Numer. Math. (2009) 112:267–293
DOI 10.1007/s00211-008-0206-2

Numerische
Mathematik

The rate of convergence of GMRES on a tridiagonal Toeplitz linear system

Ren-Cang Li · Wei Zhang

Received: 1 November 2007 / Revised: 6 August 2008 / Published online: 19 December 2008

Abstract The Generalized Minimal Residual method (GMRES) is often used to solve a nonsymmetric linear system $Ax = b$. But its convergence analysis is a rather difficult task in general. A commonly used approach is to diagonalize $A = X\Lambda X^{-1}$ and then separate the study of GMRES convergence behavior into optimizing the condition number of X and a polynomial minimization problem over A's spectrum. This artificial separation could greatly overestimate GMRES residuals and likely yields error bounds that are too far from the actual ones. On the other hand, considering the effects of both A's spectrum and the conditioning of X at the same time poses a difficult challenge, perhaps impossible to deal with in general but only possible for certain particular linear systems. This paper will do so for a (nonsymmetric) tridiagonal Toeplitz system. Sharp error bounds on and sometimes exact expressions for residuals are obtained. These expressions and/or bounds are in terms of the three parameters that define A and Chebyshev polynomials of the first kind.

Mathematics Subject Classification (2000) 65F10

1 Introduction

The Generalized Minimal Residual method (GMRES) is often used to solve a nonsymmetric linear system $Ax = b$. The basic idea is to seek approximate solutions,

R.-C. Li (✉)
Department of Mathematics, University of Texas at Arlington,
P.O. Box 19408, Arlington, TX 76019, USA
e-mail: rcli@uta.edu

W. Zhang
Department of Mathematics, University of Kentucky,
Lexington, KY 40506, USA
e-mail: wzhang@ms.uky.edu

optimal in certain sense, from the so-called Krylov subspaces. Specifically, the kth approximation x_k is sought so that the kth residual $r_k = b - Ax_k$ satisfies [21] (without loss of generality, we take initially $x_0 = 0$ and thus $r_0 = b$).

$$\|r_k\|_2 = \min_{y \in \mathcal{K}_k} \|b - Ay\|_2,$$

where the *kth Krylov subspace* $\mathcal{K}_k \equiv \mathcal{K}_k(A, b)$ of A on b is defined as

$$\mathcal{K}_k \equiv \mathcal{K}_k(A, b) \stackrel{\text{def}}{=} \operatorname{span}\{b, Ab, \ldots, A^{k-1}b\}, \tag{1.1}$$

and generic norm $\|\cdot\|_2$ is the usual ℓ_2 norm of a vector or the spectral norm of a matrix.

This paper is concerned with the convergence analysis of GMRES on linear system $Ax = b$ whose coefficient matrix A is a (nonsymmetric) tridiagonal Toeplitz coefficient matrix:

$$A = \begin{pmatrix} \lambda & \mu & & \\ \nu & \ddots & \ddots & \\ & \ddots & \ddots & \mu \\ & & \nu & \lambda \end{pmatrix},$$

where λ, μ, ν are assumed nonzero and possibly complex. When $|\mu| = |\nu|$, A is normal, including symmetric and symmetric positive definite as subcases, and convergence analysis of GMRES (or the conjugate gradient method) on the corresponding linear systems has been well studied (see [14] and the references therein). This paper will focus on the nonsymmetric case: $|\mu| \neq |\nu|$. Throughout this paper, exact arithmetic is assumed, A is N-by-N, and k is GMRES iteration index. Since in exact arithmetic GMRES computes the exact solution in at most N steps, $r_N = 0$. For this reason, we restrict $k < N$ at all times.

