# BEHAVIOR OF CG AND MINRES FOR SYMMETRIC TRIDIAGONAL TOEPLITZ 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 investigate the convergence behavior of the conjugate gradient (CG) method and the minimal residual (MINRES) method when applied to a linear algebraic system with a symmetric definite tridiagonal Toeplitz coefficient matrix $A$. Our main interest is to understand the behavior of the two methods for different right hand sides (initial residuals): The first one leads to the worst-case convergence quantity (relative $A$-norm of the error for CG, relative Euclidean residual norm for MINRES) in the next-to-last iteration step, and the second one has the property that its coordinates in the eigenvectors of $A$ are not biased towards a certain eigenvector direction (in other words, all these coordinates are of approximately equal size). We compare the results obtained for these right-hand sides with the classical convergence bound based on the condition number of $A$, and show when and why this bound is reasonably tight. For application of our results we choose the model problem of the one-dimensional Poisson equation with Dirichlet boundary conditions. For this problem we identify the data (source term and boundary conditions) that lead to the worst convergence quantities in the next-to-last steps of CG as well as MINRES when applied to the discretized problem. We also relate our results to previous work on the same model problem (particularly by Naiman, Babuška and Elman [14]).

**Key words.** Krylov subspace methods, CG, MINRES, convergence analysis, Toeplitz matrices, Poisson equation.

**AMS subject classifications.** 15A09, 65F10, 65F20.

**1. Introduction.** Among the abundance of Krylov subspace methods for solving symmetric definite linear algebraic systems of the form $Ax = b$, see, e.g., [1] or [4] for systematic classifications, the conjugate gradient (CG) method [10] and the minimal residual (MINRES) method [16] have emerged as de facto standard methods. Mathematically, these two methods are characterized by closely related minimization principles. Starting from an initial guess $x_0$, both methods compute the initial residual $r_0 = b - Ax_0$ and a sequence of iterates, $x_1, x_2, \ldots$, such that the $i$th residual $r_i = b - Ax_i$ is of the form

$$r_i \;=\; p_i(A)r_0\,, \quad p_i \in \pi_i\,,$$

where $\pi_i$ denotes the set of polynomials of degree at most $i$ and with value one at the origin. For CG, the polynomial is chosen so that the error $e_i = x - x_i = A^{-1}r_i$ is minimized in the $A$-norm ($\|y\|_A = (y^TAy)^{1/2}$),

$$\begin{aligned} \|e_i\|_A \;&=\; \min_{p\in\pi_i} \|p(A)e_0\|_A \\ \Rightarrow\quad \|e_i\|_A \,/\, \|e_0\|_A \;&\leq\; \min_{p\in\pi_i} \max_{1\leq k\leq n} |p(\lambda_k)| \qquad (\text{for CG})\,, \end{aligned} \tag{1.1}$$

while for MINRES the polynomial is chosen so that the residual $r_i$ is minimized in the Euclidean norm ($\|y\| = (y^Ty)^{1/2}$),

$$\|r_i\| \;=\; \min_{p\in\pi_i} \|p(A)r_0\|$$

†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.

$$\Rightarrow \quad \|r_i\| \,/\, \|r_0\| \;\leq\; \min_{p\in\pi_i} \max_{1\leq k\leq n} |p(\lambda_k)| \qquad \text{(for MINRES)}\,. \tag{1.2}$$

Here $\lambda_1, \ldots, \lambda_n$ denote the (distinct) eigenvalues of the symmetric definite matrix $A$.

Apparently, the two upper bounds (1.1) and (1.2) are independent of the initial residual, and hence they bound the *worst-case* relative error and residual norms, respectively, of the two methods. Both bounds have been shown to be sharp in the sense that for each iteration step $i$ there exists an $e_0^{(i)}$, respectively $r_0^{(i)}$, for which equalities hold (see [6] and [8, 11] for sharpness of (1.1) and (1.2), respectively). Hence the bounds actually *characterize* the two methods' worst-case behavior in terms of a polynomial min-max approximation problem on the matrix eigenvalues.

Several questions immediately arise: First, one wants to understand the relation between the eigenvalue distribution and the methods' worst-case behavior. While the two bounds yield some general intuition, cf. e.g. [7, Chapter 3], a systematic study of this problem is given in our previous paper [12]. There we characterize the min-max approximation problem in terms of explicit formulas involving the matrix eigenvalues. The second question is how much the worst-case and the "unbiased" behavior of each method differ from each other. By unbiased behavior we mean that each method is started with an initial residual having components in the matrix eigenvectors of (approximately) equal size (they are not biased towards a certain eigenvector direction). Our results in [12] allow to study this question, and this paper is the first application in this direction. Third, it is important to realize that the worst-case initial data for the system $Ax = b$ (i.e. the right hand side $b$ and the initial guess $x_0$), for which equality in (1.1) and (1.2) holds, are in general *different* for CG and MINRES. In the practical situation of a discretized (partial) differential equation, this means that a certain source term and/or boundary condition may lead to the worst-case behavior of CG, but not of MINRES, and vice versa. While this appears to be an easy observation, we are not aware that this difference has been systematically analyzed before.

Here we present an analysis for a particular problem class. We focus on linear algebraic systems where the coefficient matrix $A$ is a symmetric (positive) definite tridiagonal Toeplitz matrix. In our analysis we first consider the next-to-last iteration step of CG/MINRES. We use our results from [12] to estimate the convergence quantities in this step, and to characterize the initial error/residual for which the worst-case behavior is achieved. We also study how close this worst case is to the unbiased case. Additionally, we present a comparison of our results with the classical convergence bound based on the condition number of $A$. We apply our general results to the one-dimensional Poisson equation with Dirichlet boundary conditions. This model problem is frequently used for analyzing the behavior of Krylov subspace methods, particularly of CG, see, e.g., [2, 3, 14, 15]. Here we show connections between the worst-case behavior of CG and MINRES and the source terms and/or boundary conditions of the differential equation.

The paper is organized as follows. In Section 2 we collect the basic tools needed in our analysis. Section 3 studies the worst-case and the unbiased behavior of CG and MINRES for our problem class. Section 4 then applies our results to the Poisson equation model problem, giving analytical and numerical illustrations. We briefly summarize our results in Section 5. The Appendix lists all trigonometric formulas used in the proofs.

## 2. Tools.

We consider linear algebraic systems of the form

$$Ax \;=\; b\,, \tag{2.1}$$

where

$$A = \begin{bmatrix} \alpha & \beta & & \\ \beta & \ddots & \ddots & \\ & \ddots & \ddots & \beta \\ & & \beta & \alpha \end{bmatrix} \in \mathbb{R}^{n\times n} \tag{2.2}$$

satisfies

$$-\frac{\alpha}{2\beta} \equiv 1+\delta\,, \quad \text{for some} \quad \delta \geq 0\,. \tag{2.3}$$

Note that $\delta$ represents a measure for the diagonal dominance of $A$. We point out that all results in this paper also hold when our assumption (2.3), meaning that $-\alpha/(2\beta) \geq 1$, is replaced by $\alpha/(2\beta) \geq 1$. We also remark that the tridiagonal structure of $A$ is not important for deriving any of our results. In fact, all results hold for matrices that are unitarily similar to (2.2).

By $A = Q\Lambda Q^T$ we denote the eigendecomposition of $A$, where $Q^TQ = I$, and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. The eigenvalues $\lambda_k$ and the normalized eigenvectors $q_k$ of $A$ are given by

$$\lambda_k = \alpha + 2\,\beta\,\omega_k = -2\beta\,(1+\delta-\omega_k)\,, \tag{2.4}$$

$$q_k = (2h)^{1/2}\ [\sin(k\pi h)\,, \sin(2k\pi h)\,, \ldots, \sin(nk\pi h)]^T \tag{2.5}$$

where

$$\omega_k = \cos(k\pi h)\,, \quad h = (n+1)^{-1}\,, \quad k = 1, \ldots, n\,, \tag{2.6}$$

cf., e.g., [18, pp. 113–115]. Because of (2.3), all $n$ eigenvalues of $A$ are positive and distinct.

**2.1. Results for the MINRES residuals.** Suppose that we solve (2.1)–(2.3) with MINRES [16]. Since MINRES is mathematically equivalent to GMRES [17], we can apply our results from [12]. We define an $i$th worst-case MINRES residual $r_i^w$ as a MINRES residual for which the relative residual norm

$$\min_{p\in\pi_i} \frac{\|p(A)\,r_0\|}{\|r_0\|}\,, \qquad i = 1, \ldots, n-1\,, \tag{2.7}$$

is maximized over all $r_0 \neq 0$ (cf. [12, Definition 3.1]). Note that $r_n^w = 0$ due to the finite termination property of MINRES. Because of sharpness, $r_i^w$ is an $i$th MINRES residual for which the upper bound (1.2) is attained.

