while the BiCG method uses oblique projections, the smaller number of matched moments alone does not explain the observed behavior. The cost of computations cannot be evaluated using matrix-vector products due to the fact that the other costs for methods based on short recurrences (BiCG and GLSQR) and long recurrences (the Arnoldi algorithm) differ significantly. In practical applications, the cost should be measured by computer time. In any case, our experiments suggest that the newly proposed ξ_n^{B} is highly competitive.

7.7. Preconditioning. In practice, iterative methods cannot be used without efficient preconditioning. In Figure 7.6 we illustrate results of computations for the same approaches as in Figure 7.5 except for GLSQR, which was skipped due to noncompetitive performance. (This does not mean, however, that GLSQR is in general noncompetitive. We were unable to make it work for our problems; the matter needs further investigation.) For `TE2020` we used a special preconditioning tailored to the problem; for `af23560` and `young1c` we used the incomplete Cholesky preconditioning with the drop tolerances 5×10^{-2} and 10^{-2}, respectively (they were found experimentally as good compromises between performance and fill-in). For the problem `orsirr1` we used the incomplete Cholesky preconditioning with zero fill-in. We can observe that all approaches based on short recurrences, except for the scattering amplitude approximation c^*x_n computed via BiCG, are comparable (except for `young1c` they are very close to or almost coincide with ξ_n^{B}). They clearly outperform ξ_n^{A} in terms of iterations. If the number of iterations is small, the comparison on a real-world problem with a significant cost of the matrix-vector multiplication might, however, be more in favor of ξ_n^{A} computed via the Arnoldi algorithm. It is worth pointing out that due to long recurrences ξ_n^{A} can safely be computed via FOM using c^*x_n; see (3.24).

8. Concluding remarks. This paper proposes the new approximation ξ_n^{B} for the bilinear form $c^*A^{-1}b$ and compares it to the existing approaches. We have linked the presented approximations to the matching moment properties of the Krylov subspace methods. While the maximal number of moments matched at step n of the Hermitian and non-Hermitian Lanczos algorithm and BiCG is $2n$, the Arnoldi algorithm matches at step n only n moments. Matching $2n$ moments using oblique projections, however, does not necessarily mean an advantage over using the Arnoldi algorithm with orthogonal projections (at the price of computing long recurrences) and matching n moments only. In practice, the cost evaluation must take into account specifics of the given application problem which determine, e.g., the cost of the matrix-vector products in relation to the cost of the iteration updates. Therefore the choice of an optimal approach (including a choice of stopping criteria) is application-dependent. Nevertheless, the newly proposed approximation ξ_n^{B} (see (3.14)) is, in our opinion, highly competitive and can be considered a good reference standard for any other possible approach. The approximation error can be estimated using techniques based on computing d additional iterations analogously to CG [52, section 4]; see also [34, section 5.3] and, in the context of constructing stopping criteria in numerical solution of PDEs, e.g., [1, 31]. The approximation ξ_n^{B} clearly outperforms the mathematically equivalent scattering amplitude approximation c^*x_n. Scattering amplitude approximations computed via short recurrences rely upon preserving global biorthogonality among the computed vectors. Their convergence is delayed due to rounding errors much more than convergence of the approximation ξ_n^{B}, and therefore they should not be used in practical computations.

Acknowledgments. The authors are indebted to Paul Van Dooren and Volker Mehrmann for their useful comments and to Gerard Meurant for his valuable suggestions. The presentation also benefits from the comments on the previous version of the manuscript given by Karen Willcox and by the very valuable comments of three anonymous referees. The authors express their thanks to Margaret Wright for extraordinarily detailed comments and advice.

REFERENCES

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

[2] Z. Bai, *Krylov subspace techniques for reduced-order modeling of large-scale dynamical systems*, Appl. Numer. Math., 43 (2002), pp. 9–44.

[3] C. Brezinski, *Projection Methods for Systems of Equations*, Stud. Comput. Math. 7, North–Holland, Amsterdam, 1997.

[4] A. Bultheel and M. Van Barel, *Linear Algebra, Rational Approximation and Orthogonal Polynomials*, Stud. Comput. Math. 6, North–Holland, Amsterdam, 1997.

[5] D. Calvetti, S.-M. Kim, and L. Reichel, *Quadrature rules based on the Arnoldi process*, SIAM J. Matrix Anal. Appl., 26 (2005), pp. 765–781.

[6] D. Calvetti, S. Morigi, L. Reichel, and F. Sgallari, *Computable error bounds and estimates for the conjugate gradient method*, Numer. Algorithms, 25 (2000), pp. 75–88.

[7] P. Chebyshev, *Sur les fractions continues*, J. Math. Pures Appl., Ser. II, 3 (1855), pp. 289–293.

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

[9] G. Dahlquist, G. H. Golub, and S. G. Nash, *Bounds for the error in linear systems*, in Semi-infinite programming (Proc. Workshop, Bad Honnef, 1978), Lecture Notes in Control and Inform. Sci. 15, Springer, Berlin, 1979, pp. 154–172.

[10] R. Fletcher, *Conjugate gradient methods for indefinite systems*, in Numerical Analysis (Proc. 6th Biennial Dundee Conf., Univ. Dundee, Dundee, 1975), Lecture Notes in Math. 506, Springer, Berlin, 1976, pp. 73–89.

[11] R. W. Freund, *Model reduction methods based on Krylov subspaces*, Acta Numer., 12 (2003), pp. 267–319.

[12] R. W. Freund and M. Hochbruck, *Gauss quadratures associated with the Arnoldi process and the Lanczos algorithm*, in Linear Algebra for Large Scale and Real-Time Applications, NATO ASI, Ser. E 232, M. S. Moonen, G. H. Golub, and B. L. R. De Moor, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1993, pp. 377–380.

[13] K. Gallivan, E. Grimme, and P. Van Dooren, *Asymptotic waveform evaluation via a Lanczos method*, Appl. Math. Lett., 7 (1994), pp. 75–80.

[14] G. Golub and W. Kahan, *Calculating the singular values and pseudo-inverse of a matrix*, J. Soc. Indust. Appl. Math. Ser. B Numer. Anal., 2 (1965), pp. 205–224.

[15] G. H. Golub, *Some modified matrix eigenvalue problems*, SIAM Rev., 15 (1973), pp. 318–334.

[16] G. H. Golub, *Milestones in Matrix Computation: Selected Works of Gene H. Golub, with Commentaries*, R. H. Chan, C. Greif, and D. P. O'Leary, eds., Oxford Science Publications, Oxford University Press, Oxford, UK, 2007.

[17] G. H. Golub and G. Meurant, *Matrices, moments and quadrature*, in Numerical Analysis 1993 (Dundee, 1993), Pitman Res. Notes Math. Ser. 303, Longman Sci. Tech., Harlow, UK, 1994, pp. 105–156.

[18] G. H. Golub and G. Meurant, *Matrices, moments and quadrature.* II. *How to compute the norm of the error in iterative methods*, BIT, 37 (1997), pp. 687–705.

[19] G. H. Golub and G. Meurant, *Matrices, Moments and Quadrature With Applications*, Princeton University Press, Princeton, NJ, 2010.

[20] G. H. Golub, M. Stoll, and A. Wathen, *Approximation of the scattering amplitude and linear systems*, Electron. Trans. Numer. Anal., 31 (2008), pp. 178–203.

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

[22] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins Stud. Math. Sci., The Johns Hopkins University Press, Baltimore, MD, 1996.

[23] R. G. Gordon, *Error bounds in equilibrium statistical mechanics*, J. Math. Phys., 9 (1968), pp. 655–663.

[24] W. B. Gragg, *Matrix interpretations and applications of the continued fraction algorithm*, Rocky Mountain J. Math., 4 (1974), pp. 213–225.

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

[26] A. Greenbaum, *Estimating the attainable accuracy of recursively computed residual methods*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 535–551.

[27] M. H. Gutknecht, *Variants of BICGSTAB for matrices with complex spectrum*, SIAM J. Sci. Comput., 14 (1993), pp. 1020–1033.

[28] J. J. Hench and Z. Strakoš, *The RCWA method—a case study with open questions and perspectives of algebraic computations*, Electron. Trans. Numer. Anal., 31 (2008), pp. 331–357.

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

[30] K. Hoffman and R. Kunze, *Linear Algebra*, 2nd ed., Prentice–Hall, Englewood Cliffs, NJ, 1971.

[31] P. Jiránek, Z. Strakoš, and M. Vohralík, *A posteriori error estimates including algebraic error and stopping criteria for iterative solvers*, SIAM J. Sci. Comput., 32 (2010), pp. 1567–1590.

[32] P. Jordan and J. Von Neumann, *On inner products in linear, metric spaces*, Ann. of Math. (2), 36 (1935), pp. 719–723.

[33] C. Lanczos, *Solution of systems of linear equations by minimized iterations*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 33–53.

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

[35] M. Neviere and E. Popov, *Light Propagation in Periodic Media*, Marcel Dekker, New York, 2002.

[36] C. C. Paige, M. Rozložník, and Z. Strakoš, *Modified Gram–Schmidt (MGS), least squares, and backward stability of MGS-GMRES*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 264–284.

[37] C. C. Paige and M. A. Saunders, *Algorithm* 583*: LSQR: An algorithm for sparse linear equations and sparse least squares*, ACM Trans. Math. Software, 8 (1982), pp. 195–201.

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

[39] L. Reichel and Q. Ye, *A generalized LSQR algorithm*, Numer. Linear Algebra Appl., 15 (2008), pp. 643–660.

[40] W. P. Reinhardt, *l^2 discretization of atomic and molecular electronic continua: Moment, quadrature and j-matrix techniques*, Comput. Phys. Comm., 17 (1979), pp. 1–21.

[41] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed., SIAM, Philadelphia, 2003.

[42] M. A. Saunders, H. D. Simon, and E. L. Yip, *Two conjugate-gradient-type methods for unsymmetric linear equations*, SIAM J. Numer. Anal., 25 (1988), pp. 927–940.

[43] P. E. Saylor and D. C. Smolarski, *Addendum to: "Why Gaussian quadrature in the complex plane?"* [Numer. Algorithms, 26 (2001), pp. 251–280], Numer. Algorithms, 27 (2001), pp. 215–217.

[44] P. E. Saylor and D. C. Smolarski, *Why Gaussian quadrature in the complex plane?*, Numer. Algorithms, 26 (2001), pp. 251–280.

[45] L. Schlessinger and C. Schwartz, *Analyticity as a useful computational tool*, Phys. Rev. Lett., 16 (1966), pp. 1173–1174.

[46] G. L. G. Sleijpen and H. A. van der Vorst, *Maintaining convergence properties of BiCGstab methods in finite precision arithmetic*, Numer. Algorithms, 10 (1995), pp. 203–223.

[47] G. L. G. Sleijpen, H. A. van der Vorst, and D. R. Fokkema, BiCGstab(l) *and other hybrid Bi-CG methods*, Numer. Algorithms, 7 (1994), pp. 75–109.

[48] P. Sonneveld, *CGS, a fast Lanczos-type solver for nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 36–52.

[49] T. J. Stieltjes, *Recherches sur les fractions continues*, Ann. Fac. Sci. Toulouse Sci. Math. Sci. Phys., 8 (1894), pp. 1–122.

[50] M. Stoll, *Solving Linear Systems Using the Adjoint*, Ph.D. thesis, University of Oxford, Oxford, UK, 2009.

[51] Z. Strakoš, *Model reduction using the Vorobyev moment problem*, Numer. Algorithms, 51 (2009), pp. 363–379.

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

[53] Z. Strakoš and P. Tichý, *Error estimation in preconditioned conjugate gradients*, BIT, 45 (2005), pp. 789–817.

[54] H. A. van der Vorst, *Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 631–644.

[55] Y. V. Vorobyev, *Methods of Moments in Applied Mathematics*, Gordon and Breach Science Publishers, New York, 1965.

[56] K. F. Warnick and W. C. Chew, *Numerical simulation methods for rough surface scattering*, Waves Random Media, 11 (2001), pp. R1–R30.

[57] N. Young, *An Introduction to Hilbert Space*, Cambridge Math. Textbooks, Cambridge University Press, Cambridge, UK, 1988.