Our first main contribution in this paper is the following error bound (Theorem 2.1)

$$\frac{\|r_k\|_2}{\|r_0\|_2} \le \sqrt{k+1} \left[\sum_{j=0}^{k} \zeta^{2j} \left| T_j(\tau) \right|^2 \right]^{-1/2}, \tag{1.2}$$

where $T_j(t)$ is the jth Chebyshev polynomial of the first kind, and

$$\xi = -\frac{\sqrt{\mu\nu}}{\nu}, \quad \tau = \frac{\lambda}{2\sqrt{\mu\nu}}, \quad \zeta = \min\{|\xi|, |\xi|^{-1}\}.$$

We will also prove that this upper bound is nearly achieved by $b = e_1$ (the first column of the identity matrix) when $|\xi| \le 1$ or by $b = e_N$ (the last column of the identity matrix) when $|\xi| \ge 1$. By "nearly achieved", we mean it is within a factor about at most $(k+1)^{3/2}$ of the exact residual ratios.

Our second main contribution is about the worst asymptotic speed of $\|r_k\|_2$ among all possible r_0. It is proven that (Theorem 2.2)

$$\lim_{k\to\infty}\inf_{N>k}\left[\sup_{r_0}\frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k}=\lim_{k\to\infty}\sup_{N>k}\left[\sup_{r_0}\frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k}=\min\left\{(\zeta\rho)^{-1},1\right\}, \quad (1.3)$$

where $\rho=\max\left\{\left|\tau+\sqrt{\tau^2-1}\right|,\left|\tau-\sqrt{\tau^2-1}\right|\right\}$. Equivalently, the limits in (1.3) can be interpreted as follows: *for any given* $\epsilon>0$, *there exists a positive integer* K *such that*

$$\left|\left[\sup_{r_0}\frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k}-\min\left\{(\zeta\rho)^{-1},1\right\}\right|<\epsilon \quad \text{for all } N>k\geq K. \quad (1.4)$$

It is worth mentioning that $\sup_{r_0}$ among all possible r_0 in (1.3) can be replaced by $\sup_{r_0\in\{e_1,e_N\}}$. A related work that also studied asymptotic speed of convergence but for the conjugate gradient method (CG) and special right-hand sides and $\lambda=2$ and $\mu=\nu=-1$ is [2], where that N/k remains constant is required as $k\to\infty$, which is a much stronger restriction than $N>k$ needed in (1.3) and (1.4).

A by-product of (1.3) is that the worst asymptotic speed can be separated into the factor $\zeta^{-1}\geq 1$ contributed by A's departure from normality and the factor ρ^{-1} contributed by A's eigenvalue distribution. Take, for example, $\lambda=0.5$, $\mu=-0.3$, and $\nu=0.7$ which was used in [4, p. 562] to explain the effect of non-normality on GMRES convergence. We have $(\zeta\rho)^{-1}=0.90672$, whereas in [4, p. 562] it is implied $\|r_k\|_2/\|r_0\|_2\leq(0.913)^k$ for $N=50$, which is rather good, considering that $N=50$ is rather small.

Tridiagonal Toeplitz linear systems naturally arise from a convection-diffusion model problem [19]. When μ or ν is zero, A becomes a Jordan block, and GMRES residual bounds by Ipsen [12] and Liesen and Strakoš [18] suggests possibly slow convergence unless λ is much bigger in magnitude relative to the nonzero number μ or ν. This is more or less expected because otherwise A is highly non-normal, a case GMRES usually does poorly. Liesen and Strakoš also presented a detailed analysis aiming at showing that GMRES for tiny $|\mu|$ behaves much like GMRES after setting μ to 0. Their general residual expression and bound, however, involve the pseudo-inverse and the smallest singular value, respectively, of a matrix that do not seem to lead to simple bounds like ours without further, likely complicated, analysis. Ernst [9], in our notation, obtained the following inequality: *if A's field of values does not contain the origin, then*

$$\frac{\|r_k\|_2}{\|r_0\|_2}\leq\left(|\xi|^k+|\xi|^{-k}\right)\frac{\widetilde{\rho}^k}{1-\widetilde{\rho}^{2k}}, \quad (1.5)$$

where $\widetilde{\rho}=\max\left\{\left|\widetilde{\tau}+\sqrt{\widetilde{\tau}^2-1}\right|,\left|\widetilde{\tau}-\sqrt{\widetilde{\tau}^2-1}\right|\right\}$ *and* $\widetilde{\tau}=\left[\cos\frac{\pi}{N+1}\right]^{-1}\tau$. Our bound (1.2) is comparable to Ernst's bound for large N. This can be seen by noting that for N large enough, $\widetilde{\tau}\approx\tau$ and $\widetilde{\rho}\approx\rho$, and that $T_j(\tau)\approx\frac{1}{2}\rho^j$ when $\rho>1$ and