Of particular interest is the $(n-1)$st worst-case residual $r_{n-1}^w$, i.e. the situation in which MINRES is started with an initial residual $r_0^{(n-1)}$ that leads to the least progress (measured by the relative residual norm) by the next-to-last step. Since the MINRES residual norms are nonincreasing, $\|r_{n-1}^w\|/\|r_0^{(n-1)}\|$ provides a lower bound for the relative worst-case residual norms $\|r_i^w\|/\|r_0^{(i)}\|$ in every step $i = 1, \ldots n-2$. Thus a "large" $\|r_{n-1}^w\|/\|r_0^{(n-1)}\|$ implies slow convergence of the worst-case MINRES throughout the iteration. As shown in [12, Theorem 3.1], the initial residual $r_0^{(n-1)}$ leads to an $(n-1)$st worst-case MINRES residual $r_{n-1}^w$ if and only if

$$r_0^{(n-1)} = Q\,[\varrho_1^{(n-1)}, \ldots, \varrho_n^{(n-1)}]^T\,, \qquad |\varrho_k^{(n-1)}|^2 = \gamma\,L_k\,, \tag{2.8}$$

for $k = 1, \ldots, n$, and any scaling factor $\gamma > 0$, where

$$L_k = \prod_{\substack{j=1\\ j\neq k}}^{n} \frac{|\lambda_j|}{|\lambda_j - \lambda_k|}\,. \tag{2.9}$$

Moreover, the relative Euclidean norm of an $(n-1)$st worst-case MINRES residual $r^w_{n-1}$ corresponding to $r_0^{(n-1)}$ in (2.8)–(2.9) is given by

$$\frac{\|r^w_{n-1}\|}{\|r_0^{(n-1)}\|} \;=\; \left(\sum_{k=1}^{n} L_k\right)^{-1} . \tag{2.10}$$

Because of the orthogonality of the eigenvectors of $A$, we consider an initial residual with (approximately) equal components in all eigenvectors as *unbiased*. Note that this definition only depends on the given data and is independent of the solution method. Here we consider, for simplicity, the unbiased initial residual $r_0^u$ given by

$$r_0^u \;=\; Q[\varrho_1^u, \ldots, \varrho_n^u]^T , \quad \text{with} \quad \varrho_k^u = 1 , \quad k = 1, \ldots, n . \tag{2.11}$$

Obviously, $\|r_0^u\| = n^{1/2}$. When we start MINRES with $r_0^u$, the corresponding $(n-1)$st residual $r^u_{n-1}$ satisfies

$$\frac{\|r^u_{n-1}\|}{\|r_0^u\|} \;=\; \left(n \sum_{k=1}^{n} L_k^2\right)^{-1/2} , \tag{2.12}$$

see [12, Theorem 2.1].

**2.2. Analogous results for the CG errors.** Similar as for MINRES, we define an $i$th worst-case CG error $e_i^w$ as a CG error for which the relative $A$-norm of the error

$$\min_{p \in \pi_i} \frac{\|p(A)\, e_0\|_A}{\|e_0\|_A} , \qquad i = 1, \ldots, n-1 \tag{2.13}$$

is maximized over all $e_0 \neq 0$. Because of sharpness, $e_i^w$ is an $i$th CG error for which the upper bound (1.1) is attained, and $e_n^w = 0$ due to the finite termination property of CG. Note that for all $i = 0, 1, \ldots, n$,

$$\min_{p \in \pi_i} \frac{\|p(A) e_0\|_A}{\|e_0\|_A} \;=\; \min_{p \in \pi_i} \frac{\|p(A) A^{1/2} e_0\|}{\|A^{1/2} e_0\|} . \tag{2.14}$$

In other words, the relative $A$-norm of the error for CG with initial error $e_0$ is equal to the relative Euclidean residual norm for MINRES with initial residual $A^{1/2} e_0$. Thus, for $i = n-1$, the maximum on the left hand side of (2.14) is attained for the initial error $e_0^{(n-1)}$ if and only if

$$e_0^{(n-1)} \;=\; Q\,[\xi_1^{(n-1)}, \ldots, \xi_n^{(n-1)}]^T , \qquad |\xi_k^{(n-1)}|^2 = \gamma\, \lambda_k^{-1} L_k , \tag{2.15}$$

for $k = 1, \ldots, n$, and any scaling factor $\gamma > 0$, cf. (2.7)–(2.9). Moreover, the relative $A$-norm of an $(n-1)$st worst-case CG error $e^w_{n-1}$ is given by

$$\frac{\|e^w_{n-1}\|_A}{\|e_0^{(n-1)}\|_A} \;=\; \left(\sum_{k=1}^{n} L_k\right)^{-1} . \tag{2.16}$$

This obviously coincides with (2.10), which is no surprise since the right hand sides of (1.1) and (1.2) are the same, and both bounds are sharp.

The initial error corresponding to the unbiased initial residual $r_0^u$ is given by $e_0^u = A^{-1} r_0^u$, and thus has eigenvector coordinates

$$\xi_k^u \;=\; \lambda_k^{-1} , \qquad k = 1, \ldots, n . \tag{2.17}$$

The vector $e_0^u$ 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 of MINRES and CG for the same initial residual. Furthermore, this choice allows an interesting comparison with the results of [14] (see Section 4 below). Based on [12, Theorem 2.1], and using the relation between CG and MINRES, cf. (2.14), we see that

$$\begin{aligned} \frac{\|e_{n-1}^u\|_A}{\|e_0^u\|_A} &= \left(\sum_{k=1}^n \left(\frac{L_k}{\lambda_k^{1/2}\xi_k^u}\right)^2\right)^{-1/2} \left(\sum_{k=1}^n \left(\lambda_k^{1/2}\xi_k^u\right)^2\right)^{-1/2} \\ &= \left(\sum_{k=1}^n \lambda_k L_k^2\right)^{-1/2} \left(\sum_{k=1}^n \frac{1}{\lambda_k}\right)^{-1/2} . \end{aligned} \tag{2.18}$$

**2.3. Connection with Chebyshev polynomials of the second kind.** The relation of the eigenvalues of $A$, cf. (2.4) and (2.6), to the roots of the $n$th 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 $\omega_k$ in (2.6). 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

$$\begin{aligned} \prod_{k=1}^n \lambda_k &= \prod_{k=1}^n (\alpha + 2\,\beta\,\omega_k) = (-1)^n 2^n \beta^n \prod_{k=1}^n \left(-\frac{\alpha}{2\,\beta} - \omega_k\right) \\ &= (-1)^n \beta^n U_n(1+\delta)\,, \end{aligned} \tag{2.19}$$

cf. (2.3) for the definition of $\delta$.

Below we study how much the behavior of CG and MINRES change with changing $\delta$. 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

$$U_n(z) = \frac{1}{2} \frac{(z + \sqrt{z^2-1})^{n+1} - (z - \sqrt{z^2-1})^{n+1}}{\sqrt{z^2-1}}\,, \tag{2.20}$$

see, e.g., [13, 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 (2.20), $|U_n(z)|$ grows exponentially outside $(-1, 1)$. This is illustrated in Fig. 2.1, where we plot $U_n(z)/(n+1)$ for $n = 4, 6, 10$.

It is also of interest to 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 (2.3) and (2.4),

$$1 + \delta = \omega_1 \frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1} = \omega_1 \frac{\kappa(A) + 1}{\kappa(A) - 1} \equiv \omega_1\,\tau\,.$$

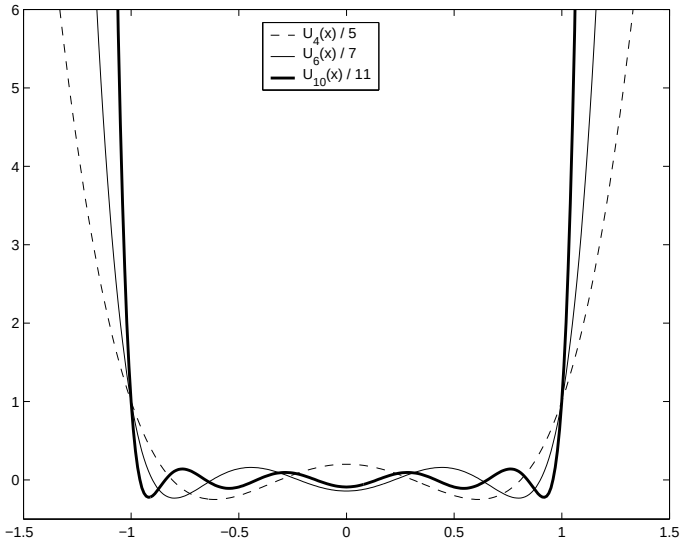


FIG. 2.1. $U_n(z)/(n+1)$ *for different* $n$.

Next,

$$\tau - \sqrt{\tau^2 - 1} \;=\; \frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \;\equiv\; \nu\,, \qquad \tau + \sqrt{\tau^2 - 1} \;=\; \nu^{-1}\,, \tag{2.21}$$

which, inserted into (2.20), yields

$$U_n(\tau) \;=\; \frac{\nu^{n+1} - \nu^{-(n+1)}}{\nu - \nu^{-1}}\,. \tag{2.22}$$

Since $U_n(z)$ is strictly monotonically increasing for $z \geq 1$, and $\omega_1 \lesssim 1$,

$$U_n(1+\delta) \;\lesssim\; U_n(\tau) \;=\; \nu^{-n} + \nu^{-n+2} + \nu^{-n+4} + \ldots + \nu^{n}\,. \tag{2.23}$$

The relation (2.23) is applied below to compare our convergence results for CG and MINRES with the classical convergence bound for these methods that is based on the matrix condition number.

**3. Worst-case and unbiased convergence quantities.** Our first goal in this section is to characterize the quantities given by (2.10), (2.12), (2.16) and (2.18). All these quantities depend in some way on the terms $L_k$, which are characterized by the following lemma.