Electronic Transactions on Numerical Analysis.
Volume 26, pp. 453-473, 2007.
Copyright © 2007, Kent State University.
ISSN 1068-9613.

ON WORST-CASE GMRES, IDEAL GMRES, AND THE POLYNOMIAL NUMERICAL HULL OF A JORDAN BLOCK*

PETR TICHÝ*, JÖRG LIESEN†, AND VANCE FABER‡

Abstract. When solving a linear algebraic system $Ax = b$ with GMRES, the relative residual norm at each step is bounded from above by the so-called ideal GMRES approximation. This worst-case bound is sharp (i.e. it is attainable by the relative GMRES residual norm) in case of a normal matrix A, but it need not characterize the worst-case GMRES behavior if A is nonnormal. Characterizing the tightness of this bound for nonnormal matrices A represents an important and largely open problem in the convergence analysis of Krylov subspace methods. In this paper we address this problem in case A is a single Jordan block. We study the relation between ideal and worst-case GMRES as well as the problem of estimating the ideal GMRES approximation. Furthermore, we prove new results about the radii of the polynomial numerical hulls of Jordan blocks. Using these, we discuss the closeness of the lower bound on the ideal GMRES approximation that is derived from the radius of the polynomial numerical hull.

Key words. GMRES convergence, ideal GMRES, polynomial numerical hull, Jordan block.

AMS subject classifications. 65F10, 65F35, 49K35.

1. Introduction.

Let a nonsingular matrix $A \in \mathbb{C}^{n\times n}$ and a vector $b \in \mathbb{C}^n$ be given. Suppose that we apply the GMRES method [14] with initial guess $x_0 = 0$ (chosen here for convenience and without loss of generality) to the linear system $Ax = b$. Then this method computes a sequence of iterates $x_1, x_2, \ldots$, so that the kth residual $r_k \equiv b - Ax_k$ satisfies

$$\|r_k\| \;=\; \min_{p\in\pi_k} \|p(A)\,b\|. \tag{1.1}$$

Here π_k denotes the set of (complex) polynomials of degree at most k and with value one at the origin, and $\|\cdot\|$ denotes the Euclidean norm. The residual r_k is uniquely determined by the minimization condition (1.1) and satisfies the equivalent orthogonality condition

$$r_k \in b + A\mathcal{K}_k(A,b), \qquad r_k \perp A\mathcal{K}_k(A,b). \tag{1.2}$$

Here $\mathcal{K}_k(A,b) \equiv \text{span}\{b, Ab, \ldots A^{k-1}b\}$ is the kth Krylov subspace generated by A and b, and $\perp$ means orthogonality with respect to the Euclidean inner product. Without loss of generality we will consider that b is a unit norm vector, i.e. $\|b\| = 1$.

A common approach for investigating the GMRES convergence behavior is to bound (1.1) independently of b, and thus to study the algorithm's worst-case behavior. In particular, for each iteration step k one may analyze the *worst-case GMRES approximation*

$$\psi_k(A) \;\equiv\; \max_{\|v\|=1}\min_{p\in\pi_k} \|p(A)v\|. \tag{1.3}$$

The quantity $\psi_k(A)$ is attainable by the GMRES residual norm in the following sense: For a given matrix A and every GMRES step k, there exists a unit norm initial vector b, for which the resulting kth GMRES residual norm is equal to $\psi_k(A)$. It should be noted, however, that

*Received March 6, 2007. Accepted for publication September 17, 2007. Recommended by M. Gutknecht. The work of the first author was supported by the Emmy Noether-Programm of the Deutsche Forschungsgemeinschaft and by the Czech National Program of Research "Information Society" under project 1ET400300415. The work of the second author was supported by the Emmy Noether-Programm of the Deutsche Forschungsgemeinschaft.

*Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 2, 18207 Prague, Czech Republic (tichy@cs.cas.cz).

†Institute of Mathematics, Technical University of Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany (liesen@math.tu-berlin.de).

‡BD Biosciences - Bioimaging Systems (vance_faber@bd.com).

for a given nonnormal matrix A and integer k the quantity $\psi_k(A)$ typically is very hard to compute. In fact, we are unaware of any efficient algorithm for performing this computation.

Using the submultiplicativity of the Euclidean norm (or by changing the order of maximization and minimization in (1.3)), we can easily find the following upper bound on (1.3),

$$\psi_k(A) \;\leq\; \min_{p\in\pi_k} \|p(A)\| \;\equiv\; \varphi_k(A). \tag{1.4}$$

The quantity $\varphi_k(A)$, called the kth *ideal GMRES approximation*, has been introduced by Greenbaum and Trefethen [7]. They argue that it is important to investigate this quantity to improve the understanding of GMRES (and matrix iterations in general) particularly in the nonnormal case, since the ideal GMRES approximation "disentangles the matrix essence of the [GMRES] process from the distracting effects of the initial vector", see [7, p. 362].

Before continuing this line of thought we have to stress a subtle point: In case $A \in \mathbb{R}^{n\times n}$ it is customary (and we will follow this custom) to assume that $b \in \mathbb{R}^n$, and to consider the approximation problem (1.3) only for $v \in \mathbb{R}^n$. In this (real) case, the values $\psi_k(A)$ and $\varphi_k(A)$ are both attained by real polynomials $p \in \pi_k$. For the worst-case GMRES approximation $\psi_k(A)$ this fact is obvious, while for the ideal GMRES approximation $\varphi_k(A)$ this has been shown in [10, Theorem 3.1].

After the 1994 paper [7], several studies have been devoted to the problem of characterizing the relation between $\psi_k(A)$ and $\varphi_k(A)$, and in particular the tightness of the inequality (1.4). The best known result is that (1.4) is an equality, i.e. $\psi_k(A) = \varphi_k(A)$ for all $k \geq 0$, whenever A is normal [6, 11]. In addition, (1.4) is an equality for arbitrary A and $k = 1$ [6, 11], for triangular Toeplitz matrices when the right hand side of (1.4) equals one [3], and also when the matrix $p_*^{(k)}(A)$ that solves the ideal GMRES approximation problem (1.4) has a simple maximal singular value [6, Lemma 2.4]. On the other hand, some examples of nonnormal matrices have been constructed, for which (1.4) is a sharp inequality [3, 17]. Despite the existence of these counterexamples, it is still an open question whether (1.4) is an equality (or at least tight inequality) for larger classes of nonnormal matrices.

Another open problem in the context of (1.4) is how to determine or estimate the value of the ideal GMRES approximation $\varphi_k(A)$ in general. A possible approach that is still under development is to associate the matrix A with some set in the complex plane and to relate the norm of the matrix polynomial to the maximum norm of the polynomial on this set. An appropriate set, designed to give useful information about the norm of functions of a matrix A, is the *polynomial numerical hull of degree* k,

$$\mathcal{H}_k(A) \;\equiv\; \{z \in \mathbb{C} : \; \|p(A)\| \geq |p(z)| \;\text{ for all } p \in \mathcal{P}_k\}, \tag{1.5}$$

introduced by Nevanlinna [13, p. 41]. Here $\mathcal{P}_k$ denotes the set of (complex) polynomials of degree at most k. Based on the definition (1.5) it is not hard to see that these sets provide a lower bound on the ideal GMRES approximation [4],

$$\min_{p\in\pi_k} \max_{z\in\mathcal{H}_k(A)} |p(z)| \;\leq\; \varphi_k(A). \tag{1.6}$$

Moreover, $\mathcal{H}_k(A)$ allows us to identify when ideal GMRES fails to converge [3, 4],

$$\varphi_k(A) = 1 \quad\Longleftrightarrow\quad 0 \in \mathcal{H}_k(A). \tag{1.7}$$

While polynomial numerical hulls appear to be a valuable tool, their determination or computation represents a difficult open problem even for simple classes of nonnormal matrices.

In summary, the investigation of worst-case and ideal GMRES as well as the polynomial numerical hulls for nonnormal matrices is at its very beginning. We believe that in this situation it is helpful to study relatively simple nonnormal matrices, for which explicit solutions of some of the open problems can be derived. Continuing the work started in [2] and [5], we here consider A being an $n \times n$ Jordan block J_λ with eigenvalue $\lambda \in \mathbb{C}$.

When experimenting with the MATLAB software SDPT3 [18] and some Jordan blocks J_λ of small size ($n = 20$, say), we observed numerically that $\psi_k(J_\lambda) = \varphi_k(J_\lambda)$ for $0 \le k \le n$. This led us to conjecture that

$$\psi_k(J_\lambda) = \varphi_k(J_\lambda) \quad \text{for all } \lambda, n \text{ and } 0 \le k \le n.$$

At first sight, proving this conjecture looks not too difficult; after all, one just has to deal with a single Jordan block. However, it turns out that the approximation problems behind the quantities $\psi_k(A)$ and $\varphi_k(A)$ as well as the exact determination of $\mathcal{H}_k(A)$ are highly nontrivial even in case $A = J_\lambda$. When trying to prove our conjecture we found that numerous cases need to be distinguished, and in the end we were unable to prove all of them. Nevertheless, we believe that the work presented here has been worthwhile. In particular, it uncovered a previously unknown structure behind the worst-case and ideal GMRES approximation problems in case $A = J_\lambda$, it extended the recent results of [2, 5] on the polynomial numerical hulls of Jordan blocks, and it led to new results about the bound (1.6).

Since the presentation below is rather technical, we give a detailed overview of the sections and the corresponding results in this paper:

- In section 2 we summarize known results on worst-case and ideal GMRES as well as the polynomial numerical hull.
- In section 3 we show that $\psi_k(J_\lambda) = \varphi_k(J_\lambda)$ for $0 \le k < n/2$ and whenever $|\lambda|$ is outside a small interval on the positive real line.
- In section 4 we study the structure of the polynomials that solve the ideal GMRES approximation problem, i.e. the polynomials for which the value $\varphi_k(J_\lambda)$ is attained. This allows us to show that $\varphi_k(J_\lambda) = \psi_k(J_\lambda)$ for all λ in case k divides n. Moreover, we establish a relationship between the radii of polynomial numerical hulls of J_λ.
- In section 5 we analyze the quantities $\psi_{n-1}(J_\lambda)$ and $\varphi_{n-1}(J_\lambda)$. This allows us to show that $\varphi_{n-k}(J_\lambda) = \psi_{n-k}(J_\lambda)$ whenever $|\lambda| \ge 1$ and k divides n.
- Finally, in section 6 we apply results of the previous sections to analyze the closeness of the bound (1.6) on the kth ideal GMRES approximation. We are unaware that any theoretical results in this direction have been obtained previously.

2. Notation and theoretical background.

The following result collects a number of basic results concerning the quantities $\psi_k(A)$ and $\varphi_k(A)$. These results are either easy to verify, or they have been published in [10, Theorem 3.1] or [3, Proposition 2.1].

LEMMA 2.1. *Let $A \in \mathbb{C}^{n\times n}$ be a matrix with minimal polynomial degree $d(A)$. Then the following hold:*

1. *$\psi_k(A)$ and $\varphi_k(A)$ are both nonincreasing in k.*
2. *$\psi_0(A) = \varphi_0(A) = 1$.*
3. *$0 < \psi_k(A) \le \varphi_k(A)$ for $1 \le k \le d(A) - 1$.*
4. *If A is nonsingular, then $\psi_k(A) = \varphi_k(A) = 0$ for all $k \ge d(A)$.*
5. *If A is singular, then $\psi_k(A) = \varphi_k(A) = 1$ for all $k \ge 0$.*

The previous theorem shows that to investigate the relation between worst-case and ideal GMRES, one only has to consider nonsingular matrices A and positive integers $k < d(A)$. In this case $\varphi_k(A) > 0$, and the polynomial that solves the ideal GMRES approximation problem (1.4) is uniquely determined [7, Theorem 2]. This gives rise to the following definition.

DEFINITION 2.2. *For a nonsingular matrix $A \in \mathbb{C}^{n\times n}$, and a positive integer $k < d(A)$, the uniquely determined polynomial $p_*^{(k)} \in \pi_k$ that satisfies*

$$\|p_*^{(k)}(A)\| = \varphi_k(A) = \min_{p\in\pi_k} \|p(A)\|,$$

is called the kth ideal GMRES polynomial of A, and the matrix $p_*^{(k)}(A)$ is called the kth ideal GMRES residual matrix of A.
The matrix A is called ideal of degree k, when $\varphi_k(A) = \psi_k(A)$, and A is called ideal, when $\varphi_k(A) = \psi_k(A)$ for $k = 1, \ldots, d(A) - 1$.