$|\zeta|^{-k} \le |\xi|^k + |\xi|^{-k} \le 2|\zeta|^{-k}$. Ernst's bound also leads to

$$\limsup_{k\to\infty} \sup_{N>k} \left[\sup_{r_0} \frac{\|r_k\|_2}{\|r_0\|_2} \right]^{1/k} \le \min\left\{ (\zeta\rho)^{-1}, 1 \right\}. \tag{1.6}$$

In differentiating our contributions here from Ernst's, we use a different technique to arrive at (1.2) and (1.3). While our approach is not as elegant as Ernst's which was based on A's field of values (see also [5]), it allows us to establish both lower and upper bounds on relative residuals for special right-hand sides to conclude that our bound is nearly achieved. Also (1.3) is an equality while only an inequality (1.6) can be deduced from Ernst's bound and approach.

We also obtain residual bounds and exact expressions especially for right-hand sides $b = e_1$ and $b = e_N$ (Theorem 2.3). They suggest, besides the sharpness of (1.2), an interesting GMRES convergence behavior. For $b = e_1$, that $|\xi| > 1$ speeds up GMRES convergence, and in fact $\|r_k\|_2$ is roughly proportional to $|\xi|^{-k}$. So the bigger the $|\xi|$ is, the faster the convergence will be. Note as $|\xi|$ gets bigger, A gets further away from a normal matrix. Thus, loosely speaking, the non-normality contributes to the convergence rate in the positive way. Nonetheless this does not contradict our usual perception that high non-normality is bad for GMRES if the worst behavior of GMRES among all b is considered. This mystery can be best explained by looking at the extreme case: $|\xi| = \infty$, i.e., $\nu = 0$, for which $b = e_1$ is an eigenvector (and convergence occurs in just one step). In general for $\nu \ne 0$, as $|\xi|$ gets bigger and bigger, roughly speaking $b = e_1$ comes closer and closer to A's invariant subspaces of lower dimensions and consequently speedier convergence is witnessed. Similar comments apply to the case when $b = e_N$.

The rest of this paper is organized as follows. We state our main results in Sect. 2. Tedious proofs that rely on residual reformulation involving rectangular Vandermonde matrices and complicated analysis will be presented separately in Sect. 3. Exact residual norm formulas for two special right-hand sides $b = e_1$ and e_N are established in Sect. 4. Finally in Sect. 5 we present our concluding remarks.

Notation Throughout this paper, $\mathbb{K}^{n\times m}$ is the set of all $n \times m$ matrices with entries in $\mathbb{K}$, where $\mathbb{K}$ is $\mathbb{C}$ (the set of complex numbers) or $\mathbb{R}$ (the set of real numbers), $\mathbb{K}^n = \mathbb{K}^{n\times 1}$, and $\mathbb{K} = \mathbb{K}^1$.

I_n (or simply I if its dimension is clear from the context) is the $n \times n$ identity matrix, and e_j is its jth column. The superscript "$\cdot^*$" takes conjugate transpose while "$\cdot^{\mathrm{T}}$" takes transpose only. $\sigma_{\min}(X)$ denotes the smallest singular value of X.

We shall also adopt MATLAB-like convention to access the entries of vectors and matrices. The set of integers from i to j inclusive is $i : j$. For a vector u and a matrix X, $u_{(j)}$ is u's jth entry, $X_{(i,j)}$ is X's (i, j)th entry, $\mathrm{diag}(u)$ is the diagonal matrix with $(\mathrm{diag}(u))_{(j,j)} = u_{(j)}$; X's submatrices $X_{(k:\ell,i:j)}$, $X_{(k:\ell,:)}$, and $X_{(:,i:j)}$ consists of intersections of row k to row ℓ and column i to column j, row k to row ℓ and all columns, and all rows and column i to column j, respectively. Finally $\|\cdot\|_p$ $(1 \le p \le \infty)$ is the ℓ_p norm of a vector or the ℓ_p operator norm of a matrix, defined as

$$\|u\|_p = \left(\sum_j |u_{(j)}|^p\right)^{1/p}, \quad \|X\|_p = \max_{\|u\|_p=1} \|Xu\|_p.$$

$\lfloor\alpha\rfloor$ be the largest integer that is no bigger than α, and $\lceil\alpha\rceil$ the smallest integer that is no less than α.