LEMMA 3.1. *Suppose that* $\lambda_1, \ldots, \lambda_n$ *are given by* (2.4) *and* (2.6) *satisfying the assumption* (2.3). *Then* $L_k$ *as defined in* (2.9) *satisfies*

$$L_k = \frac{U_n\,(1+\delta)}{n+1} \cdot \frac{\sin^2\,(k\pi h)}{\delta + 2\sin^2\left(\frac{k\pi h}{2}\right)}\,. \tag{3.1}$$

*In particular, for* $\delta = 0$,

$$L_k \;=\; 2\cos^2\left(\frac{k\pi h}{2}\right). \tag{3.2}$$

*Proof.* According to (2.19),

$$\prod_{\substack{j=1\\ j\neq k}}^{n} |\lambda_j| = \frac{1}{|\lambda_k|}\,|\beta|^n U_n\,(1+\delta) = \frac{|\beta|^{n-1} U_n\,(1+\delta)}{2\,\left(\delta + 2\sin^2\left(\frac{k\pi h}{2}\right)\right)}\,. \tag{3.3}$$

The denominator in (2.9) can be written as

$$\begin{aligned}\prod_{\substack{j=1\\ j\neq k}}^{n} |\lambda_j - \lambda_k| \;&=\; |\beta|^{n-1} \prod_{\substack{j=1\\ j\neq k}}^{n} |2\,\omega_k - 2\,\omega_j| \\ &=\; 2^{2n-2}|\beta|^{n-1} \prod_{\substack{j=1\\ j\neq k}}^{n} \left|\sin^2\left(\frac{jh\pi}{2}\right) - \sin^2\left(\frac{kh\pi}{2}\right)\right| \\ &=\; |\beta|^{n-1}\frac{n+1}{2\sin^2\left(k\pi h\right)}\,,\end{aligned}$$

cf. identity (6.1), and hence (3.1) follows.

Considering $\delta = 0$ and using $U_n(1) = n+1$, (3.1) can be written as

$$L_k \;=\; \frac{\sin^2\left(k\pi h\right)}{2\sin^2\left(\frac{k\pi h}{2}\right)} \;=\; 2\cos^2\left(\frac{k\pi h}{2}\right)\,,$$

which finishes the proof. ∎

We point out that this lemma gives explicit expressions for the coefficients (2.8) and (2.15) leading to $(n-1)$st worst-case MINRES residuals and CG errors, respectively. We continue with deriving bounds on the norms (2.10)/(2.16) and (2.12).

THEOREM 3.2. *Suppose that MINRES is applied to a system $Ax = b$ where $A$ as in* (2.2) *with $n \geq 2$ has entries satisfying* (2.3)*. Then*

$$3^{-1}\,\frac{2+\delta}{U_n\,(1+\delta)} \;<\; \frac{\|r^u_{n-1}\|}{\|r^u_0\|} \;<\; \frac{\|r^w_{n-1}\|}{\|r^{(n-1)}_0\|} \;\leq\; 3\,\frac{2+\delta}{U_n\,(1+\delta)}\,. \tag{3.4}$$

*In particular, for $\delta = 0$,*

$$\frac{1}{n}\sqrt{\frac{2}{3}} \;<\; \sqrt{\frac{2}{3n^2-n}} \;=\; \frac{\|r^u_{n-1}\|}{\|r^u_0\|} \;<\; \frac{\|r^w_{n-1}\|}{\|r^{(n-1)}_0\|} \;=\; \frac{1}{n}\,. \tag{3.5}$$

*Proof.* We first prove (3.4). The middle inequality is trivial. To show the leftmost inequality it suffices to use the relation (2.12) and to find an upper bound on the sum of the $L_k^2$. Using (3.1) and (6.6),

$$\begin{aligned}\sum_{k=1}^{n} L_k^2 \;&\leq\; \frac{U_n^2\,(1+\delta)}{(n+1)^2(\frac{\delta}{2}+1)^2}\sum_{k=1}^{n}\frac{\sin^4(k\pi h)}{4\sin^4\left(\frac{k\pi h}{2}\right)} \\ &=\; \frac{16\,U_n^2\,(1+\delta)}{(n+1)^2(\delta+2)^2}\sum_{k=1}^{n}\cos^4\left(\frac{k\pi h}{2}\right) \\ &=\; \frac{(6\,n-2)\,U_n^2\,(1+\delta)}{(n+1)^2(\delta+2)^2}\,. \qquad (3.6)\end{aligned}$$

Then (2.12) implies

$$\left( n \sum_{k=1}^{n} L_k^2 \right)^{-1/2} \;\geq\; \frac{(n+1)(\delta+2)}{\sqrt{(6\,n-2)n}\,U_n\,(1+\delta)} \;>\; \frac{1}{3}\,\frac{\delta+2}{U_n\,(1+\delta)}\,.$$

Next note that, using (6.5),

$$\begin{aligned} \sum_{k=1}^{n} L_k \;&\geq\; \frac{U_n\,(1+\delta)}{\delta+2} \sum_{k=1}^{n} \frac{\sin^2\,(k\pi h)}{n+1} \;=\; \frac{1}{2}\,\frac{U_n\,(\delta+1)}{\delta+2}\,\frac{n}{n+1} \\ &\geq\; \frac{1}{3}\,\frac{U_n\,(\delta+1)}{\delta+2}\,, \end{aligned} \tag{3.7}$$

and thus the rightmost inequality in (3.4) follows from applying (3.7) to (2.10).

For $\delta = 0$ we have

$$\sum_{k=1}^{n} L_k = 2 \sum_{k=1}^{n} \cos^2\left(\frac{k\pi h}{2}\right) \;=\; n\,, \tag{3.8}$$

cf. (6.5), and

$$\begin{aligned} \sum_{k=1}^{n} L_k^2 \;&=\; \frac{U_n^2\,(1)}{(n+1)^2} \sum_{k=1}^{n} \frac{\sin^4(k\pi h)}{4\sin^4\left(\frac{k\pi h}{2}\right)} \\ &=\; 4 \sum_{k=1}^{n} \cos^4\left(\frac{k\pi h}{2}\right) \;=\; \frac{3\,n-1}{2}\,, \end{aligned} \tag{3.9}$$

cf. (6.6). Substituting (3.8) and (3.9) into (2.10) and (2.12), we obtain (3.5). ∎

The theorem has interesting implications. First, for $\delta = 0$, MINRES in the worst case decreases the relative residual norm in the first $n-1$ iteration steps only to $n^{-1}$, cf. the rightmost equation in (3.5). But since $\|r_{n-1}^w\|/\|r_0^{(n-1)}\| \approx (1+\delta)/U_n(1+\delta)$, for all $\delta$, the $(n-1)$st worst-case MINRES residual norm in fact decreases dramatically when we increase $\delta$, and hence the diagonal dominance of $A$ (see Fig. 2.1 and the corresponding discussion).

Second, the progress MINRES has made in the next-to-last iteration step for the unbiased initial residual could hardly be any worse, since the relative residual norm is at most a *constant factor* (less than 1/9) apart from the worst case. This is somewhat surprising since in general the two cases may differ by a factor of up to $n^{1/2}$, see [12, Section 5], [9, Section 5].

Numerical illustrations are given in Fig. 3.1, where we plot the MINRES residual norms for 40 by 40 systems of the form (2.1)-(2.3) with different $\delta$ and the initial residuals $r_0^{(n-1)}$ (solid line) and $r_0^u$ (corresponding dashed line). The figure shows the increasing speed of convergence for both types of initial residuals resulting from the increase of $\delta$. Furthermore, the figure reveals that not only $\|r_{n-1}^w\|/\|r_0^{(n-1)}\| \approx \|r_{n-1}^u\|/\|r_0^u\|$, as predicted by Theorem 3.2, but that in fact

$$\frac{\|r_i^{(n-1)}\|}{\|r_0^{(n-1)}\|} \;\approx\; \frac{\|r_i^u\|}{\|r_0^u\|}\,, \quad \text{for} \quad i = 0, 1, \ldots\,,$$

where $r_i^{(n-1)}$ denotes the $i$th MINRES residual corresponding to $r_0^{(n-1)}$ (and hence $r_{n-1}^{(n-1)} = r_{n-1}^w$). In fact, it seems that the MINRES convergence for $r_0^u$ closely corresponds to the convergence behavior for which the worst possible residual norm is attained in the next-to-last step.

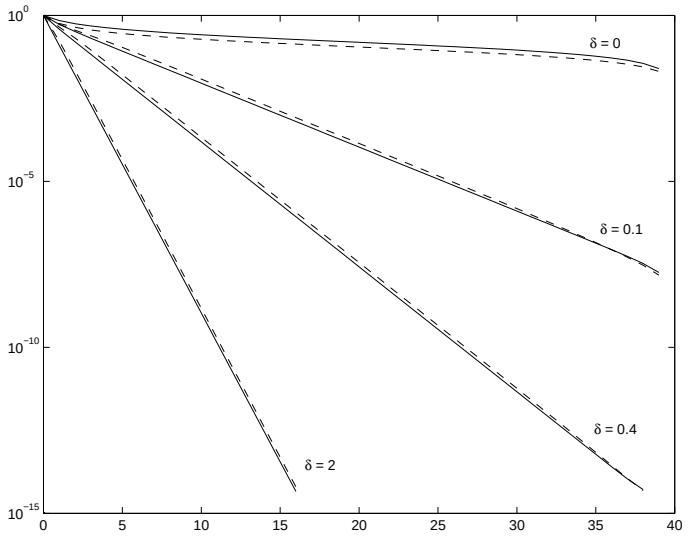


FIG. 3.1. *MINRES residual norms* $\|r_i^{(n-1)}\|/\|r_0^{(n-1)}\|$ *(solid) vs.* $\|r_i^u\|/\|r_0^u\|$ *(dashed) for different* $\delta$.

Because of the equality of (2.12) and (2.16), Theorem 3.2 also characterizes $\|e_{n-1}^w\|/\|e_0^{(n-1)}\|$, the $(n-1)$st worst-case relative $A$-norm of the error for CG. The theorem does not characterize, however, the case of CG for the initial error $e_0^u$. This is done in the following result.