We point out that if A is ideal of some degree k, then this does not necessarily imply that A is ideal of any other degree. In fact, it would be interesting to characterize necessary and sufficient conditions on A that allow one to conclude from idealness of some degree to idealness of other degrees.

In general it is an open problem which properties of A are necessary and sufficient for A to be ideal. Below we summarize the most important results for our context. Proofs of all of these statements can be found in [6, 11].

LEMMA 2.3. *Any nonsingular matrix $A \in \mathbb{C}^{n\times n}$ is ideal of degree $k = 1$. Moreover:*

1. *If A is normal, then A is ideal.*
2. *If $p_*^{(k)}(A)$ has a simple maximal singular value, then A is ideal of degree k.*

Let us discuss the condition in the second item. If A is a normal matrix with (distinct) eigenvalues $\lambda_1, \ldots, \lambda_{d(A)}$, then the ideal GMRES approximation problem is a (scalar) min-max problem on the set of the eigenvalues,

$$\varphi_k(A) \;=\; \min_{p\in\pi_k} \|p(A)\| \;=\; \min_{p\in\pi_k} \max_{\lambda_i} |p(\lambda_i)|.$$

It is well known that the corresponding min-max polynomial of degree k attains its maximum value on at least $k + 1$ of the eigenvalues, see, e.g., [1, Chapter 3, §4]. Hence in this case the multiplicity of the maximal singular value of $p_*^{(k)}(A)$ is at least $k + 1$. Since any normal matrix is ideal, we see that the condition in the second item is *not necessary*.

This fact has already been noted, and explained by a similar argument, by Greenbaum and Trefethen [7]. Based on some numerical observations, they consider the case in which $p_*^{(k)}(A)$ for a nonnormal matrix A has a simple maximal singular value the "generic case", see [7, p. 366]. However, we believe that the situation of $p_*^{(k)}(A)$ having a multiple maximal singular value can be quite frequent also for nonnormal A. For a clear example see Fig. 4.1 below, which shows that for the 20 × 20 Jordan block J_λ with $\lambda = 1$, only 9 out of 19 matrices $p_*^{(k)}(J_\lambda)$ have a simple maximal singular value.

We denote the maximal singular value of a matrix B by $\sigma_{max}(B)$, and we define the linear space

$$\Sigma(B) \;\equiv\; \text{span}\{v : v \text{ is a right singular vector of } B \text{ corresponding to } \sigma_{max}(B)\}.$$

We use such spaces in the next result, which gives a further characterization of the case $\psi_k(A) = \varphi_k(A)$. This result can be found in a more general form in [3, Lemma 2.16], but we formulate and prove it here independently of [3].

LEMMA 2.4. *Suppose that a nonsingular matrix A and a positive integer $k < d(A)$ are given. Then $\psi_k(A) = \varphi_k(A)$ if and only if there exist a polynomial $q \in \pi_k$ and a unit norm vector $b \in \Sigma(q(A))$, such that*

$$q(A)b \perp A\mathcal{K}_k(A, b). \tag{2.1}$$

If such q and b exist, then $q = p_*^{(k)}$.

Proof. If $\psi_k(A) = \varphi_k(A)$, then there exist a unit norm vector b and a polynomial $q \in \pi_k$ satisfying (2.1), cf. (1.2), such that $\|p_*^{(k)}(A)\| = \|q(A)b\|$. Since $\|p_*^{(k)}(A)b\| \le \|p_*^{(k)}(A)\|$ and $\|q(A)b\|$ is minimal,

$$\|p_*^{(k)}(A)b\| = \|p_*^{(k)}(A)\| = \|q(A)b\|. \tag{2.2}$$

But this means that $b \in \Sigma(p_*^{(k)}(A))$. Moreover, since $1 \le k \le d(A) - 1$, we know that $\psi_k(A) > 0$ by Lemma 2.1, and thus the kth GMRES polynomial is unique, cf. [7, Theorem 2]. Therefore $p_*^{(k)} = q$, and hence $b \in \Sigma(q(A))$.

Now assume that there exist a polynomial $q \in \pi_k$ and a unit norm vector b such that (2.1) holds and $b \in \Sigma(q(A))$. Then

$$\|q(A)\| = \|q(A)b\| = \min_{p \in \pi_k} \|p(A)b\| \le \|p_*^{(k)}(A)\|. \tag{2.3}$$

Since $p_*^{(k)}$ is the ideal GMRES polynomial, $\|q(A)\| < \|p_*^{(k)}(A)\|$ is impossible, and therefore equality holds in (2.3). But then $\psi_k(A) = \varphi_k(A)$, and from uniqueness of $p_*^{(k)}$ it follows that $q = p_*^{(k)}$. □

In [3], the k-dimensional generalized field of values of A,

$$F(\{A^i\}_{i=1}^k) \equiv \left\{ \begin{pmatrix} v^*Av \\ \vdots \\ v^*A^kv \end{pmatrix} : v^*v = 1 \right\},$$

is used to characterize when worst-case or ideal GMRES do not converge.

THEOREM 2.5. *For a nonsingular matrix $A \in \mathbb{C}^{n\times n}$ the following hold:*

1. $\psi_k(A) = 1$ *if and only if* $0 \in F(\{A^i\}_{i=1}^k)$.
2. $\varphi_k(A) = 1$ *if and only if* $0 \in \mathrm{cvx}(F(\{A^i\}_{i=1}^k))$, *the convex hull of* $F(\{A^i\}_{i=1}^k)$.

Note that when $A \in \mathbb{R}^{n\times n}$ is real, one can take the *real* k-dimensional generalized field of values A defined over $v \in \mathbb{R}^n$, $v^Tv = 1$.

The k-dimensional generalized field of values of any triangular Toeplitz matrix $T \in \mathbb{C}^{n\times n}$ is a convex set [3], and, therefore,

$$\psi_k(T) = 1 \quad \Longleftrightarrow \quad \varphi_k(T) = 1, \tag{2.4}$$

i.e. T is ideal of degree k in case of stagnation. However, it is in general still an open problem, originally posed in [3, p. 722], whether T is ideal of degree k when ideal GMRES converges, i.e. when $\varphi_k(T) < 1$.

In this paper we concentrate on an $n \times n$ Jordan block

$$J_\lambda = \begin{pmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{pmatrix} \equiv \lambda I_n + E_n. \tag{2.5}$$

Apart from the identity matrix I_n and the shift E_n, we will use the backward identity I_n^B and the diagonal matrix $I_n^\pm$ defined by

$$I_n^B \equiv \begin{pmatrix} & & 1 \\ & \cdot^{\cdot^{\cdot}} & \\ 1 & & \end{pmatrix} \in \mathbb{R}^{n\times n}, \quad I_n^\pm \equiv \mathrm{diag}(1, -1, \ldots, (-1)^{n-1}). \tag{2.6}$$

As explained above, the singular case ($\lambda = 0$) is uninteresting, so we only consider the *nonsingular case*, i.e. $\lambda \neq 0$. Each $\lambda \in \mathbb{C}$ can be written as $\lambda = |\lambda| e^{\mathbf{i}\alpha}$, and it holds that

$$J_\lambda \;=\; e^{\mathbf{i}\alpha} U J_{|\lambda|} U^H, \qquad U \equiv \mathrm{diag}(e^{\mathbf{i}\alpha}, e^{\mathbf{i}2\alpha}, \ldots, e^{\mathbf{i}n\alpha}). \tag{2.7}$$

Since J_λ is unitarily similar to $J_{|\lambda|}$, and the values of the approximation problems we deal with are unitarily invariant, it suffices to consider *real and positive* λ. All results can be easily extended to all $\lambda \in \mathbb{C}$ using the unitary similarity transformation defined by (2.7). Since $d(J_\lambda) = n$, we will consider $k = 1, \ldots, n-1$, so that $0 < \psi_k(J_\lambda) \le \varphi_k(J_\lambda)$, and the corresponding ideal GMRES polynomials are well defined in the sense of Definition 2.2.

As mentioned in the Introduction, the polynomial numerical hull (1.5) appears to be useful in the analysis of ideal GMRES. As shown in [2], for each $k = 1, \ldots, n-1$, $\mathcal{H}_k(J_\lambda)$ is a circle around the eigenvalue λ with some radius $\varrho_{k,n}$, where

$$0 < \varrho_{n-1,n} < \cdots < \varrho_{1,n} < 1,$$

and $\varrho_{k,n}$ is independent of the eigenvalue λ. In particular, the authors of [2] concentrate on determining the radii $\varrho_{1,n}$ and $\varrho_{n-1,n}$. Since $\mathcal{H}_1(J_\lambda)$ is equal to the field of values of J_λ, it holds that

$$\varrho_{1,n} = \cos\left(\tfrac{\pi}{n+1}\right), \tag{2.8}$$

cf. [2, p. 235]. The problem of determining $\varrho_{n-1,n}$ is equivalent to a classical problem in complex approximation theory, closely related to the Carathéodory-Fejér interpolation problem. Using this connection it is shown in [2, p. 238], that $\varrho_{n-1,n}$ is a solution of a certain nonlinear problem and can be bounded by

$$1 - \tfrac{\log(2n)}{n} \le \varrho_{n-1,n} \le 1 - \tfrac{\log(2n)}{n} + \tfrac{\log(\log(2n))}{n}. \tag{2.9}$$

Continuing this work, Greenbaum [5, p. 88] combines (1.6), (2.4) and results of [2] to prove that for $k = 1, \ldots, n-1$,

$$\varrho_{k,n}^k \lambda^{-k} \le \varphi_k(J_\lambda) \;\le\; \lambda^{-k} \quad \text{for} \quad \lambda \ge \varrho_{k,n}, \tag{2.10}$$

$$\psi_k(J_\lambda) \;=\; \varphi_k(J_\lambda) \;=\; 1 \quad \Longleftrightarrow \quad \lambda \le \varrho_{k,n}. \tag{2.11}$$

The upper bound on $\varphi_k(J_\lambda)$ in (2.10) can be replaced by 1 if $\lambda \le 1$. The lower bound in (2.10) is a special case of the general lower bound (1.6) on the ideal GMRES approximation based on the polynomial numerical hull. The closeness of this lower bound is examined in section 6 below.

We point out that the lower bound on $\varrho_{n-1,n}$ in (2.9) approaches 1 as $n \to \infty$. Hence the equivalence (2.11) implies that for each λ with $0 < |\lambda| < 1$, there exists a positive integer $n = n_\lambda$ such that for the $n \times n$ Jordan block J_λ, $\psi_{n-1}(J_\lambda) = \varphi_{n-1}(J_\lambda) = 1$. In other words, both worst-case and ideal GMRES stagnate completely for each Jordan block J_λ corresponding to an eigenvalue λ inside the unit circle, provided that J_λ is sufficiently large. The more interesting cases are therefore the Jordan blocks J_λ with $|\lambda| \ge 1$.

3. Worst-case and ideal GMRES for $k < n/2$. In this section we show that if $|\lambda|$ is outside a small interval around one, then J_λ is ideal of degree k for $1 \le k < n/2$. We start with a general characterization of the radius $\varrho_{k,n}$ of the polynomial numerical hull of degree k of J_λ.

LEMMA 3.1. *A positive real number ϱ satisfies $\varrho \le \varrho_{k,n}$, if and only if there exists a real unit norm vector b such that*

$$b^T E_n^j b = (-\varrho)^j, \qquad j = 1, \ldots, k. \tag{3.1}$$

Proof. A positive real number ϱ satisfies $\varrho \leq \varrho_{k,n}$, if and only if an $n \times n$ Jordan block J_ϱ satisfies $\psi_k(J_\varrho) = \varphi_k(J_\varrho) = 1$, cf. (2.11). This is equivalent with the existence of a real unit norm vector b such that

$$b \perp J_\varrho \mathcal{K}_k(J_\varrho, b) \;=\; \mathcal{K}_k(E_n, J_\varrho b) \;=\; \mathcal{K}_k(E_n, \varrho\, b + E_n b). \tag{3.2}$$

But the orthogonality of b to the space $\mathcal{K}_k(E_n, \varrho\, b + E_n b)$ means that

$$0 \;=\; b^T \left(\varrho\, E_n^{j-1} b + E_n^j b\right), \quad j = 1, \ldots, k,$$

which can be written in the equivalent form (3.1). □

THEOREM 3.2. *An $n \times n$ Jordan block J_λ with $\lambda > 0$ is ideal of degree k with the kth ideal GMRES polynomial given by*

$$q(z) = (1 - \lambda^{-1} z)^k \tag{3.3}$$

if and only if $1 \leq k < n/2$ *and* $\lambda \;\geq\; \varrho_{k,n-k}^{-1}$.