2 Main results

2.1 Setting the stage

An $N \times N$ tridiagonal Toeplitz A takes this form

$$A = \begin{pmatrix} \lambda & \mu & & \\ \nu & \ddots & \ddots & \\ & \ddots & \ddots & \mu \\ & & \nu & \lambda \end{pmatrix} \in \mathbb{C}^{N\times N}. \tag{2.1}$$

Throughout the rest of this paper, ν, λ, and μ are reserved as the defining parameters of A, and set

$$\xi = -\frac{\sqrt{\mu\nu}}{\nu}, \quad \tau = \frac{\lambda}{2\sqrt{\mu\nu}}, \quad \zeta = \min\left\{|\xi|, \frac{1}{|\xi|}\right\}, \tag{2.2}$$

$$\rho = \max\left\{\left|\tau + \sqrt{\tau^2-1}\right|, \left|\tau - \sqrt{\tau^2-1}\right|\right\}. \tag{2.3}$$

Matrix A is diagonalizable when $\mu \neq 0$ and $\nu \neq 0$. In fact [22, pp. 113–115] (see also [9,18]),

$$A = X\Lambda X^{-1}, \quad X = SZ, \quad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_N), \tag{2.4}$$

$$\lambda_j = \lambda - 2\sqrt{\mu\nu}\, t_j, \quad t_j = \cos\theta_j, \quad \theta_j = \frac{j\pi}{N+1}, \tag{2.5}$$

$$Z_{(:,j)} = \sqrt{\frac{2}{N+1}}\,(\sin j\theta_1, \ldots, \sin j\theta_N)^{\mathrm{T}}, \tag{2.6}$$

$$S = \mathrm{diag}(1, \xi^{-1}, \ldots, \xi^{-N+1}). \tag{2.7}$$

It can be verified that $Z^{\mathrm{T}}Z = I_N$; So A is normal if $|\xi| = 1$, i.e., $|\mu| = |\nu| > 0$. Set

$$\omega = -2\sqrt{\mu\nu}.$$

By (2.5), we have

$$\lambda_j = \omega(t_j - \tau), \quad 1 \le j \le N. \tag{2.8}$$

Any branch of $\sqrt{\mu\nu}$, once picked and fixed, is a valid choice in this paper. Note $\rho \ge 1$ always because $(\tau + \sqrt{\tau^2-1})(\tau - \sqrt{\tau^2-1}) = 1$. In particular if $\lambda \in \mathbb{R}$, $\mu < 0$ and $\nu > 0$, then $\rho = |\tau| + \sqrt{|\tau|^2+1}$.

A common starting point in existing quantitative analysis for GMRES [10, Page 54] on $Ax = b$ with diagonalizable $A = X\Lambda X^{-1}$ and $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_N)$ is

$$\|r_k\|_2/\|r_0\|_2 \le \kappa(X) \min_{\phi_k(0)=1} \max_i |\phi_k(\lambda_i)|. \tag{2.9}$$

as it seems that there is no easy way to do otherwise. It simplifies the analysis by separating the study of GMRES convergence behavior into optimizing the condition number of the eigenvector matrix X and a polynomial minimization problem over A's spectrum, but it could potentially overestimate GMRES residuals. This is partly because, as observed by Liesen and Strakoš [18], possible cancelations of huge components in X and/or X^{-1} were artificially ignored for the sake of the convergence analysis. For tridiagonal Toeplitz matrix A we are interested in here, however, rich structure allows us to do differently, as we shall do later.