THEOREM 3.3. *Suppose that CG is applied to a system* $Ax = b$ *where* $A$ *as in* (2.2) *with* $n \geq 2$ *has entries satisfying* (2.3). *Then*

$$3^{-1}\,\frac{\delta}{U_n\,(1+\delta)} \;<\; \frac{\|e_{n-1}^u\|_A}{\|e_0^u\|_A} \;<\; 3\,\frac{2+\delta}{U_n\,(1+\delta)}\,. \tag{3.10}$$

*For* $\delta < 1/4$,

$$3^{-1}\,\frac{\delta+2}{n^{1/2}U_n\,(1+\delta)} \;<\; \frac{\|e_{n-1}^u\|_A}{\|e_0^u\|_A}\,, \tag{3.11}$$

*and for* $\delta = 0$,

$$\frac{\|e_{n-1}^u\|_A}{\|e_0^u\|_A} \;=\; \frac{\sqrt{6}}{\sqrt{n(n+1)(n+2)}} \;>\; n^{-3/2}\,. \tag{3.12}$$

*Proof.* The second inequality in (3.10) follows easily from (3.4). To prove the first inequality it suffices to bound the term

$$K \equiv \left(\sum_{k=1}^{n} \lambda_k L_k^2\right)\left(\sum_{k=1}^{n} \frac{1}{\lambda_k}\right)$$

from above, cf. (2.18). Using Cauchy's inequality

$$K \;\leq\; \left(\sum_{k=1}^{n} L_k^4\right)^{1/2}\left(\sum_{k=1}^{n} \lambda_k^2\right)^{1/2}\left(\sum_{k=1}^{n} \frac{1}{\lambda_k}\right)\,. \tag{3.13}$$

Since $\lambda_n$ is the largest eigenvalue,

$$\left(\sum_{k=1}^{n}\lambda_k^2\right)^{1/2}\left(\sum_{k=1}^{n}\frac{1}{\lambda_k}\right) \;<\; n^{1/2}\lambda_n\left(\sum_{k=1}^{n}\frac{1}{\lambda_k}\right) \;<\; n^{3/2}\kappa(A)\,, \tag{3.14}$$

where $\kappa(A)$ denotes the condition number of the symmetric positive definite matrix $A$,

$$\kappa(A) = \frac{\lambda_n}{\lambda_1} = \frac{1+\delta+\omega_1}{1+\delta-\omega_1} > \frac{2+\delta}{\delta}\,.$$

It remains to find a bound on the sum of the $L_k^4$. Using (3.1) and (6.7),

$$\begin{aligned}\sum_{k=1}^{n} L_k^4 \;&\le\; \frac{U_n^4\,(1+\delta)}{(n+1)^4(\frac{\delta}{2}+1)^4}\sum_{k=1}^{n}\frac{\sin^8(k\pi h)}{2^4\sin^8\left(\frac{k\pi h}{2}\right)}\\ &=\; 2^8\frac{U_n^4\,(1+\delta)}{(n+1)^4(\delta+2)^4}\sum_{k=1}^{n}\cos^8\left(\frac{k\pi h}{2}\right)\\ &<\; 3^4\frac{n\,U_n^4\,(1+\delta)}{(n+1)^4(\delta+2)^4}\,. \qquad (3.15)\end{aligned}$$

From (3.13)–(3.15) we now obtain (3.10).

Now consider the case $\delta < 1/4$. Then

$$\begin{aligned}\left(\sum_{k=1}^{n}\lambda_k^2\right)^{1/2}\sum_{k=1}^{n}\frac{1}{\lambda_k} \;&=\; \left(\sum_{k=1}^{n}(1+\delta-\omega_k)^2\right)^{1/2}\sum_{k=1}^{n}\frac{1}{1+\delta-\omega_k}\\ &<\; \left(\sum_{k=1}^{n}(5/4-\omega_k)^2\right)^{1/2}\sum_{k=1}^{n}\frac{1}{1-\omega_k}\\ &=\; \left(\frac{33}{16}n-\frac{1}{2}\right)^{1/2}\sum_{k=1}^{n}\frac{1}{2\sin^2\left(\frac{k\pi h}{2}\right)}\\ &<\; \left(\frac{36}{16}n\right)^{1/2}\frac{n(n+2)}{3} \;=\; \frac{n^{1/2}n(n+2)}{2}\\ &<\; \frac{n^{1/2}(n+1)^2}{2}\,, \qquad (3.16)\end{aligned}$$

where we have used the identities (6.9) and (6.10). Then (3.11) follows from (3.13), (3.15) and (3.16).

Let $\delta = 0$. To prove (3.12) we use (6.8) and (6.9),

$$\begin{aligned}\frac{\|e_0^u\|_A^2}{\|e_{n-1}^u\|_A^2} \;&=\; \left(\sum_{k=1}^{n}4\sin^2\left(\frac{k\pi h}{2}\right)4\cos^4\left(\frac{k\pi h}{2}\right)\right)\left(\sum_{k=1}^{n}\frac{1}{4\sin^2\left(\frac{k\pi h}{2}\right)}\right)\\ &=\; (n+1)\left(\frac{n(n+2)}{6}\right)\,.\end{aligned}$$

□

A comparison of Theorem 3.2 and Theorem 3.3 shows that for small $\delta$,

$$\text{(MINRES)}\qquad \frac{\|r_{n-1}^u\|}{\|r_0^u\|} \;\approx\; n^{1/2}\,\frac{\|e_{n-1}^u\|_A}{\|e_0^u\|_A}\qquad \text{(CG)}\,.$$

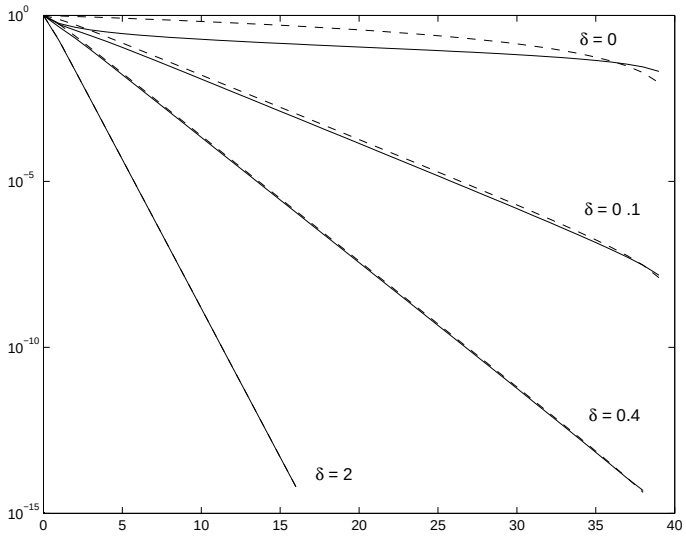


FIG. 3.2. *MINRES residual norms* $\|r_i^u\|/\|r_0^u\|$ *(solid) vs. CG error norms* $\|e_i^u\|_A/\|e_0^u\|_A$ *(dashed) for different* $\delta$.

For larger $\delta$, this difference is much less pronounced, and these MINRES and CG quantities are at most a small constant apart from each other. For a numerical illustration see Fig. 3.2, where we plot the MINRES and CG convergence curves for 40 by 40 systems of the form (2.1)-(2.3) with different $\delta$ and the initial residual $r_0^u$. Also see Fig. 4.1 for an example with $\delta = 0$ and $n = 120$.

We next compare our results with the classical bound on the worst-case convergence values based on the condition number of $A$, given by

$$\min_{p\in\pi_i} \max_{1\le k\le n} |p(\lambda_k)| \;\le\; \frac{2\nu^i}{1+\nu^{2i}}\,, \qquad i = 0,\ldots,n\,, \tag{3.17}$$

where $\nu = (\sqrt{\kappa(A)}-1)/(\sqrt{\kappa(A)}+1) < 1$ as in (2.21), cf., e.g., [7, Theorem 3.1.1]. The proof of this bound is based on the idea of replacing the min-max problem on the eigenvalues of $A$ by the min-max problem on the interval $[\lambda_1, \lambda_n]$. The latter is solved by the scaled and shifted Chebyshev polynomials of the first kind.

In the notation established above,

$$\frac{2\nu^{n-1}}{1+\nu^{2(n-1)}} \;\ge\; \frac{\|r_{n-1}^w\|}{\|r_0^{(n-1)}\|} \;=\; \frac{\|e_{n-1}^w\|}{\|e_0^{(n-1)}\|} \tag{3.18}$$

$$\gtrsim\; \frac{\|r_{n-1}^u\|}{\|r_0^u\|} \;\approx\; \frac{4}{\omega_1}\,\frac{2+\delta}{U_n\,(1+\delta)} \tag{3.19}$$

$$\gtrsim\; \frac{4\tau}{U_n(\tau)} \tag{3.20}$$

$$\gtrsim\; \frac{2}{\nu\,U_n(\tau)} \;=\; \frac{2\,\nu^{n-1}}{1+\nu^2+\ldots+\nu^{2(n-1)}+\nu^{2n}}\,, \tag{3.21}$$

where "$\gtrsim$" means that the inequality is close. In (3.18) we use (3.17) for $i = n-1$, and in (3.19) we use (3.4), where the unimportant multiplicative factor (between 1/3 and 3) was replaced by $4/\omega_1$ for convenience. Next, in (3.20) we use (2.23) as well as the relation $\tau = (1+\delta)/\omega_1$, from which we receive (3.21) using (2.23) and the inequality $2\,\tau \ge \nu^{-1} \ge \tau$.

The main point in the above derivation is that the actual convergence quantities on the right hand side of the inequality in (3.18) are always quite close to (3.21), i.e.

$$\frac{\|r_{n-1}^w\|}{\|r_0^{(n-1)}\|} \;=\; \frac{\|e_{n-1}^w\|}{\|e_0^{(n-1)}\|} \;\approx\; \frac{2\,\nu^{n-1}}{1+\nu^2+\ldots+\nu^{2(n-1)}+\nu^{2n}}\,.$$

The tightness of the *upper* bound (3.18) to the actual convergence quantities therefore depends on the size of $\nu$, which is related to the condition number of $A$. By definition, we always have $0 < \nu < 1$. Moreover, $\nu$ approaches zero when $\kappa(A)$ approaches one, and $\nu$ approaches one, when $\kappa(A)$ approaches infinity. Also note that $\kappa(A)$ for a fixed matrix size $n$ is a strictly decreasing function of the parameter $\delta \geq 0$.