Proof. Since

$$q(J_\lambda) = (-1)^k \lambda^{-k} E_n^k,$$

each $w \in \Sigma(q(J_\lambda))$ has to be of the form

$$w = (0, \ldots, 0, b_1, \ldots, b_{n-k})^T, \tag{3.4}$$

and hence

$$q(J_\lambda) w = (-1)^k \lambda^{-k} (b_1, \ldots, b_{n-k}, 0, \ldots, 0)^T.$$

Using Lemma 2.4 and the previous observation, J_λ is ideal of degree k and q is both the worst-case and ideal GMRES polynomial if and only if there exists a unit norm vector w of the form (3.4) such that

$$q(J_\lambda) w \perp J_\lambda \mathcal{K}_k(J_\lambda, w) = \mathcal{K}_k(E_n, \lambda w + E_n w), \tag{3.5}$$

i.e.

$$\lambda\, w^T (E_n^{j-1})^T E_n^k w + w^T (E_n^j)^T E_n^k w = 0, \qquad j = 1, \ldots, k. \tag{3.6}$$

Since w has the special form (3.4), it holds that $w^T (E_n^j)^T E_n^k w \;=\; b^T E_{n-k}^{k-j} b$, where $b \;\equiv\; (b_1, \ldots, b_{n-k})^T$. Then, (3.6) is equivalent to

$$\lambda^{-1} b^T E_{n-k}^{k-j} b + b^T E_{n-k}^{k-j+1} b = 0, \qquad j = 1, \ldots, k. \tag{3.7}$$

Writing the equations (3.7) in the reverse order (for $j = k, \ldots, 1$), we obtain

$$\lambda^{-1}\, b^T E_{n-k}^{j-1} b + b^T E_{n-k}^{j} b = 0, \qquad j = 1, \ldots, k, \tag{3.8}$$

or, equivalently,

$$b^T E_{n-k}^{j} b = (-\lambda^{-1})^j, \qquad j = 1, \ldots, k. \tag{3.9}$$

Clearly, if $k \geq n/2$, then E_{n-k}^k is the zero matrix. In this case at least one of the conditions in (3.9) takes the form $0 = (-\lambda)^k$, and the system (3.9) does not have a solution for any positive λ. On the other hand, for $1 \leq k < n/2$, the system (3.9) has a solution if and only if $\lambda^{-1} \leq \varrho_{k,n-k}$, cf. Lemma 3.1, which completes the proof. □

We summarize what we have seen so far in the following corollary.

COROLLARY 3.3. *For an $n \times n$ Jordan block J_λ with eigenvalue $\lambda \in \mathbb{C}$, and $1 \leq k < n/2$ the following hold:*

1. *If $|\lambda| \leq \varrho_{k,n}$, then J_λ is ideal of degree k with $\psi_k(J_\lambda) = \varphi_k(J_\lambda) = 1$.*
2. *If $|\lambda| \geq \varrho_{k,n-k}^{-1}$, then J_λ is ideal of degree k with $\psi_k(J_\lambda) = \varphi_k(J_\lambda) = \lambda^{-k}$.*

The first item already was shown in (2.11), the second follows from Theorem 3.2. In summary, for $1 \leq k < n/2$ and $|\lambda| \geq 0$, we completely understand the situation except for the cases

$$\varrho_{k,n} \;<\; |\lambda| \;<\; \varrho_{k,n-k}^{-1}. \tag{3.10}$$

The lower bound in (3.10) is bounded from below by $1/2$, and it approaches 1 for $n \to \infty$, while the upper bound in (3.10) is bounded from above by 2.

4. Structure of the ideal GMRES residual matrices for a Jordan block.

In this section we analyze the special structure of the ideal GMRES residual matrices for a Jordan block, which we originally discovered numerically when experimenting with the semidefinite programming package SDPT3 [18]. Since the development below is quite technical, we start with a high-level description of a simple example.

Consider the 6×6 Jordan block J_λ with $\lambda = 1$. As shown below, its second, third and fourth ideal GMRES residual matrices are upper triangular Toeplitz matrices of the form

$$\underbrace{\begin{pmatrix} \bullet & \circ & \bullet & & & \\ & \bullet & \circ & \bullet & & \\ & & \bullet & \circ & \bullet & \\ & & & \bullet & \circ & \bullet \\ & & & & \bullet & \circ \\ & & & & & \bullet \end{pmatrix}}_{p_*^{(2)}(J_1)}, \quad \underbrace{\begin{pmatrix} \bullet & \circ & \circ & \bullet & & \\ & \bullet & \circ & \circ & \bullet & \\ & & \bullet & \circ & \circ & \bullet \\ & & & \bullet & \circ & \circ \\ & & & & \bullet & \circ \\ & & & & & \bullet \end{pmatrix}}_{p_*^{(3)}(J_1)}, \quad \underbrace{\begin{pmatrix} \bullet & \circ & \bullet & \circ & \bullet & \\ & \bullet & \circ & \bullet & \circ & \bullet \\ & & \bullet & \circ & \bullet & \circ \\ & & & \bullet & \circ & \bullet \\ & & & & \bullet & \circ \\ & & & & & \bullet \end{pmatrix}}_{p_*^{(4)}(J_1)},$$

where "$\bullet$" stands for a nonzero entry and "$\circ$" represents a zero entry. It is easy to see that there exist permutation matrices P_2, P_3 and P_4 that transform $p_*^{(2)}(J_1)$, $p_*^{(3)}(J_1)$ and $p_*^{(4)}(J_1)$ into block diagonal matrices with upper triangular Toeplitz blocks,

$$\underbrace{\begin{pmatrix} \bullet & \bullet & & & & \\ & \bullet & \bullet & & & \\ & & \bullet & & & \\ & & & \bullet & \bullet & \\ & & & & \bullet & \bullet \\ & & & & & \bullet \end{pmatrix}}_{P_2^T p_*^{(2)}(J_1) P_2}, \quad \underbrace{\begin{pmatrix} \bullet & \bullet & & & & \\ & \bullet & & & & \\ & & \bullet & \bullet & & \\ & & & \bullet & & \\ & & & & \bullet & \bullet \\ & & & & & \bullet \end{pmatrix}}_{P_3^T p_*^{(3)}(J_1) P_3}, \quad \underbrace{\begin{pmatrix} \bullet & \bullet & \bullet & & & \\ & \bullet & \bullet & & & \\ & & \bullet & & & \\ & & & \bullet & \bullet & \bullet \\ & & & & \bullet & \bullet \\ & & & & & \bullet \end{pmatrix}}_{P_4^T p_*^{(4)}(J_1) P_4}.$$

Since the transformation $p_*^{(k)}(J_1) \to P_k^T p_*^{(k)}(J_1) P_k$ is orthogonal, and all diagonal blocks of $P_k^T p_*^{(k)}(J_1) P_k$ are equal, the ideal GMRES approximation $\|p_*^{(k)}(J_1)\|$ equals the norm of any diagonal block of $P_k^T p_*^{(k)}(J_1) P_k$.

These observations are the key to analyzing the kth and $(n-k)$th ideal GMRES approximations for J_1 and, more generally, for any Jordan block J_λ, when k divides n. The following lemma formalizes the just described orthogonal transformation and shows the connection between the singular value decompositions of $p_*^{(k)}(J_\lambda)$ and of a diagonal block of $P_k^T p_*^{(k)}(J_\lambda) P_k$.

LEMMA 4.1. *Let n and k be positive integers, $n > k$, and let d be their greatest common divisor. Define $m \equiv n/d$ and $\ell = k/d$. Consider the $m \times m$ upper triangular Toeplitz matrix B,*

$$B \equiv \sum_{j=0}^{\ell} b_j E_m^j, \quad \textit{and let} \quad B = USV^T \tag{4.1}$$

be its singular value decomposition. Then the singular value decomposition of the $n \times n$ matrix G,

$$G \equiv \sum_{j=0}^{\ell} b_j E_n^{jd} \quad \textit{is given by} \quad G = (U \otimes I_d)(S \otimes I_d)(V \otimes I_d)^T. \tag{4.2}$$

Proof. Define the $n \times n$ matrix P by $P \equiv [I_m \otimes e_1, \ldots, I_m \otimes e_d]$, then

$$P^T G P = I_d \otimes B = I_d \otimes (USV^T) = (I_d \otimes U)(I_d \otimes S)(I_d \otimes V)^T,$$

and hence

$$\begin{aligned} G &= P(I_d \otimes U)(I_d \otimes S)(I_d \otimes V)^T P^T \\ &= [P(I_d \otimes U)P^T]\,[P(I_d \otimes S)P^T]\,[P(I_d \otimes V)P^T]^T \\ &= (U \otimes I_d)(S \otimes I_d)(V \otimes I_d)^T. \end{aligned}$$

In the last equation we have used [8, Corollary 4.3.10]. □

As outlined above, our strategy is as follows: Having an ideal GMRES residual matrix G of the special form (4.2), we can find a permutation matrix P such that $P^T G P = I \otimes B$ (where I and B have the appropriate sizes), and then investigate the norm and properties of G through the norm and properties of the block B.

THEOREM 4.2. *Let n and k be positive integers, $n > k$, and let d be their greatest common divisor. Let $\lambda > 0$ and define $m \equiv n/d$, $\ell \equiv k/d$,*

$$J_\lambda \equiv \lambda I_n + E_n\,, \qquad J_\mu \equiv \mu I_m + E_m, \qquad \mu \equiv \lambda^d.$$

Suppose that the ℓth ideal GMRES polynomial $p_*^{(\ell)}$ of J_μ is of the form

$$p_*^{(\ell)}(z) = \sum_{j=0}^{\ell} c_j (\mu - z)^j. \tag{4.3}$$

If J_μ is ideal of degree ℓ, then J_λ is ideal of degree k, and

$$\psi_\ell(J_\mu) = \varphi_\ell(J_\mu) \quad = \quad \psi_k(J_\lambda) = \varphi_k(J_\lambda).$$

Moreover, the kth ideal GMRES polynomial $p_*^{(k)}$ of J_λ is given by

$$p_*^{(k)}(z) = \sum_{j=0}^{\ell} c_j (\lambda - z)^{jd}. \tag{4.4}$$

Proof. Given the ℓth ideal GMRES polynomial $p_*^{(\ell)} \in \pi_\ell$ of J_μ as in (4.3), we define the polynomial

$$q(z) \;\equiv\; \sum_{j=0}^{\ell} c_j (\lambda - z)^{jd} \;\in\; \pi_k. \tag{4.5}$$

Our goal is to show that this polynomial q, which is equal to $p_*^{(k)}$ in (4.4), is the kth ideal GMRES polynomial of J_λ. We will show this by constructing a unit norm vector $b \in \Sigma(q(J_\lambda))$, such that the condition (2.1) is satisfied.

From

$$p_*^{(\ell)}(J_\mu) \;=\; \sum_{j=0}^{\ell} c_j (-E_m)^j, \qquad q(J_\lambda) \;=\; \sum_{j=0}^{\ell} c_j (-E_n)^{jd}, \tag{4.6}$$

we see that the matrices $p_*^{(\ell)}(J_\mu)$ and $q(J_\lambda)$ have a similar structure as the matrices B and G, respectively, in Lemma 4.1 (up to the sign in case d is even).