Let $V_{k+1,N}$ be the $(k+1) \times N$ rectangular Vandermonde matrix

$$V_{k+1,N} \stackrel{\text{def}}{=} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \cdots & \lambda_N \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_1^k & \lambda_2^k & \cdots & \lambda_N^k \end{pmatrix} \tag{2.10}$$

having nodes $\{\lambda_j\}_{j=1}^N$, and

$$Y = X \operatorname{diag}(X^{-1}b). \tag{2.11}$$

Using $(b, Ab, \dots, A^k b) = Y\, V_{k+1,N}^{\mathrm{T}}$ [25, Lemma 2.1], we have for GMRES

$$\|r_k\|_2 = \min_{u_{(1)}=1} \|(b, Ab, \dots, A^k b)u\|_2 = \min_{u_{(1)}=1} \|Y\, V_{k+1,N}^{\mathrm{T}} u\|_2. \tag{2.12}$$

Recall Chebyshev polynomials of the first kind:

$$T_m(t) = \cos(m \arccos t) \qquad \text{for real } t \text{ and } |t| \le 1, \tag{2.13}$$

$$= \frac{1}{2}\left(t + \sqrt{t^2-1}\right)^m + \frac{1}{2}\left(t - \sqrt{t^2-1}\right)^m, \tag{2.14}$$

and define the mth *Translated Chebyshev Polynomial* in z of degree m as

$$T_m(z;\omega,\tau) \stackrel{\text{def}}{=} T_m(z/\omega + \tau) \tag{2.15}$$

$$= a_{mm} z^m + a_{m-1\,m} z^{m-1} + \cdots + a_{1m} z + a_{0m}, \tag{2.16}$$

where $a_{jm} \equiv a_{jm}(\omega, \tau)$ are functions of ω and τ, and upper triangular $R_m \in \mathbb{C}^{m\times m}$, a matrix-valued function in ω and τ, too, as

$$R_m \equiv R_m(\omega,\tau) \stackrel{\text{def}}{=} \begin{pmatrix} a_{00} & a_{01} & a_{02} & \cdots & a_{0\,m-1} \\ & a_{11} & a_{12} & \cdots & a_{1\,m-1} \\ & & a_{22} & \cdots & a_{2\,m-1} \\ & & & \ddots & \vdots \\ & & & & a_{m-1\,m-1} \end{pmatrix}, \tag{2.17}$$

i.e., the jth column consists of the coefficients of $T_{j-1}(z;\omega,\tau)$. Set

$$\boldsymbol{T}_N \stackrel{\text{def}}{=} \begin{pmatrix} T_0(t_1) & T_0(t_2) & \cdots & T_0(t_N) \\ T_1(t_1) & T_1(t_2) & \cdots & T_1(t_N) \\ \vdots & \vdots & & \vdots \\ T_{N-1}(t_1) & T_{N-1}(t_2) & \cdots & T_{N-1}(t_N) \end{pmatrix} \tag{2.18}$$

and $V_N = V_{N,N}$ for short. Then

$$V_N^{\mathrm{T}} R_N = \boldsymbol{T}_N^{\mathrm{T}}. \tag{2.19}$$

Equation (2.19) yields $V_N^{\mathrm{T}} = \boldsymbol{T}_N^{\mathrm{T}} R_N^{-1}$. Extracting the first $k+1$ columns from both sides of $V_N^{\mathrm{T}} = \boldsymbol{T}_N^{\mathrm{T}} R_N^{-1}$ yields

$$V_{k+1,N}^{\mathrm{T}} = \boldsymbol{T}_{k+1,N}^{\mathrm{T}}\, R_{k+1}^{-1}, \tag{2.20}$$

where $\boldsymbol{T}_{k+1,N} = (\boldsymbol{T}_N)_{(1:k+1,:)}$. Now notice $Y = X\,\mathrm{diag}(X^{-1}b)$ and $X = SZ$ with Z in (2.6) being real and orthogonal to get