When $\nu$ is close to zero, equivalently $\kappa(A)$ is small or $\delta$ is bounded away from zero, then there is no significant quantitative difference between the classical bound (3.18) and (3.21). In such cases the classical bound in fact provides accurate information about the actual convergence quantities of CG and MINRES in (3.18) and (3.19). On the other hand, when $\nu$ is close to one, equivalently $\kappa(A)$ is large or $\delta$ is close to zero, then the lower bound (3.21), and with it the CG and MINRES convergence quantities will be smaller (up to the factor $n^{-1}$) than predicted by the classical upper bound (3.18). Note that in this case the CG error norm for $e_0^u$ may be well below the other three convergence quantities, cf. our above discussion of Theorem 3.3. The case $\delta = 0$ appears to be the most interesting case, and we study it in more detail in the following section.

**4. Model problem: The one-dimensional Poisson equation.** We will now apply the results developed above to a model problem, namely the one-dimensional Poisson equation,

$$-u''(z) = f(z), \quad z \in (0,1), \tag{4.1}$$

with Dirichlet boundary conditions

$$u(0) = u_0, \quad u(1) = u_1\,. \tag{4.2}$$

Then for each positive integer $n$, the central finite difference approximation on the uniform grid $kh$, $k = 1,\ldots,n$, $h = 1/(n+1)$, leads to a linear algebraic system of the form

$$\begin{bmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix} x \;=\; h^2 \begin{bmatrix} f(h) \\ \vdots \\ \vdots \\ f(nh) \end{bmatrix} + \begin{bmatrix} u_0 \\ \\ \\ u_1 \end{bmatrix} \equiv b\,. \tag{4.3}$$

The coefficient matrix $A$ is of the form (2.2) with $\alpha = 2$ and $\beta = -1$, so that $\delta = 0$ in (2.3), i.e. $A$ is only weakly diagonally dominant. The eigenvalues of $A$ are given by

$$\lambda_k \;=\; 2 - 2\,\cos(k\pi h) = 4\,\sin^2\left(\frac{k\pi h}{2}\right)\,, \quad k = 1,\ldots,n\,, \tag{4.4}$$

and the eigenvectors are the same as in (2.5).

REMARK 4.1. The results developed in the previous sections could also be used to study discretized one-dimensional elliptic differential equations that are more general than (4.1). For example, consider the elliptic equation $-u''(z) + \sigma u(z) = f(z)$ with some parameter $\sigma \geq 0$. Then the coefficient matrix resulting from a central finite difference discretization as described above is of the form (2.2) with $\alpha = 2 + \sigma h^2$, and

$\beta = -1$, and hence $\delta = (\sigma h^2)/2 \geq 0$. In a nutshell, our results in Section 3 show that increasing $\sigma$ will increase the convergence speed of MINRES and CG (assuming that $h$ is fixed), cf. Figs. 3.1 and 3.2 and the corresponding analyses. ∎

In [14], see also [15], the authors study the convergence of CG for the system (4.3) for a certain set of solutions $x = A^{-1}b$ dependent on a parameter. Assuming that $x_0 = 0$, and hence $e_0 = x$, they present exact analytic expressions for the relative $A$-norm of the CG errors. Two of these solutions are of particular interest in our context.

The first solution, here denoted by $x^{(M)}$, is defined by

$$x^{(M)} = Q[\xi_1^{(M)}, \ldots, \xi_n^{(M)}]^T, \qquad \xi_k^{(M)} = \sin^{-1}\left(\frac{k\pi h}{2}\right), \tag{4.5}$$

for $k = 1, \ldots, n$. If $x_0 = 0$, equivalently $e_0^{(M)} = x^{(M)}$, then the relative $A$-norm of the corresponding $i$th CG error $e_i^{(M)}$ fulfills

$$\frac{\|e_i^{(M)}\|_A}{\|e_0^{(M)}\|_A} \;=\; \left[\frac{(n-i)+(n-i)^2}{n(n+1)+2ni(n-i)}\right]^{1/2} \;\equiv\; \varphi_{M}(n,i)\,, \tag{4.6}$$

see [14, p. 222].

Similarly, for the solution $x^{(C)}$ defined by

$$x^{(C)} = Q[\xi_1^{(C)}, \ldots, \xi_n^{(C)}]^T, \qquad \xi_k^{(C)} = \sin^{-2}\left(\frac{k\pi h}{2}\right), \tag{4.7}$$

for $k = 1, \ldots, n$, the corresponding CG errors for $x_0 = 0$, i.e. $e_0^{(C)} = x^{(C)}$, satisfy

$$\frac{\|e_i^{(C)}\|_A}{\|e_0^{(C)}\|_A} \;=\; \left[\frac{(n-i)^3+3(n-i)^2+2(n-i)}{n(n+1)(n+2)}\right]^{1/2} \;\equiv\; \varphi_{C}(n,i)\,, \tag{4.8}$$

see [14, p. 229].

REMARK 4.2. The solutions $x^{(M)}$ and $x^{(C)}$ correspond to choosing the parameter in [14, formula (2.7)] equal to one and two, respectively. The authors solve the resulting minimization problems using Lagrange multipliers. They do not comment on their motivation for choosing these particular solution vectors, but according to Howard Elman (private communication) they were mainly motivated by numerical experiments. ∎

What is the meaning of the solutions $x^{(M)}$ in (4.5) and $x^{(C)}$ in (4.7) in our context? First consider the solution $x^{(C)}$. If $x_0 = 0$, then initial error is $e_0^{(C)} = x^{(C)}$. Comparing (2.17) and (4.5) shows that

$$e_0^{(C)} \;=\; 4\, e_0^{u}\,.$$

Hence (4.8) gives the exact convergence curve for the relative $A$-norm of the error, when CG is applied to (4.3) and the initial error is $e_0^u$.

Next note that the convergence of CG with initial error $x^{(M)}$ (i.e. solution $x^{(M)}$ and $x_0 = 0$) corresponds to the convergence of MINRES with initial residual $A^{1/2}x^{(M)}$, cf. (2.14). Now a simple calculation shows that the coordinates $\varrho_k^{(M)}$ of $A^{1/2}x^{(M)}$ in the eigenvectors of $A$ are given by $\varrho_k^{(M)} = 2$, $k = 1, \ldots, n$. Equivalently,

$$A^{1/2}x^{(M)} \;=\; 2\, r_0^{u}\,,$$

cf. (2.11). Hence (4.6) gives the exact convergence curve for the relative residual norms, when MINRES is applied to (4.3) and the initial residual is $r_0^u$.

We summarize these considerations in the following proposition.

PROPOSITION 4.3. *Suppose that CG and MINRES are applied to the system* (4.3) *and the respective initial error and residual are given by* $e_0^u$ *and* $r_0^u$. *Then the resulting CG errors* $e_i^u$ *and MINRES residuals* $r_i^u$ *satisfy*

$$\frac{\|e_i^u\|_A}{\|e_0^u\|_A} \;=\; \varphi_{C}(n,i), \qquad \frac{\|r_i^u\|}{\|r_0^u\|} \;=\; \varphi_{M}(n,i)\,, \quad i=0,\ldots,n\,,$$

*where* $\varphi_{C}(n,i)$ *and* $\varphi_{M}(n,i)$ *are defined by* (4.6) *and* (4.8).

Our next goal is to characterize the initial residuals for which the bounds (1.1) and (1.2) are attained in step $n-1$, and to relate these initial residuals to actual right hand sides $f$ and/or boundary conditions in our model problem (4.1)–(4.2).

**4.1. Worst data for CG.** The eigenvector coordinates $\xi_k^{(n-1)}$ of the initial error $e_0^{(n-1)}$ that yields the maximal relative $A$-norm of the error in the step $n-1$ of CG are determined by (2.15). For example, choosing them all positive, and using (3.2) and (4.4), yields

$$\xi_k^{(n-1)} \;=\; \left(\frac{\gamma}{2}\right)^{1/2} \cot\left(\frac{k\pi h}{2}\right) \tag{4.9}$$

for $k=1,\ldots,n$, and an arbitrary $\gamma>0$. For simplicity consider $x_0=0$, then $Ae_0^{(n-1)}$ is the right hand side of (4.3), which is given by

$$Ae_0^{(n-1)} \;=\; \left(\frac{\gamma}{h}\right)^{1/2} Q\,q_1 = \left(\frac{\gamma}{h}\right)^{1/2} e_1\,, \tag{4.10}$$

where $e_1$ is the first unit vector, $e_1=[1,0,\ldots,0]^T$. Of course, the scaling of the right hand side makes no difference for the relative $A$-norm of the error. Hence any right hand side that is a (nonzero) multiple of $e_1$ leads to the worst possible relative $A$-norm of the error in the next-to-last step of CG (with $x_0=0$).

Another example that leads to an $(n-1)$st worst-case CG error is to choose the coefficients $\xi_k^{(n-1)}$ similar to (4.9), but with alternating sign, i.e.

$$\xi_k^{(n-1)} \;=\; (-1)^{k+1} \cot\left(\frac{k\pi h}{2}\right).$$

This yields, using the relation $(-1)^{k+1}\sin(k\pi h) = \sin(nk\pi h)$ and $x_0=0$, a right hand side that is a (nonzero) multiple of the $n$th unit vector $e_n$.

Both examples show that the initial data leading to the very unfavorable convergence behavior of CG may look rather unsuspicious at first sight. In terms of the model problem (4.1)–(4.2), the worst possible relative $A$-norm of the $(n-1)$st error in CG (for $x_0=0$) is obtained simply by

$$f=0 \quad \text{and} \quad u_0=c,\; u_1=0, \quad \text{or} \quad u_0=0,\; u_1=c\,, \tag{4.11}$$

for any nonzero constant $c$.