By assumption, $\psi_\ell(J_\mu) = \varphi_\ell(J_\mu) > 0$, and hence by Lemma 2.4 there exists a unit norm vector $w \in \Sigma(p_*^{(\ell)}(J_\mu))$, such that

$$p_*^{(\ell)}(J_\mu) w \;\perp\; J_\mu \mathcal{K}_\ell(J_\mu, w). \tag{4.7}$$

Define $S_\mu \in \mathbb{R}^{m\times m}$, $v \in \mathbb{R}^m$, and $B \in \mathbb{R}^{m\times m}$ by

$$S_\mu \;\equiv\; \left\{ \begin{array}{l} J_\mu, \\ I_m^{\pm} J_\mu I_m^{\pm}, \end{array} \right. \qquad v \equiv \left\{ \begin{array}{ll} w\,, & \text{if } d \text{ is odd,} \\ I_m^{\pm}\, w, & \text{if } d \text{ is even,} \end{array} \right. \tag{4.8}$$

$$B \;\equiv\; p_*^{(\ell)}(S_\mu). \tag{4.9}$$

Then it easily follows that

$$Bv \;\perp\; S_\mu \mathcal{K}_\ell(S_\mu, v), \tag{4.10}$$

and $v \in \Sigma(B)$. Since B is a Toeplitz matrix, the matrix $I_m^B B$ is symmetric, and hence unitarily diagonalizable, $I_m^B B = V \Lambda V^T$. Therefore, there exists a diagonal matrix $\hat{I}_m^{\pm}$ having entries 1 and -1 on its diagonal, such that

$$B \;=\; (I_m^B V \hat{I}_m^{\pm})\,(\hat{I}_m^{\pm} \Lambda)\, V^T$$

is the singular value decomposition of B. If $z \in \Sigma(B)$ is a right singular vector, then the corresponding left singular vector is given either by $I_m^B z$ or by $-I_m^B z$. Since $v \in \Sigma(B)$, we can decompose this vector as $v = v^+ + v^-$. Here v^+ resp. v^- are the orthogonal projections of v onto the space spanned by right singular vectors $z \in \Sigma(B)$ with the corresponding left singular vector equal to $I_m^B z$ resp. $-I_m^B z$.

Denoting by δ the maximal singular value of $p_*^{(\ell)}(J_\mu)$,

$$Bv = \delta\, I_m^B (v^+ - v^-), \qquad \text{and} \qquad \delta \equiv \|p_*^{(\ell)}(J_\mu)\| = \|B\| = \|q(J_\lambda)\|, \tag{4.11}$$

where we have applied Lemma 4.1 to obtain the last equality.

Since $v \in \Sigma(B)$, Lemma 4.1 implies that $v \otimes e_j \in \Sigma(q(J_\lambda))$, where e_j denotes the jth standard basis vector for $j = 1, \ldots, d$. Now define $e_\lambda \equiv [1, -\lambda, \ldots, (-\lambda)^{d-1}]^T$, and

$$b \;\equiv\; \gamma \sum_{j=1}^{d} (-\lambda)^{j-1}\, v \otimes e_j \;=\; \gamma\, (v \otimes e_\lambda), \tag{4.12}$$

where γ is chosen so that $\|b\| = 1$. Clearly, $b \in \Sigma(q(J_\lambda))$, and b can be decomposed as

$$b = b^+ + b^-, \qquad b^+ \equiv \gamma\,(v^+ \otimes e_\lambda), \qquad b^- \equiv \gamma\,(v^- \otimes e_\lambda),$$

with $q(J_\lambda)b^+ = \delta I_n^B b^+$, $q(J_\lambda)b^- = -\delta I_n^B b^-$. Hence, using the first expression in (4.11),

$$\begin{aligned} q(J_\lambda)b &= \gamma q(J_\lambda)\,(b^+ + b^-) = \gamma\delta I_n^B(b^+ - b^-) \\ &= \gamma\delta I_n^B\,((v^+ - v^-) \otimes e_\lambda) = \gamma\delta\,((I_m^B(v^+ - v^-)) \otimes (I_d^B e_\lambda)) \\ &= \gamma\,((Bv) \otimes (I_d^B e_\lambda)). \end{aligned} \tag{4.13}$$

We next show that

$$q(J_\lambda)b \;\perp\; J_\lambda^j b, \quad j = 1, \ldots, k, \tag{4.14}$$

i.e. that q is a GMRES polynomial for J_λ and the initial vector b. Since

$$\text{span}\{J_\lambda b, \ldots, J_\lambda^k b\} \;=\; \text{span}\{E_n^0 J_\lambda b, \ldots, E_n^{k-1} J_\lambda b\}, \tag{4.15}$$

the relation (4.14) holds if and only if

$$q(J_\lambda)b \;\perp\; E_n^j J_\lambda b, \quad j = 0, \ldots, k-1. \tag{4.16}$$

Let us decompose the index j as

$$j = sd + t, \qquad s = 0, \ldots, l-1, \qquad t = 0, \ldots, d-1. \tag{4.17}$$

An elementary computation shows that

$$J_\lambda b \;=\; \gamma\, J_\lambda(v \otimes e_\lambda) \;=\; \gamma\,((S_\mu v) \otimes e_d).$$

Multiplication of $J_\lambda b$ from the left by E_n^j shifts all entries of $J_\lambda b$ upwards by j positions. Using (4.17), $E_n^j J_\lambda b$ can be written as

$$E_n^j J_\lambda b \;=\; \gamma\, E_n^{sd}((S_\mu v) \otimes e_{d-t}) = \gamma\,((E_m^s S_\mu v) \otimes e_{d-t}). \tag{4.18}$$

Now from (4.13) and (4.18) we obtain

$$\begin{aligned} (q(J_\lambda)b)^T(E_n^j J_\lambda b) &= \gamma^2\,((Bv) \otimes (I_d^B e_\lambda))^T\,((E_m^s S_\mu v) \otimes e_{d-t}) \\ &= \gamma^2\,[(Bv)^T E_m^s S_\mu v]\,[e_\lambda^T I_d^B e_{d-t}]. \end{aligned}$$

Similar as in (4.15), $E_m^s S_\mu v \in S_\mu \mathcal{K}_\ell(S_\mu, v)$ for $s = 0, \ldots, l-1$. Since Bv is orthogonal to $S_\mu \mathcal{K}_\ell(S_\mu, v)$, cf. (4.10), it holds that $(Bv)^T E_m^s S_\mu v = 0$ for $s = 0, \ldots, l-1$. In other words, we just proved (4.14).

Summarizing, q is the kth GMRES polynomial for the matrix J_λ and the initial vector $b \in \Sigma(q(J_\lambda))$. Using Lemma 2.4, it holds that $\psi_k(J_\lambda) = \varphi_k(J_\lambda)$ and, therefore, q is the kth ideal GMRES polynomial of J_λ. Moreover, Lemma 4.1 implies that the ideal GMRES residual matrices $p_*^{(\ell)}(J_\mu)$ and $p_*^{(k)}(J_\lambda)$ have the same norm and thus $\psi_\ell(J_\mu) = \varphi_\ell(J_\mu) = \psi_k(J_\lambda) = \varphi_k(J_\lambda)$. □

Note that the integers ℓ and m defined in Theorem 4.2 are relatively prime. The assertion of this theorem is quite tricky, so some explanation is appropriate. Suppose we know that an $m \times m$ Jordan block J_μ is ideal of degree ℓ, where ℓ and m are relatively prime. Then by Theorem 4.2, an $n \times n$ Jordan block J_λ is ideal of degree k, where $n \equiv dm$, $k \equiv d\ell$, $\lambda \equiv \mu^d$, and d is *any* positive integer. Therefore, to prove that any Jordan block is ideal, it would be

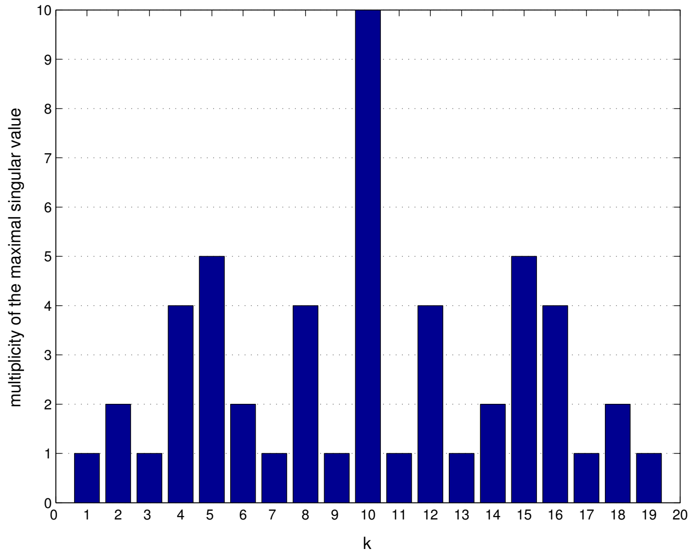


FIG. 4.1. *Multiplicity of the maximal singular value of* $p_*^{(k)}(J_1)$ *for the* 20×20 *Jordan block* J_1 *and* $k = 1, \ldots, 19$.

sufficient to show that any Jordan block is ideal of degree k whenever k and the size of the Jordan block are relatively prime; all the other cases are then covered by Theorem 4.2. In other words, Theorem 4.2 reduces the question of idealness of Jordan blocks to block sizes n and steps k, where k and n are relatively prime.

EXAMPLE 1. Consider the 20×20 Jordan block J_1. In Fig 4.1 we plot the multiplicity of the maximal singular value of $p_*^{(k)}(J_1)$ for $k = 1, \ldots, 19$. Apparently, the multiplicity is equal to the greatest common divisor of n and k. In particular, at steps k such that k and n are relatively prime, the maximal singular value of $p_*^{(k)}(J_1)$ is simple. (The same phenomenon can be observed numerically also for other choices of n.) By the second item in Lemma 2.3, J_1 is ideal of degree k in the steps where k and n are relatively prime. Then Theorem 4.2 implies that J_1 is ideal.

Theorem 4.2 also allows us to prove the following result about the radii of the polynomial numerical hulls of Jordan blocks.

THEOREM 4.3. *Let n and k be positive integers, $n > k$, and let d be their greatest common divisor. Define $m \equiv n/d$, $\ell \equiv k/d$. Then the radius $\varrho_{k,n}$ of the kth polynomial numerical hull of an $n \times n$ Jordan block satisfies*

$$\varrho_{k,n} = \varrho_{\ell,m}^{1/d}. \tag{4.19}$$

Proof. Let $\lambda > 0$ and consider Jordan blocks

$$J_\lambda \equiv \lambda I_n + E_n, \qquad J_\mu \equiv \mu I_m + E_m, \qquad \mu \equiv \lambda^d.$$

We prove the following equivalence

$$\mu \le \varrho_{\ell,m} \overset{\mathbf{A}}{\Longleftrightarrow} \psi_\ell(J_\mu) = \varphi_\ell(J_\mu) = 1 \overset{\mathbf{B}}{\Longleftrightarrow} \psi_k(J_\lambda) = \varphi_k(J_\lambda) = 1 \overset{\mathbf{C}}{\Longleftrightarrow} \lambda \le \varrho_{k,n}.$$

The equivalences **A** and **C** follow from (2.11), so we only have to prove the equivalence **B**. From Theorem 4.2,

$$\psi_\ell(J_\mu) = \varphi_\ell(J_\mu) = 1 \quad \Longrightarrow \quad \psi_k(J_\lambda) = \varphi_k(J_\lambda) = 1.$$

On the other hand, suppose that $\psi_k(J_\lambda) = \varphi_k(J_\lambda) = 1$. Consider the polynomial $p_*^{(\ell)}$ of the form (4.3). Then, similarly as in the proof of Theorem 4.2, the polynomial q defined by (4.5) satisfies $q \in \pi_k$ and $\|q(J_\lambda)\| = \|p_*^{(\ell)}(J_\mu)\|$, cf. (4.11). Now if $\varphi_k(J_\mu) = \|p_*^{(\ell)}(J_\mu)\| < 1$, then $\|q(J_\lambda)\| < 1 = \varphi_k(J_\lambda)$, which contradicts the optimality property of the kth ideal GMRES polynomial $p_*^{(k)}$ of J_λ. Therefore $\varphi_k(J_\mu) = 1$, which implies that $\psi_k(J_\mu) = 1$, cf. (2.4), and thus **B** must hold.

Consequently, for each $\lambda > 0$, $\lambda^d \le \varrho_{\ell,m} \Longleftrightarrow \lambda \le \varrho_{k,n}$, which implies (4.19). □

COROLLARY 4.4. *Consider an $n \times n$ Jordan block J_λ with $\lambda > 0$. Let $k < n$ be a positive integer dividing n. Then $\psi_k(J_\lambda) = \varphi_k(J_\lambda)$, and if $\lambda \ge \varrho_{k,n}$, then*

$$\lambda^{-k} \cos\left(\tfrac{\pi}{n/k+1}\right) \;\le\; \varphi_k(J_\lambda) \;\le\; \lambda^{-k}. \tag{4.20}$$

The kth ideal GMRES polynomial $p_*^{(k)}$ of J_λ is of the form

$$p_*^{(k)}(z) = c_0 + c_1\,(\lambda - z)^k, \tag{4.21}$$

where c_0 and c_1 are the coefficients of the first ideal GMRES polynomial (4.3) of the $\frac{n}{k} \times \frac{n}{k}$ Jordan block J_{λ^k}. Moreover, it holds that

$$\varrho_{k,n} = \varrho_{1,n/k}^{1/k} = \left[\cos\left(\tfrac{\pi}{n/k+1}\right)\right]^{1/k}. \tag{4.22}$$

Proof. All results follow from Theorem 4.2 and Theorem 4.3. If k divides n, then $d = k$ is their greatest common divisor, and $m = n/k$, $\ell = 1$. For $\ell = 1$, the assumption $\psi_\ell(J_\mu) = \varphi_\ell(J_\mu) > 0$ in Theorem 4.2 is satisfied and therefore $\psi_k(J_\lambda) = \varphi_k(J_\lambda)$. In (4.22) we used Theorem 4.3 and the explicit form of the radius $\varrho_{1,n/k}$, cf. (2.8). The bound (4.20) is just the bound (2.10), where for $\varrho_{k,n}$ we substituted its exact value on the right hand side of (4.22). □

If k divides n, then $p_*^{(k)}(z) = 1$ for $\lambda \le \varrho_{n,k}$ and $p_*^{(k)}(z) = c_0 + c_1(\lambda - z)^k$ for $\lambda > \varrho_{n,k}$. For $\lambda \ge \varrho_{k,n-k}^{-1}$ we know that $c_0 = 0$ and $c_1 = \lambda^{-k}$, cf. Theorem 3.2. Moreover, since $\psi_k(J_\lambda) = \varphi_k(J_\lambda)$, it follows from (2.11) and Theorem 3.2 that $c_0 \ne 0$ and $c_1 \ne 0$ whenever $\varrho_{k,n} < \lambda < \varrho_{k,n-k}^{-1}$. Then, from the form of the kth ideal GMRES polynomial (4.21) it is easy to see that the k roots of $p_*^{(k)}$ are uniformly distributed on the circle around λ with radius $|c_0/c_1|^{1/k}$.

EXAMPLE 2. Consider an $n \times n$ Jordan block J_λ with $\lambda > 0$, n even and $k = n/2$. This gives $d = n/2$, $m = 2$, $\ell = 1$, and $\mu = \lambda^{n/2}$ in Theorem 4.2. Since for the 2×2 Jordan block J_μ, $\psi_1(J_\mu) = \varphi_1(J_\mu) > 0$, Theorem 4.2 implies that $\psi_1(J_\mu) = \varphi_1(J_\mu) = \psi_{n/2}(J_\lambda) = \varphi_{n/2}(J_\lambda)$. Theorem 4.3 shows that

$$\varrho_{n/2,n} = \varrho_{1,2}^{1/k} = 2^{-2/n}. \tag{4.23}$$

Moreover, by a direct computation of the first ideal GMRES approximation for the 2×2 Jordan block J_μ with $\mu = \lambda^{n/2}$, we obtain that for $\lambda \ge 2^{-2/n}$,

$$c_0 = \frac{2}{4\lambda^n + 1}, \qquad c_1 = \frac{1}{\lambda^{n/2}}\,\frac{4\lambda^n - 1}{4\lambda^n + 1}, \qquad \varphi_{n/2}(J_\lambda) \;=\; \frac{4\lambda^{n/2}}{4\lambda^n + 1}. \tag{4.24}$$

Using (2.10) and the fact that $\varrho_{k,n}^{k} \geq \varrho_{n/2,n}^{k} = 2^{-2k/n} \geq 2^{-1}$ for $k \leq n/2$, we get the bound

$$\frac{1}{2}\lambda^{-k} \;\leq\; \varphi_k(J_\lambda) \;\leq\; \lambda^{-k}, \qquad k \;\leq\; n/2. \tag{4.25}$$

5. The next-to-last worst-case and ideal GMRES approximations. In this section we consider the $(n-1)$st worst-case and ideal GMRES approximations for an $n \times n$ Jordan block J_λ with $\lambda > 0$. Our main result, stated in Theorem 5.5 below, is that $\psi_{n-1}(J_\lambda) = \varphi_{n-1}(J_\lambda)$ for $\lambda \geq 1$. We also give an explicit expression for $\varphi_{n-1}(J_\lambda)$ in terms of the eigenvalue λ. The proof of this result will make use of three technical lemmas. To simplify the notation, we define the vector

$$e_\lambda^{(n)} \;\equiv\; I_n^{\pm}[1, \lambda, \ldots, \lambda^{n-1}]^T$$

and the Hankel matrix

$$H(v_1, \ldots, v_n) \;\equiv\; \begin{pmatrix} v_1 & v_2 & \cdots & v_n \\ v_2 & & \iddots & \\ \vdots & \iddots & & \\ v_n & & & \end{pmatrix}. \tag{5.1}$$

The first lemma is a slight reformulation of [12, Corollary 2.2].

LEMMA 5.1. *Consider the linear algebraic system $J_\lambda x = b$, with an $n \times n$ Jordan block J_λ, and a right hand side vector $b = [b_1, \ldots, b_n]^T$ such that $b_n \neq 0$. If $x_0 = 0$, then the $(n-1)$st GMRES residual r_{n-1} is uniquely determined by the linear system*

$$\|r_{n-1}\|^{-2}\, H(b_1, \ldots, b_n)\, r_{n-1} \;=\; e_\lambda^{(n)}. \tag{5.2}$$

LEMMA 5.2. *Let $\lambda > 0$ be given and let $b \in \mathbb{R}^n$ be the unit norm vector*

$$b \;\equiv\; (-1)^{n-1}\|\xi\|^{-1}\, I_n^{B}\xi, \tag{5.3}$$

where $\xi = [\xi_1, \ldots, \xi_n]^T$ has the components

$$\xi_{i+1} \;=\; \lambda^{\frac{n-1}{2}-i}\frac{(-1)^i}{4^i}\binom{2i}{i}, \qquad i = 0, \ldots, n-1. \tag{5.4}$$

Then the $(n-1)$st GMRES residual r_{n-1} for the $n \times n$ Jordan block J_λ and the initial vector b is given by $r_{n-1} = \|\xi\|^{-3}\,\xi$, and hence

$$\|r_{n-1}\| = \|\xi\|^{-2} = \frac{1}{\lambda^{n-1}}\left[\sum_{i=0}^{n-1}(4\lambda)^{-2i}\binom{2i}{i}^2\right]^{-1}. \tag{5.5}$$

Proof. Since the last component of $b = (-1)^{n-1}\|\xi\|^{-1}[\xi_n, \ldots, \xi_1]^T$ is nonzero, Lemma 5.1 implies that the $(n-1)$st GMRES residual for J_λ and b satisfies

$$\frac{(-1)^{n-1}}{\|\xi\|\,\|r_{n-1}\|^2}\,\widehat{H}\,r_{n-1} \;=\; e_\lambda^{(n)}, \tag{5.6}$$

where $\widehat{H} = H(\xi_n, \ldots, \xi_1)$. Using the definition (5.4), the numbers ξ_{i+1} satisfy for $j = 0, \ldots, n-1$,

$$\sum_{i=0}^{j} \xi_{i+1}\xi_{j-i+1} = \frac{(-1)^j}{4^j}\lambda^{n-j-1} \sum_{i=0}^{j} \binom{2i}{i}\binom{2(j-i)}{j-i} = (-1)^j \lambda^{n-j-1}.$$

In the last equality we use the fact that the sum of the products of the given binomial coefficients is equal to 4^j, see e.g. [15, p. 44]. The n previous equations can be written in matrix form as

$$\widehat{H}\,\xi \;=\; (-1)^{n-1} e_{\lambda}^{(n)}. \tag{5.7}$$

A comparison of (5.7) and (5.6) shows that $\xi = \|r_{n-1}\|^{-2} r_{n-1} \|\xi\|^{-1}$ and, therefore, $\|\xi\|^{-2} = \|r_{n-1}\|$. Finally, $r_{n-1} = \xi\,\|\xi\|\,\|r_{n-1}\|^2 = \xi\,\|\xi\|^{-3}$. A straightforward computation shows that $\|r_{n-1}\|$ is given by (5.5). □

REMARK 5.3. It is not hard to check that ξ_{i+1} defined in (5.4) can be computed by the recurrence

$$\xi_1 = \lambda^{\frac{n-1}{2}}, \qquad \xi_{i+1} = -\xi_i \lambda^{-1} \frac{2i-1}{2i}, \qquad i = 1, \ldots, n-1. \tag{5.8}$$

LEMMA 5.4. *Let $\lambda > 0$ be given and let $\xi^+ \equiv I_n^{\pm}\xi$, where the vector ξ is defined as in Lemma 5.2. Then there exists an uniquely determined Hankel matrix $\hat{H}$ such that*

$$\xi^+ = \hat{H}\,\xi^+. \tag{5.9}$$

If $\lambda \geq 1$, the matrix $\hat{H}$ is primitive and has only one eigenvalue of maximum modulus. This eigenvalue is equal to 1, and ξ^+ is the corresponding eigenvector.

Proof. First note that since the entries of ξ alternate in sign and $\xi_1 > 0$, all components of $\xi^+ = [\xi_1^+, \ldots, \xi_n^+]^T$ are positive. We are now going to construct the Hankel matrix $\hat{H}$ of the form $\hat{H} = H(h_n, \ldots, h_1)$.

The nth equation in $\xi^+ = \hat{H}\xi^+$ is $h_1\xi_1^+ = \xi_n^+$, i.e. $h_1 = \xi_n^+/\xi_1^+$. Therefore, h_1 is well-defined and positive. Considering the equations $n-1, \ldots, 1$ it is clear that the entries $h_2, \ldots, h_n$ of $\hat{H}$ are uniquely determined.

To show the remaining part of the lemma, we will first prove by induction that for $\lambda \geq 1$, $\hat{H}$ is nonnegative with $h_i > 0$, $i = 1, \ldots, n$. We already know that $h_1 > 0$. Now suppose that $h_1 > 0, \ldots, h_j > 0$ for some $j \geq 1$. The $(n-j)$th equation in $\xi^+ = \hat{H}\xi^+$ is of the form

$$\xi_{n-j}^+ = h_{j+1}\xi_1^+ + \sum_{i=2}^{j+1} h_{j-i+2}\xi_i^+ = h_{j+1}\xi_1^+ + \sum_{i=1}^{j} h_{j-i+1}\xi_{i+1}^+.$$

Using the definitions of ξ_{i+1}^+ and ξ_{i+1}, cf. (5.4) and (5.8), it holds that

$$\xi_{i+1}^+ = \lambda^{-1}\left(\xi_i^+ - \frac{\xi_i^+}{2i}\right)$$

and, therefore,

$$\begin{aligned}
\xi_{n-j}^+ &= h_{j+1}\xi_1^+ + \lambda^{-1}\sum_{i=1}^{j} h_{j-i+1}\xi_i^+ - \lambda^{-1}\sum_{i=1}^{j} h_{j-i+1}\frac{\xi_i^+}{2i} \\
&= h_{j+1}\xi_1^+ + \lambda^{-1}\xi_{n-j+1}^+ - \lambda^{-1}\sum_{i=1}^{j} h_{j-i+1}\frac{\xi_i^+}{2i}.
\end{aligned}$$

Finally,

$$h_{j+1} = (\xi_1^+)^{-1}\left(\xi_{n-j}^+ - \lambda^{-1}\xi_{n-j+1}^+ + \left[\lambda^{-1}\sum_{i=1}^{j} h_{j-i+1}\frac{\xi_i^+}{2i}\right]\right). \tag{5.10}$$

The term in the square brackets is positive according to the induction hypothesis. Moreover, since the sequence $\xi_1^+, \xi_2^+, \ldots$ is decreasing for $\lambda \geq 1$, it holds that $\xi_{n-j}^+ > \lambda^{-1}\xi_{n-j+1}^+$, i.e. $h_{j+1} > 0$.