For the initial error $e_0^{(n-1)}$ in (4.9) it is also possible to determine the exact values of the relative $A$-norm of the error in every step of the CG method. This can be done using the same techniques as in [14] based on Lagrange multipliers. This technique is quite involved, and the full proof would take us several pages to state. We here mention only the final result, and justify it numerically in Section 4.3: The exact

relative $A$-norm of the CG error $e_i^{(n-1)}$ resulting from the initial error $e_0^{(n-1)}$ in (4.9) is given by

$$\frac{\|e_i^{(n-1)}\|_A}{\|e_0^{(n-1)}\|_A} \;=\; \left[\frac{n-i}{n\,(i+1)}\right]^{1/2} \;\equiv\; \varphi_W(n,i)\,, \quad i=0,\ldots,n\,. \tag{4.12}$$

Comparing (4.12) and (4.6) it can be easily shown that

$$\varphi_M(n,i) \;<\; \varphi_W(n,i) \;<\; \sqrt{2}\,\varphi_M(n,i)\,, \quad i=1,\ldots,n-1\,. \tag{4.13}$$

**4.2. Worst data for MINRES.** The eigenvector coordinates $\varrho_k^{(n-1)}$ of the initial residual $r_0^{(n-1)}$ that yields the maximal relative Euclidean residual norm in the step $n-1$ of MINRES are determined by (2.8) and (3.2). Let us choose $\varrho_k^{(n-1)} = \cos\left(\frac{k\pi h}{2}\right)$, i.e. all $\varrho_k^{(n-1)}$ positive and $\gamma = 1/2$, for $k = 1,\ldots,n$. As shown in (2.14), MINRES for this $r_0^{(n-1)}$ is equivalent to CG for the initial *error* $A^{-1/2}r_0^{(n-1)}$. But a simple calculation shows that the eigenvector coordinates of this initial error are given by $\frac{1}{2}\cot\left(\frac{k\pi h}{2}\right)$. Hence $A^{-1/2}r_0^{(n-1)}$ is a multiple of the error vector $e_0^{(n-1)}$ described by (4.9). As a consequence, the relative MINRES residual norms for the initial residual $r_0^{(n-1)}$ also satisfy (4.12), i.e.

$$\frac{\|r_i^{(n-1)}\|}{\|r_0^{(n-1)}\|} \;=\; \varphi_W(n,i)\,, \quad i=0,\ldots,n\,. \tag{4.14}$$

Similar to CG, we now determine data for (4.1)–(4.2), which yields the vector $r_0^{(n-1)}$. First note that

$$r_0^{(n-1)} \;=\; \sum_{k=1}^{n} \varrho_k^{(n-1)}\, q_k \;=\; (2h)^{1/2} \sum_{k=1}^{n} \cos\left(\frac{k\pi h}{2}\right) \begin{bmatrix} \sin(k\pi h) \\ \sin(2k\pi h) \\ \vdots \\ \sin(nk\pi h) \end{bmatrix}. \tag{4.15}$$

Now the $j$th entry of $r_0^{(n-1)}$, denoted by $r_{0,j}^{(n-1)}$ for $j = 1,\ldots,n$, satisfies

$$\begin{aligned} r_{0,j}^{(n-1)} \;&=\; (2h)^{1/2} \sum_{k=1}^{n} \cos\left(\frac{k\pi h}{2}\right) \sin(jk\pi h)\,, \\ &=\; (2h)^{1/2} \sum_{k=1}^{n} \left[\sin\left(kj\pi h + \frac{k\pi h}{2}\right) + \sin\left(kj\pi h - \frac{k\pi h}{2}\right)\right] \\ &=\; (2h)^{1/2} \frac{\sin(j\pi h)}{\cos\left(\frac{\pi h}{2}\right) - \cos(j\pi h)}\,, \end{aligned} \tag{4.16}$$

cf. formula (6.2) for the last equality.

The dependence of $r_0^{(n-1)}$ on $h$ is a striking difference from the CG case, where $r_0^{(n-1)}$ is just a scalar multiple of the first or last standard unit vector. In the MINRES case it is therefore not as straightforward to find data for (4.1)–(4.2) that leads to the worst case in step $n-1$. One possibility is to choose $u_0 = 0$, $u_1 = 0$, and the source function

$$f_n(z) \;\equiv\; \frac{\sin(\pi z)}{\cos\left(\frac{\pi h}{2}\right) - \cos(\pi z)}\,.$$

(Note that the scaling factor $(2h)^{1/2}$ in (4.16) does not play any role since we consider the relative residual norm.) This $f_n(z)$ has a pole at $z = \frac{h}{2}$. Given that for CG we have the nicely continuous source function $f(z) = 0$, this is somewhat unsatisfactory. However, it is not difficult to show that

$$\frac{\sin(\pi z)}{1-\cos(\pi z)} \;=\; \cot\left(\frac{\pi z}{2}\right) \;\le\; f_n(z) \;\le\; \sqrt{2}\,\cot\left(\frac{\pi z}{2}\right), \tag{4.17}$$

for any $z \in [h, 1]$. Hence $f_n(z)$ is closely approximated by the function $f(z) = \cot\left(\frac{\pi z}{2}\right)$, that is continuous in $(0, 1]$. In particular, (4.17) holds at any mesh point $kh$, $k = 1, \ldots, n$. Hence the problem (4.1)–(4.2) with homogeneous Dirichlet boundary conditions, and with the source term $f(z) = \cot\left(\frac{\pi z}{2}\right)$, leads to MINRES convergence (for $x_0 = 0$) that is very close to the one obtained for the initial residual $r_0^{(n-1)}$ in (4.15), see Fig. 4.1 in Section 4.3.

**4.3. Numerical Experiments.** Above we explored the convergence of CG and MINRES for the model problem (4.3). We now illustrate our theoretical results by numerical experiments. In all experiments we use $n = 120$, and the initial guess $x_0 = 0$. The experiments are performed in MATLAB 6.5, Release 13, on an AMD Athlon XP 2100+ personal computer with machine precision $\varepsilon \sim 10^{-16}$.

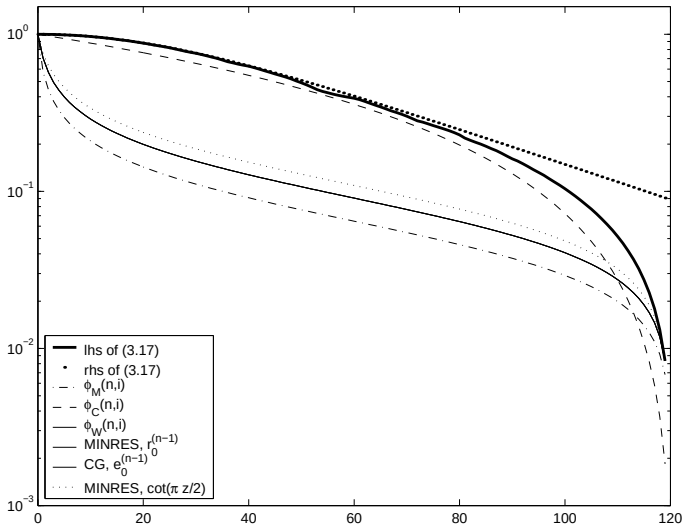


FIG. 4.1. *CG and MINRES convergence curves, and both sides of* (3.17).

*Figure 4.1*: As shown in Proposition 4.3, the CG and MINRES convergence curves for $e_0^u$ and $r_0^u$ are given by $\varphi_C(n, i)$ (dashed) and $\varphi_M(n, i)$ (dashed dotted), respectively. These curves may in general for small $\delta$ be quite different from each other, cf. the discussion after Theorem 3.3, and in fact they are in this example (where $\delta = 0$). Here also we notice that the convergence curve $\varphi_C(n, i)$ exhibits a "superlinear" behavior, i.e. that the error norm reduction per iteration step increases with increasing $i$. This behavior can be easily explained by considering the form of $\varphi_C(n, i)$, cf. (4.8). It can be shown by an elementary computation, that

$$\frac{\varphi_C(n,i)}{\varphi_C(n,i-1)} \;=\; \left(\frac{n-i}{n-i+3}\right)^{1/2}, \qquad i = 1, \ldots, n\,,$$

which represents a strictly decreasing function of the iteration step $i$. The superlinear effect can also be related to the distribution of the eigenvector coordinates of the initial error $e_0^u$. As proved asymptotically by Beckermann and Kuijlaars [3], CG may for the model problem (4.3) converge superlinearly, when the initial error exhibits a certain distribution of eigencomponents that is far from a uniform distribution. This is precisely the case in our example, where $e_0^u$ is *biased*, cf. (2.17).

The convergence curves of CG with right-hand side (4.10), MINRES with right hand side (4.16), as well as the curve $\varphi_W(n,i)$ defined by the right equality in (4.12) are plotted by solid lines. The three lines coincide, which gives a numerical justification of the left equality in (4.12), and of (4.14). As predicted by (4.13), the curves $\varphi_M(n,i)$ (MINRES for $r_0^u$) and $\varphi_W(n,i)$ (CG for $e_0^{(n-1)}$ and MINRES for $r_0^{(n-1)}$) are very close.

Applying MINRES to the system (4.3) with the initial residual having components $r_{0,j} = \cot(\frac{j\pi h}{2})$ instead of (4.16) yields the convergence curve plotted by dots. Because of (4.17), the dotted and the solid curve do not differ significantly.

The left hand side of the classical convergence bound (3.17), computed by the function `cheby0` of the semidefinite programming package SDPT3 [19], is plotted by the bold line, and the right hand side is plotted by bold dots. It is quite surprising how close the curve $\varphi_C(n,i)$ (CG for $e_0^u$) is to the worst-case values (bold dots) during the iterations $i \le n/2$. On the other hand, the convergence curve $\varphi_W(n,i)$ (solid), that attains the worst-case value in the $(n-1)$st step, differs from the bound (1.1) significantly in most iterations.

The bound (3.17) is tight in the step $i$, if there exist $i-1$ eigenvalues of $A$, that closely approximate extrema of the $i$th scaled and shifted Chebyshev polynomial of the first kind. As shown by Fig. 4.1, this is apparently true in our case for $i \le n/2$. For $i > n/2$, the left and right hand sides of (3.17) start to differ significantly. In particular, for $i = n-1$, and $n$ reasonably large,

$$\min_{p\in\pi_{n-1}} \max_{1\le k\le n} |p(\lambda_k)| \;=\; \frac{1}{n} \;\ll\; \frac{2\nu^{n-1}}{1+\nu^{2(n-1)}} \;\stackrel{n\to\infty}{\longrightarrow}\; \frac{2e^{\pi}}{1+e^{2\pi}}\,.$$

This clearly demonstrates that for reasonably large $n$ the classical bound (3.17) cannot describe the worst-case convergence values of CG or MINRES in later iterations. As shown by our discussion of the relations (3.18)–(3.21) at the end of Section 3, this is due to the large condition number of $A$ (note that $\delta = 0$ in this example). Asymptotically (for $n \to \infty$) the weakness of the classical bound in this context has also been noticed before by Axelsson [2, Example 13.7] and others.

*Figure 4.2.* In this experiment we demonstrate that the initial data (initial residual or initial error) that lead to the worst-case convergence quantity for one method in the $(n-1)$st iteration step does not lead (in general) to similar convergence for the other method. As mentioned before, MINRES for $r_0^{(n-1)}$ and CG for $e_0^{(n-1)}$ have the same convergence curve (solid). However, the curves of MINRES for the initial residual $Ae_0^{(n-1)}$ (dashed) and CG for the initial error $A^{-1}r_0^{(n-1)}$ (dashed dotted) differ significantly (by orders of magnitude) from the solid curve. An interesting numerical observation is that both curves (dashed and dashed dotted) end with the same convergence value in the $(n-1)$st iteration (we did not investigate this theoretically).

*Figure 4.3.* This experiment studies to which extent the initial residual $r_0^u$ in (2.11) can be considered a representative for a general initial residual vector having *approximately* equal components in the eigenvectors of $A$. We generate 1000 right hand sides by the MATLAB-command `b = Q * randn(n,1)`, i.e. the vector of eigenvector coordinates is generated by a random number generator with normal distribution. The corresponding CG and MINRES convergence curves with $x_0 = 0$ are plotted by dotted lines in the left and right part of Fig. 4.3. By the solid line we plot the left hand side of (3.17). The bold line in the left part represents the CG convergence

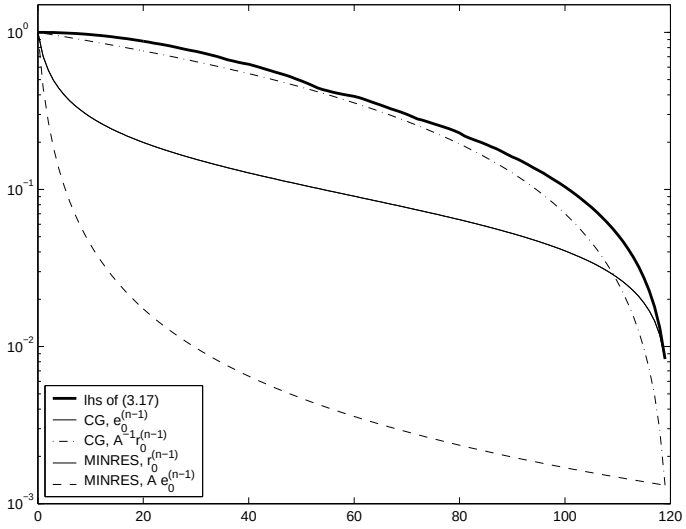


FIG. 4.2. *CG and MINRES convergence curves and the left hand side of* (3.17).

curve for $e_0^u = A^{-1}r_0^u$, i.e. the values $\varphi_C(n,i)$, and the bold line in the right part is the MINRES convergence curve for $r_0^u$, i.e. the values $\varphi_M(n,i)$. For CG, the bold solid line represents de facto (up to some small inaccuracies) an upper bound for all other CG convergence curves. In this sense $r_0^u$ represents for CG an extreme case of an unbiased initial residual vector. The MINRES convergence curve for $r_0^u$ describes the other MINRES convergence curves reasonably well, except for the final stage of the iteration.

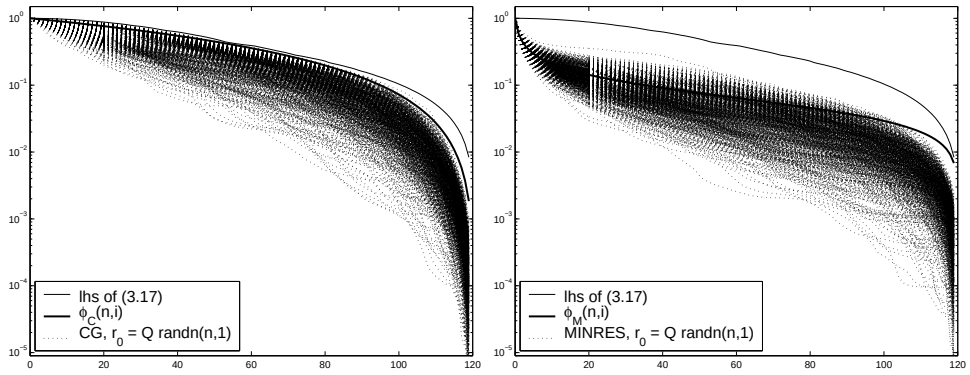


FIG. 4.3. *CG and MINRES convergence curves and the left hand side of* (3.17).

**5. Summary.** Our results in [12] allow to study the $(n-1)$st CG and MINRES iteration step. This approach provides additional information about the convergence of these methods. As demonstrated in this paper, such information can have interesting implications for understanding the convergence itself as well as connections between a differential equation and the convergence of the linear solver for the discretized problem.

Here we consider linear algebraic systems with symmetric (positive) definite tridiagonal Toeplitz matrices. We concentrate on the worst case and the case of an unbiased initial residual. It turns out that for MINRES these two cases are essentially

the same, as the corresponding relative residual norms differ by at most a small constant independent of the matrix size. On the other hand, for CG the two cases may differ more significantly, with the size of the difference depending on the degree of diagonal dominance of $A$. When $A$ is only weakly diagonally dominant, then CG for the unbiased initial residual may even exhibit a "superlinear" convergence behavior, leading to a much faster error norm reduction in later stages of the iteration. This may significantly outperform the worst-case behavior. The reason for this difference is that the initial error corresponding to the unbiased initial residual is in fact biased towards the eigenvalue distribution of $A$.

Additionally, a comparison of our results with the classical convergence bound based on the condition number of $A$ shows that this bound is reasonable when the matrix $A$ from our problem class is well conditioned, and that it otherwise fails to provide good information about the worst-case convergence quantities. Moreover, we show quantitatively how the speed of convergence increases with increasing diagonal dominance of $A$. For the one-dimensional Poisson equation, we identify the source terms and/or boundary conditions of the differential equation that lead to the worst CG and MINRES convergence for the discretized problem.

**Acknowledgments.** We thank Miro Rozložník for his helpful comments.

**6. Appendix.** Let $h = (n+1)^{-1}$, $n \in \mathbb{N}$. Then the following identities hold:

$$\frac{n+1}{2^{2n-1}} \frac{1}{\sin^2(k\pi h)} = \prod_{\substack{j=1 \\ j\neq k}}^{n} \left| \sin^2\left(\frac{j\pi h}{2}\right) - \sin^2\left(\frac{k\pi h}{2}\right) \right|, \tag{6.1}$$

$$\frac{\sin(k\pi h)}{\cos\left(\frac{\pi h}{2}\right) - \cos(k\pi h)} = \sum_{j=1}^{n} \left[ \sin\left(jk\pi h + \frac{j\pi h}{2}\right) + \sin\left(jk\pi h - \frac{j\pi h}{2}\right) \right], \tag{6.2}$$

$$\frac{n+1}{2^n} = \prod_{j=1}^{n} \sin(j\pi h), \tag{6.3}$$

$$\frac{\sin\left(\frac{ny}{2}\right)\sin\left(ny+\frac{y}{2}\right)}{\sin\left(\frac{y}{2}\right)} = \sum_{j=1}^{n} \sin(j\,y), \tag{6.4}$$

$$\frac{n}{2} = \sum_{j=1}^{n} \cos^2(j\pi h) = \sum_{j=1}^{n} \sin^2(j\pi h), \tag{6.5}$$

$$\frac{3\,n-1}{2^3} = \sum_{j=1}^{n} \cos^4\left(\frac{j\pi h}{2}\right), \tag{6.6}$$

$$\frac{35\,n-29}{2^7} = \sum_{j=1}^{n} \cos^8\left(\frac{j\pi h}{2}\right), \tag{6.7}$$

$$\frac{n+1}{16} = \sum_{j=1}^{n} \sin^2\left(\frac{j\pi h}{2}\right) \cos^4\left(\frac{j\pi h}{2}\right), \tag{6.8}$$

$$\frac{2\,n(n+2)}{3} = \sum_{j=1}^{n} \sin^{-2}\left(\frac{j\pi h}{2}\right), \tag{6.9}$$

$$\frac{33}{16}\,n - \frac{1}{2} = \sum_{j=1}^{n} \left(\frac{5}{4} - \cos(j\pi h)\right)^2. \tag{6.10}$$

The identities (6.3)–(6.10) are either standard identities, see, e.g., [5], or they

can be derived using standard symbolic computation software such as MAPLE [20]. Below we will only give proofs of the identities (6.1) and (6.2).