Summarizing, $\hat{H}$ is nonnegative and $\xi^+ > 0$ is an eigenvector of $\hat{H}$ corresponding to the eigenvalue 1. Therefore, 1 must be an eigenvalue of maximum modulus [9, Corollary 8.1.30.]. Moreover, since $\hat{H}^2 > 0$, $\hat{H}$ is primitive, cf. [9, Theorem 8.5.2.], and there exists only one eigenvalue of maximum modulus. □

Now we can state and prove the main result of this section.

THEOREM 5.5. *Consider an $n \times n$ Jordan block J_λ with $\lambda \geq 1$. Then the unit norm vector b defined in (5.3)–(5.4) solves the worst-case GMRES approximation problem (1.3) for J_λ and $k = n-1$, and it holds that*

$$\psi_{n-1}(J_\lambda) \;=\; \varphi_{n-1}(J_\lambda) \;=\; \frac{1}{\lambda^{n-1}}\left[\sum_{i=0}^{n-1}(4\lambda)^{-2i}\binom{2i}{i}^2\right]^{-1}. \tag{5.11}$$

Proof. Consider the $(n-1)$st GMRES residual r_{n-1} for J_λ and the initial vector b defined in (5.3)–(5.4), and denote by p_{n-1} the corresponding GMRES polynomial, i.e.

$$r_{n-1} \;=\; p_{n-1}(J_\lambda)\, b. \tag{5.12}$$

Using (5.5), $\|r_{n-1}\|$ is equal to the rightmost expression in (5.11). To prove the assertion it suffices to show that b is a maximal right singular vector of the matrix $p_{n-1}(J_\lambda)$, cf. Lemma 2.4. Since $p_{n-1}(J_\lambda)$ is an upper triangular Toeplitz matrix, the matrix $p_{n-1}(J_\lambda)I_n^{B}$, where I_n^{B} is defined in (2.6), is symmetric, and hence unitarily diagonalizable. Denote its eigendecomposition by $p_{n-1}(J_\lambda)I_n^{B} = UDU^T$, where D is a nonsingular real diagonal matrix, and $U^TU = UU^T = I_n$. Given D, there exists a (uniquely determined) diagonal matrix $\hat{I}_n^{\pm}$ having entries 1 or -1 on its diagonal such that $S \equiv D\hat{I}_n^{\pm}$ is a real diagonal matrix with positive diagonal entries. Then

$$p_{n-1}(J_\lambda) \;=\; U\,(D\hat{I}_n^{\pm})\,(\hat{I}_n^{\pm}U^TI_n^{B}) \;=\; U\,S\,(\hat{I}_n^{\pm}U^TI_n^{B}), \tag{5.13}$$

and the rightmost expression is the singular value decomposition of $p_{n-1}(J_\lambda)$.

Substituting (5.3), (5.5) and (5.13) into (5.12), we obtain

$$\xi = (-1)^{n-1}\|\xi\|^2US\hat{I}_n^{\pm}U^T\xi. \tag{5.14}$$

Similarly as in Lemma 5.4, denote $\xi^+ \equiv I_n^{\pm}\xi > 0$. Multiplying both sides of (5.14) from the left by $I_n^{\pm}$ we receive

$$\begin{aligned}\xi^+ \;=\; \hat{H}\xi^+\,, \quad \hat{H} &\equiv (-1)^{n-1}\|\xi\|^2(I_n^{\pm}U)\,S\hat{I}_n^{\pm}\,(I_n^{\pm}U)^T \\ &= (-1)^{n-1}\|\xi\|^2(I_n^{\pm}p_{n-1}(J_\lambda)I_n^{B}I_n^{\pm}).\end{aligned} \tag{5.15}$$

Since $p_{n-1}(J_\lambda)$ is an upper triangular Toeplitz matrix, the expression (5.15) shows that $\hat{H}$ is a Hankel matrix. Considering the eigenvalue decomposition $\hat{H} = Q\Lambda Q^T$ it is easy to see that

$$Q = I_n^{\pm}U, \qquad \Lambda = (-1)^{n-1}\|\xi\|^2S\hat{I}_n^{\pm}. \tag{5.16}$$

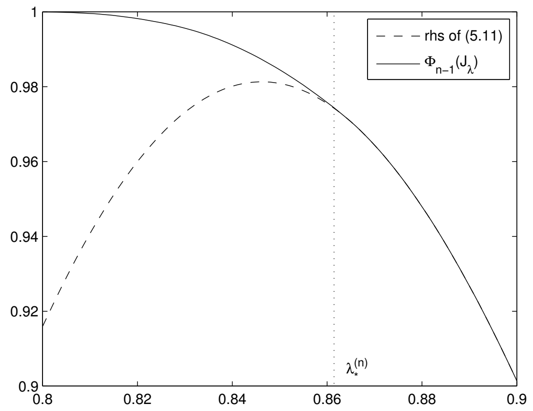


FIG. 5.1. *The right hand side of (5.11) and $\varphi_{n-1}(J_\lambda)$ plotted as a function of λ.*

Therefore, the modulus of any eigenvalue of $\hat{H}$ is a $\|\xi\|^2$-multiple of some singular value of $p_{n-1}(J_\lambda)$. Consequently, ξ^+ in (5.15) is an eigenvector corresponding to the eigenvalue of maximum modulus of $\hat{H}$ if and only if b is a right singular vector corresponding to the maximal singular value of $p_{n-1}(J_\lambda)$. By Lemma 5.4, $\hat{H}$ has only one eigenvalue of maximum modulus, and ξ^+ is the corresponding eigenvector. Hence b is the maximal right singular vector of $p_{n-1}(J_\lambda)$. □

In the previous theorem we use the assumption $\lambda \geq 1$. It is natural to ask about the relation between worst-case and ideal GMRES for $\varrho_{n-1,n} < \lambda < 1$, and whether for such λ the right hand side of (5.11) still characterizes $\psi_{n-1}(J_\lambda)$ and $\varphi_{n-1}(J_\lambda)$. While our numerical experiments predict that $\psi_{n-1}(J_\lambda) = \varphi_{n-1}(J_\lambda)$ also for $\varrho_{n-1,n} < \lambda < 1$, for each integer n there seems to exist a $\lambda_*^{(n)}$, $\varrho_{n-1,n} < \lambda_*^{(n)} < 1$, such that $\varphi_{n-1}(J_\lambda)$ is larger than the right hand side of (5.11) whenever $\lambda < \lambda_*^{(n)}$. In other words, the right hand side of (5.11) does *not* characterize $\psi_{n-1}(J_\lambda)$ and $\varphi_{n-1}(J_\lambda)$ for all $\lambda \geq \varrho_{n-1,n}$. This situation is demonstrated in Fig. 5.1, which shows a numerical experiment with $n = 10$, giving $\varrho_{n-1,n} \approx 0.8$. The dashed line shows the right hand side of (5.11), and the solid line shows the ideal GMRES approximation $\varphi_{n-1}(J_\lambda)$, both as a function of λ.

In the following corollary we combine results of Theorem 4.2, Theorem 4.3 and Theorem 5.5.

COROLLARY 5.6. *Consider an $n \times n$ Jordan block J_λ with $\lambda \geq 1$. If $k < n$ is a positive integer dividing n, then*

$$\psi_{n-k}(J_\lambda) = \varphi_{n-k}(J_\lambda) = \frac{1}{\lambda^{n-k}} \left[\sum_{i=0}^{n/k-1} \lambda^{-2ki} 4^{-2i} \binom{2i}{i}^2 \right]^{-1} . \tag{5.17}$$

and

$$\varrho_{n-k,n} = \varrho_{n/k-1,n/k}^{1/k}. \tag{5.18}$$

Proof. The parameters in Theorem 4.2 and Theorem 4.3 are given by $d = k$, $m = n/k$, $\ell = m - 1$ and $\mu = \lambda^k$. Applying Theorem 5.5 to the $m \times m$ Jordan block J_μ we see that

$\varphi_{m-1}(J_\mu) = \psi_{m-1}(J_\mu)$, and this quantity is positive. Hence the assumption of Theorem 4.2 is satisfied. Therefore, $\varphi_{m-1}(J_\mu) = \varphi_{n-k}(J_\lambda) = \psi_{n-k}(J_\lambda)$. The value of $\varphi_{m-1}(J_\mu)$ (and also of $\varphi_{n-k}(J_\lambda)$) is given by (5.11), where n and λ have to be replaced by m and λ^k, respectively. □

For example, if $n \geq 4$ is even and $k = 2$, then $m = n/2$ and (5.18) means that $\varrho_{n-2,n} = \varrho_{m-1,m}^{1/2}$. Using a completely different and highly nontrivial proof technique based on complex analysis, the same result has been obtained in [2, p. 241]. Tight bounds on $\varrho_{m-1,m}$ are given by (2.9). Note that for n even and $k = n/2$, it can be easily checked that the rightmost expression in (5.17) agrees with the rightmost expression in (4.24).

6. Polynomial numerical hulls and the ideal GMRES convergence. In [4, Section 3], some numerical examples with nonnormal matrices A of (small) size n are given, for which

$$\varphi_{n-1}(A) \leq C \min_{p\in\pi_{n-1}} \max_{z\in\mathcal{H}_{n-1}(A)} |p(z)|,$$

where C is a moderate size constant. It is not shown, however, whether the constant depends on n, or how close the bound (1.6) may be for a general nonnormal matrix A. As we are unaware of any such results in the literature, we here study this question using our above results for an $n \times n$ Jordan block J_λ. We concentrate on the case $\lambda = 1$. We need the following lemma, which can be proven by a straightforward computation; see also [16].

LEMMA 6.1. *The singular value decomposition of the $n \times n$ Jordan block J_1 is given by $J_1 = USV^T$, where*

$$V = \{v_{ij}\}_{i,j=1}^n, \qquad v_{ij} = \tfrac{2}{\sqrt{2n+1}} \sin\left(\tfrac{2i-1}{2n+1} j\pi\right), \tag{6.1}$$

$$U = \{u_{ij}\}_{i,j=1}^n, \qquad u_{ij} = \tfrac{2}{\sqrt{2n+1}} \sin\left(\tfrac{2i}{2n+1} j\pi\right), \tag{6.2}$$

$$S = \mathrm{diag}(\sigma_i), \qquad \sigma_i = 2\cos\left(\tfrac{i\pi}{2n+1}\right), \quad i = 1,\ldots,n. \tag{6.3}$$

THEOREM 6.2. *Consider the $n \times n$ Jordan block J_1, and let $k < n$ be a positive integer dividing n. Then the ideal GMRES approximations $\varphi_k(J_1)$ and $\varphi_{n-k}(J_1)$ are bounded by*

$$\cos\left(\tfrac{\pi}{2n/k}\right) \leq \varphi_k(J_1) \leq \cos\left(\tfrac{\pi}{2n/k+1}\right), \tag{6.4}$$

$$\left[1 + \tfrac{1}{2}\log(n/k)\right]^{-1} \leq \varphi_{n-k}(J_1) \leq \left[1 + \tfrac{1}{4}\log(n/k)\right]^{-1}. \tag{6.5}$$

Proof. We first prove (6.4). In the notation of Theorem 4.2, $m \equiv n/k$ and $\ell = 1$. Denote by J the $m \times m$ Jordan block with the eigenvalue one. Since $\psi_1(J) = \varphi_1(J) > 0$, Theorem 4.2 implies that $\varphi_k(J_1) = \varphi_1(J)$. It therefore suffices to bound $\|p_*^{(1)}(J)\|$.

The upper bound in (6.4) follows from

$$\|p_*^{(1)}(J)\| \leq \|I - \frac{1}{2}J\| = \frac{1}{2}\|J\| = \cos\left(\tfrac{\pi}{2m+1}\right),$$

where $\|J\| = \sigma_1(J)$ is known, cf. Lemma 6.1. For $\omega \in \mathbb{R}$, define the polynomial

$$p_\omega(z) \equiv 1 - \omega z.$$

The norm of $p_\omega(J)$ is the square root of the maximal eigenvalue of

$$p_\omega(J)^T p_\omega(J) = \begin{pmatrix} \gamma_\omega & -\beta_\omega & & \\ -\beta_\omega & \alpha_\omega & \ddots & \\ & \ddots & \ddots & -\beta_\omega \\ & & -\beta_\omega & \alpha_\omega \end{pmatrix},$$

where $\alpha_\omega \equiv \omega^2 + (1-\omega)^2$, $\beta_\omega \equiv (1-\omega)\,\omega$, $\gamma_\omega \equiv (1-\omega)^2$. Next, define the $m \times m$ matrix $T_{\omega,m} \equiv \text{tridiag}(-\beta_\omega, \alpha_\omega, -\beta_\omega)$. Denote the characteristic polynomials of $p_\omega(J)^T p_\omega(J)$ and $T_{\omega,m}$ by $\eta_{\omega,m}(z) \equiv \det(zI_m - p_\omega(J)^T p_\omega(J))$ and $\tau_{\omega,m}(z) \equiv \det(zI_m - T_{\omega,n})$, respectively. It is not hard to see that

$$\eta_{\omega,m}(z) = \tau_{\omega,m}(z) + \omega^2 \tau_{\omega,m-1}(z).$$

Using results of classical polynomial theory, the roots of the polynomials $\tau_{\omega,m}$ and $\tau_{\omega,m-1}$ interlace. Therefore, the maximal root of $\eta_{\omega,m}$ (equal to $\|p_\omega(J)\|^2$) must lay between the maximal roots of $\tau_{\omega,m}$ and $\tau_{\omega,m-1}$ (between the maximal eigenvalues of $T_{\omega,m}$ and $T_{\omega,m-1}$). It is well known that the eigenvalues of $T_{\omega,m-1}$ are given by

$$\lambda^{(j)}_{\omega,m-1} = \alpha_\omega - 2\beta_\omega \cos\left(\tfrac{j\pi}{m}\right), \quad j = 1, \ldots, m-1.$$

Considering these eigenvalues as a function of ω, and taking derivatives with respect to ω, shows that the minimum is obtained for $\omega = 1/2$. Therefore,

$$\|p_\omega(J)\|^2 \;\geq\; \max_j \lambda^{(j)}_{\frac{1}{2},m-1} = \frac{1}{2} + \frac{1}{2}\cos\left(\tfrac{\pi}{m}\right) = \cos^2\left(\tfrac{\pi}{2m}\right).$$

Taking square roots, we obtain the lower bound in (6.4).

We next prove (6.5). Using (5.17), the value of $\varphi_{n-k}(J_1)$ is given by

$$\varphi_{n-k}(J_1) = \left[\sum_{i=0}^{m-1} \vartheta_{i+1}\right]^{-1}, \quad \vartheta_{i+1} \equiv \frac{1}{4^{2i}}\binom{2i}{i}^2. \tag{6.6}$$

We first prove that for $j \geq 2$ it holds that

$$\frac{1}{4(j-1)} \leq \vartheta_j \leq \frac{1}{2j}. \tag{6.7}$$

For $j = 2$, $\vartheta_2 = \frac{1}{4}$ and (6.7) holds. Suppose that (6.7) is satisfied for some $j \geq 2$. We show that this inequality holds also for $j+1$. For ϑ_{j+1} we obtain

$$\begin{aligned} \vartheta_{j+1} &= \left(1 - \frac{1}{2j}\right)^2 \vartheta_j \leq \frac{1}{2j}\left(1 - \frac{1}{2j}\right)^2 \frac{j+1}{j+1} \\ &= \frac{1}{2(j+1)}\left(1 - \frac{3}{4j^2} + \frac{1}{4j^3}\right) \leq \frac{1}{2(j+1)}. \end{aligned}$$

Similarly,

$$\vartheta_{j+1} \geq \frac{1}{4(j-1)}\left(1 - \frac{1}{2j}\right)^2 \frac{4j}{4j} = \frac{1}{4j}\left(1 + \frac{1}{4j^2} + \frac{1}{4j^2(j-1)}\right) \geq \frac{1}{4j},$$

and (6.7) holds. Now, we can find upper and lower bounds on $\varphi_{n-k}(J_\lambda)$,

$$\sum_{i=0}^{m-1} \vartheta_{i+1} = 1 + \sum_{j=2}^{m} \vartheta_j \leq 1 + \frac{1}{2}\sum_{j=2}^{m}\frac{1}{j} \leq 1 + \frac{1}{2}\int_1^m x^{-1}\,dx = 1 + \frac{1}{2}\log(m),$$
$$\sum_{i=0}^{m-1} \vartheta_{i+1} = 1 + \sum_{j=2}^{m} \vartheta_j \geq 1 + \frac{1}{4}\sum_{j=2}^{m}\frac{1}{j-1} \geq 1 + \frac{1}{4}\int_1^m x^{-1}\,dx = 1 + \frac{1}{4}\log(m).$$

Using these inequalities and (6.6) we obtain (6.5). □

For simplicity, let us assume that n is even. The bounds (6.4) and (6.5) predict that the convergence of ideal GMRES for J_1 has two phases:

$$\varphi_k(J_1) \;\sim\; \cos\left(\tfrac{\pi}{2n/k+1}\right), \qquad \text{for } k \leq n/2,\ k \text{ divides } n, \tag{6.8}$$

$$\varphi_{n-k}(J_1) \;\sim\; \left[1+\log(n/k)\right]^{-1}, \qquad \text{for } n-k > n/2,\ k \text{ divides } n. \tag{6.9}$$

The convergence bound based on the polynomial numerical hull, i.e. (1.6), which is the lower bound in (2.10) in case of a Jordan block, is $\varphi_k(J_1) \geq \varrho_{k,n}^k$. For k dividing n we know $\varrho_{k,n}$ explicitly, and this lower bound can be evaluated, cf. (4.20). For other k one can use the explicit value of $\varrho_{n/2,n}$ resp. the lower bound on $\varrho_{n-1,n}$, cf. (4.25) resp. [5, p. 88], giving

$$\frac{1}{2} \;\leq\; 2^{-2k/n} \leq \varphi_k(J_1), \qquad \text{for } k = 1,\dots,n/2, \tag{6.10}$$

$$\left[1 - \tfrac{\log(2n)}{n}\right]^k \leq \varphi_k(J_1), \qquad \text{for } k = n/2+1,\dots,n-1. \tag{6.11}$$

Comparing (6.10) and (6.8) shows that the lower bound in (6.10) is a tight approximation of the actual ideal GMRES approximations. Hence the polynomial numerical hull of J_1 gives good information about the first phase of the ideal GMRES convergence. However, the information is less reliable in the second phase. In particular, consider the ideal GMRES approximation for $n-1$. Then (6.9) shows that

$$\varphi_{n-1}(J_1) \;\sim\; \left[1+\log n\right]^{-1},$$

while the lower bound (6.11) yields

$$\left[1 - \tfrac{\log(2n)}{n}\right]^{n-1} \;\leq\; \varphi_{n-1}(J_1).$$

A real analysis exercise shows that

$$\lim_{n\to\infty} 2n\left[1 - \tfrac{\log(2n)}{n}\right]^{n-1} \;=\; 1.$$

Hence for large n and $k = n-1$, the value on the right hand side of the lower bound (6.11) is of order $\mathcal{O}(1/n)$, while the actual ideal GMRES approximation $\varphi_{n-1}(J_1)$ is of order $\mathcal{O}(1/\log(n))$. Note that since

$$\lim_{n\to\infty} \tfrac{2n}{\log(n)}\left[1 - \tfrac{\log(2n)}{n} + \tfrac{\log(\log(2n))}{n}\right]^{n-1} \;=\; 1,$$

an approximation of $\varphi_{n-1}(J_1)$ based on the upper bound on $\varrho_{n-1,n}$, cf. (2.9), also would fail to predict the correct order of magnitude of the ideal GMRES approximation.

As shown by this example, the bound (1.6) on the kth ideal GMRES approximation for a general nonnormal matrix A based on the polynomial numerical hull of A of degree k, cannot be expected to be tight for all k.

7. Concluding remarks. Motivated by the (in general) open question of how to characterize the convergence of the GMRES method in the nonnormal case, we have studied the behavior of worst-case and ideal GMRES for an $n \times n$ Jordan block J_λ with eigenvalue $\lambda \in \mathbb{C}$. We conjecture that any such J_λ is ideal. We have shown in this paper that J_λ is ideal of degree k if any of the following conditions is satisfied:

1. $|\lambda| \leq \varrho_{k,n}$,
2. k divides n,
3. $k < n/2$ and $|\lambda| \geq \varrho_{k,n-k}^{-1}$,
4. $k \geq n/2$, $n - k$ divides n and $|\lambda| \geq 1$.

Apart from studying the idealness of J_λ, we have extended the results of [2, 5] by proving new results about the radii of the polynomial numerical hulls of Jordan blocks. Using these, we discussed the closeness of (1.6), i.e. the lower bound on ideal GMRES based on polynomial numerical hull.

Acknowledgments. We thank Anne Greenbaum for many helpful comments, and in particular for pointing out that Theorem 4.2 can be used for investigating the radii of polynomial numerical hulls of Jordan blocks. We also thank Jurjen Duintjer Tebbens for helpful comments and suggestions.

REFERENCES

[1] R. A. DeVore and G. G. Lorentz, *Constructive Approximation*, vol. 303 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Springer-Verlag, Berlin, 1993.
[2] V. Faber, A. Greenbaum, and D. E. Marshall, *The polynomial numerical hulls of Jordan blocks and related matrices*, Linear Algebra Appl., 374 (2003), pp. 231–246.
[3] V. Faber, W. Joubert, E. Knill, and T. Manteuffel, *Minimal residual method stronger than polynomial preconditioning*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 707–729.
[4] A. Greenbaum, *Generalizations of the field of values useful in the study of polynomial functions of a matrix*, Linear Algebra Appl., 347 (2002), pp. 233–249.
[5] ———, *Some theoretical results derived from polynomial numerical hulls of Jordan blocks*, Electron. Trans. Numer. Anal., 18 (2004), pp. 81–90. http://etna.math.kent.edu/vol.18.2004/pp81-90.dir/pp81-90.html.
[6] A. Greenbaum and L. Gurvits, *Max-min properties of matrix factor norms*, SIAM J. Sci. Comput., 15 (1994), pp. 348–358.
[7] A. Greenbaum and L. N. Trefethen, *GMRES/CR and Arnoldi/Lanczos as matrixapproximation problems*, SIAM J. Sci. Comput., 15 (1994), pp. 359–368.
[8] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, 1994. Corrected reprint of the 1991 original.
[9] ———, *Matrix Analysis*, Cambridge University Press, Cambridge, 1999. Corrected reprint of the 1985 original.
[10] W. Joubert, *On the convergence behavior of the restarted GMRES algorithm for solving nonsymmetric linear systems*, Numer. Linear Algebra Appl., 1 (1994), pp. 427–447.
[11] ———, *A robust GMRES-based adaptive polynomial preconditioning algorithm for nonsymmetric linear systems*, SIAM J. Sci. Comput., 15 (1994), pp. 427–439.
[12] J. Liesen and Z. Strakoš, *Convergence of GMRES for tridiagonal Toeplitz matrices*, SIAM J. Matrix Anal. Appl., 26 (2004), pp. 233–251 (electronic).
[13] O. Nevanlinna, *Convergence of Iterations for Linear Equations*, Lectures in Mathematics ETH Zürich, Birkhäuser Verlag, Basel, 1993.
[14] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetriclinear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.
[15] R. P. Stanley, *Enumerative Combinatorics. Vol. 1*, vol. 49 of Cambridge Studies in Advanced Mathematics, Cambridge University Press, Cambridge, 1997. Corrected reprint of the 1986 original.
[16] P. Tichý and J. Liesen, *Worst-case and ideal GMRES for a Jordan block*, Preprint 19-2005, Institute of Mathematics, Technical University of Berlin, 2005.
[17] K. C. Toh, *GMRES vs. ideal GMRES*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 30–36.
[18] K. C. Toh, M. J. Todd, and R. H. Tütüncü, *SDPT3 – a Matlab software package for semidefinite programming, version 2.3. Interior point methods.* June 2001.