We start with a proof of (6.1). Using a simple algebraic manipulation and the standard formulas we obtain

$$\begin{aligned}
&\prod_{\substack{j=1\\ j\neq k}}^{n}\left[\sin^2\left(\frac{j\pi h}{2}\right)-\sin^2\left(\frac{k\pi h}{2}\right)\right]\\
&=\prod_{\substack{j=1\\ j\neq k}}^{n}\left[\sin\left(\frac{j\pi h}{2}\right)-\sin\left(\frac{k\pi h}{2}\right)\right]\left[\sin\left(\frac{j\pi h}{2}\right)+\sin\left(\frac{k\pi h}{2}\right)\right]\\
&=\prod_{\substack{j=1\\ j\neq k}}^{n}2\,\cos\left(\frac{(j+k)\pi h}{4}\right)\sin\left(\frac{(j-k)\pi h}{4}\right)2\,\sin\left(\frac{(j+k)\pi h}{4}\right)\cos\left(\frac{(j-k)\pi h}{4}\right)\\
&=\prod_{\substack{j=1\\ j\neq k}}^{n}\sin\left(\frac{(j+k)\pi h}{2}\right)\sin\left(\frac{(j-k)\pi h}{2}\right)\\
&=\prod_{\substack{j=1\\ j\neq k}}^{n}\sin\left(\frac{(j+k)\pi h}{2}\right)\prod_{\substack{j=1\\ j\neq n+1-k}}^{n}\cos\left(\frac{(j+k)\pi h}{2}\right).
\end{aligned}$$

Now we distinguish two situations, either $kh=\frac{1}{2}$, or not. If $kh=\frac{1}{2}$ then $n+1-k=k$ and the product in (6.1) takes the form

$$\begin{aligned}
\prod_{\substack{j=1\\ j\neq k}}^{n}\left|\sin\left(\frac{(j+k)\pi h}{2}\right)\cos\left(\frac{(j+k)\pi h}{2}\right)\right| &= \frac{1}{2^{n-1}}\prod_{\substack{j=1\\ j\neq k}}^{n}|\sin\left((j+k)\pi h\right)|\\
&= \frac{1}{2^{n-1}}\prod_{j=1}^{n}\sin\left(j\pi h\right) \;=\; \frac{n+1}{2^{2n-1}},
\end{aligned}$$

cf. (6.3). Clearly, (6.1) holds since $\sin^2(k\pi h)=1$ for $kh=\frac{1}{2}$.

When $kh\neq\frac{1}{2}$, the product in (6.1) can be written as

$$\begin{aligned}
&|\cos(k\pi h)|\prod_{\substack{j=1\\ j\neq k\\ j\neq n+1-k}}^{n}\left|\sin\left(\frac{(j+k)\pi h}{2}\right)\cos\left(\frac{(j+k)\pi h}{2}\right)\right|\\
&=\frac{|\cos(k\pi h)|}{2^{n-2}}\prod_{\substack{j=1\\ j\neq k\\ j\neq n+1-k}}^{n}|\sin\left((j+k)\pi h\right)|\\
&=\frac{|\cos(k\pi h)|}{2^{n-2}|\sin(2k\pi h)|}\cdot\prod_{\substack{j=1\\ j\neq n+1-k}}^{n}|\sin\left((j+k)\pi h\right)|\\
&=\frac{2\sin(k\pi h)\cos(k\pi h)}{2^{n-1}\sin(k\pi h)\sin(2k\pi h)}\cdot\frac{1}{\sin(k\pi h)}\prod_{j=1}^{n}\sin\left(j\pi h\right)\\
&=\frac{n+1}{2^{2n-1}}\,\frac{1}{\sin^2\left(k\pi h\right)},
\end{aligned}$$

and (6.1) holds.

Next, we will prove the identity (6.2). Denoting

$$y_+ \;=\; k\pi h + \frac{\pi h}{2}\,, \qquad y_- \;=\; k\pi h - \frac{\pi h}{2}\,,$$

and using the formula (6.4) we obtain

$$\begin{aligned}
&\sum_{j=1}^{n} \sin\left(j\, y_+\right) + \sum_{j=1}^{n} \sin\left(j\, y_-\right) \\
&= \frac{\sin\left(\frac{n y_+}{2}\right) \sin\left(\frac{(n+1)\, y_+}{2}\right)}{\sin\left(\frac{y_+}{2}\right)} + \frac{\sin\left(\frac{n y_-}{2}\right) \sin\left(\frac{(n+1)\, y_-}{2}\right)}{\sin\left(\frac{y_-}{2}\right)} \\
&= \frac{\sin\left(\frac{y_-}{2}\right) \sin\left(\frac{n y_+}{2}\right) \sin\left(\frac{(n+1)\, y_+}{2}\right) + \sin\left(\frac{y_+}{2}\right) \sin\left(\frac{n y_-}{2}\right) \sin\left(\frac{(n+1)\, y_-}{2}\right)}{\sin\left(\frac{y_+}{2}\right) \sin\left(\frac{y_-}{2}\right)}\,. \qquad (6.11)
\end{aligned}$$

The first product in numerator of (6.11) has the form

$$\begin{aligned}
&4 \sin\left(\frac{y_-}{2}\right) \sin\left(\frac{n y_+}{2}\right) \sin\left(\frac{(n+1)\, y_+}{2}\right) \\
&= \sin\left(k\pi h\right) - \sin\left(\pi h\right) - \sin\left(k\pi + \frac{n\pi h}{2}\right) + \sin\left(nk\pi h + \frac{\pi}{2}\right) \qquad (6.12)
\end{aligned}$$

and similarly, the second product in numerator can be written as

$$\begin{aligned}
&4 \sin\left(\frac{y_+}{2}\right) \sin\left(\frac{n y_-}{2}\right) \sin\left(\frac{(n+1)\, y_-}{2}\right) \\
&= \sin\left(k\pi h\right) + \sin\left(\pi h\right) - \sin\left(k\pi - \frac{n\pi h}{2}\right) + \sin\left(n\, k\pi h - \frac{\pi}{2}\right)\,. \qquad (6.13)
\end{aligned}$$

Substituting (6.12) and (6.13) into (6.11) we obtain

$$\begin{aligned}
\sum_{j=1}^{n} \sin\left(j\, y_+\right) + \sum_{j=1}^{n} \sin\left(j\, y_-\right) \;&=\; \frac{\sin\left(k\pi h\right) - \frac{1}{2}\left[\sin\left(k\pi + \frac{n\pi h}{2}\right) + \sin\left(k\pi - \frac{n\pi h}{2}\right)\right]}{\cos\left(\frac{\pi h}{2}\right) - \cos\left(k\pi h\right)} \\
&=\; \frac{\sin\left(k\pi h\right) - \sin\left(k\pi\right) \cos\left(\frac{n\pi h}{2}\right)}{\cos\left(\frac{\pi h}{2}\right) - \cos\left(k\pi h\right)} \\
&=\; \frac{\sin\left(k\pi h\right)}{\cos\left(\frac{\pi h}{2}\right) - \cos\left(k\pi h\right)}\,,
\end{aligned}$$

which completes the proof.

## REFERENCES


[1] 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.

[2] O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, 1994.

[3] B. Beckermann and A. B. J. Kuijlaars, *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).

[4] M. Eiermann and O. G. Ernst, *Geometric aspects of the theory of Krylov subspace methods*, Acta Numer., 10 (2001), pp. 251–312.
[5] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*, Academic Press [Harcourt Brace Jovanovich Publishers], New York, 1980. Corrected and enlarged edition edited by Alan Jeffrey, Incorporating the fourth edition edited by Yu. V. Geronimus [Yu. V. Geronimus] and M. Yu. Tseytlin [M. Yu. Tseĭtlin], Translated from the Russian.
[6] A. Greenbaum, *Comparison of splittings used with the conjugate gradient algorithm*, Numer. Math., 33 (1979), pp. 181–193.
[7] ———, *Iterative methods for solving linear systems*, vol. 17 of Frontiers in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.
[8] A. Greenbaum and L. Gurvits, *Max-min properties of matrix factor norms*, SIAM J. Sci. Comput., 15 (1994), pp. 348–358.
[9] A. Greenbaum and L. N. Trefethen, *GMRES/CR and Arnoldi/Lanczos as matrix approximation problems*, SIAM J. Sci. Comput., 15 (1994), pp. 359–368.
[10] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bureau Standarts, 49 (1952), pp. 409–435.
[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 P. Tichý, *The worst-case GMRES for normal matrices*, BIT Numerical Mathematics, 44 (2004), pp. 79–98.
[13] J. C. Mason and D. C. Handscomb, *Chebyshev polynomials*, Chapman & Hall/CRC, Boca Raton, FL, 2003.
[14] A. E. Naiman, I. M. Babuška, and H. C. Elman, *A note on conjugate gradient convergence*, Numer. Math., 76 (1997), pp. 209–230.
[15] A. E. Naiman and S. Engelberg, *A note on conjugate gradient convergence. II, III*, Numer. Math., 85 (2000), pp. 665–683, 685–696.
[16] C. C. Paige and M. A. Saunders, *Solutions of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.
[17] 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.
[18] 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.
[19] K. Toh, M. Todd, and R. Tütüncü, *SDPT3 – a Matlab software package for semidefinite programming, version 2.1. Interior point methods.* The software is available on the site `http://www.math.nus.edu.sg/~mattohkc/sdpt3.html`, June 2001.
[20] Waterloo Maple Inc., *Maple 8.0.* Copyright (c) 1981-2002 by Waterloo Maple Inc. Maple and Maple V are registered trademarks of Waterloo Maple Inc., 2002.