$$\begin{aligned} Y\,V_{k+1,N}^{\mathrm{T}} &= SZ\,\mathrm{diag}(Z^{\mathrm{T}}S^{-1}b)\,(\boldsymbol{T}_N^{\mathrm{T}})_{(:,1:k+1)} R_{k+1}^{-1} \\ &= SM_{(:,1:k+1)} R_{k+1}^{-1} \qquad (2.21)\\ &= SM_{(:,1:k+1)} S_{k+1}^{-1}\, S_{k+1} R_{k+1}^{-1}, \qquad (2.22) \end{aligned}$$

where $S_{k+1} = S_{(1:k+1,1:k+1)}$, the $(k+1)$th leading principle submatrix of S,

$$M = Z\,\mathrm{diag}(Z^{\mathrm{T}}S^{-1}b)\,\boldsymbol{T}_N^{\mathrm{T}}. \tag{2.23}$$

It follows from (2.12) and (2.22) that

$$\sigma_{\min}(SM_{(:,1:k+1)}S_{k+1}^{-1}) \le \frac{\|r_k\|_2}{\min_{u_{(1)}=1}\|S_{k+1}R_{k+1}^{-1}u\|_2} \le \|SM_{(:,1:k+1)}S_{k+1}^{-1}\|_2. \tag{2.24}$$

The second inequality in (2.24) is our foundation to bound GMRES residuals for a general right-hand side, through explicitly expressing $\min_{u_{(1)}=1}\|S_{k+1}R_{k+1}^{-1}u\|_2$ in terms of

Chebyshev polynomials and the parameters τ and ξ and bounding $\|SM_{(:,1:k+1)}S_{k+1}^{-1}\|_2$ in terms of $\|b\|_2$.

2.2 An equivalence principle

We shall now describe an equivalence principle between the linear system with A and b and the one with A^{T} and a permuted b. The principle will simplify our analysis by allowing us to focus on the case $|\xi| \le 1$ only. Let $\Pi = (e_N, \dots, e_2, e_1) \in \mathbb{R}^{N\times N}$ be the permutation matrix. Notice $\Pi^{\mathrm{T}} A \Pi = A^{\mathrm{T}}$ and thus $Ax = b$ is equivalent to

$$A^{\mathrm{T}}\Pi^{\mathrm{T}}x = (\Pi^{\mathrm{T}}A\Pi)(\Pi^{\mathrm{T}}x) = \Pi^{\mathrm{T}}b. \tag{2.25}$$

Note $\mathcal{K}_k(A^{\mathrm{T}}, \Pi^{\mathrm{T}}b) = \mathcal{K}_k(\Pi^{\mathrm{T}}A\Pi, \Pi^{\mathrm{T}}b) = \Pi^{\mathrm{T}}\mathcal{K}_k(A, b)$, and

$$\begin{aligned} \|r_k\|_2 &= \min_{y\in\mathcal{K}_k(A,b)} \|b - Ay\|_2 \\ &= \min_{\Pi^{\mathrm{T}}y\in\Pi^{\mathrm{T}}\mathcal{K}_k(A,b)} \|\Pi^{\mathrm{T}}(b - A\Pi\,\Pi^{\mathrm{T}}y)\|_2 \\ &= \min_{w\in\mathcal{K}_k(A^{\mathrm{T}},\Pi^{\mathrm{T}}b)} \|\Pi^{\mathrm{T}}b - A^{\mathrm{T}}w\|_2. \end{aligned} \tag{2.26}$$

The last expression is the GMRES residual for the equivalent system (2.25). Therefore

> *Any result on GMRES for $Ax = b$ leads to one for $A^{\mathrm{T}}y = \Pi^{\mathrm{T}}b$ after performing the following substitutions*
>
> $$\mu \leftarrow \nu,\ \nu \leftarrow \mu,\ \xi \leftarrow \xi^{-1},\ b \leftarrow \Pi^{\mathrm{T}}b. \tag{2.27}$$

2.3 General right-hand sides

Our first main result is given in Theorem 2.1 whose proof, along with the proofs of other results in the section, involve complicated computations and will be postponed to Sect. 3.

Theorem 2.1 *For $Ax = b$, where A is tridiagonal Toeplitz as in (2.1) with nonzero (real or complex) parameters ν, λ, and μ. Then the kth GMRES residual r_k satisfies for $1 \le k < N$*

$$\frac{\|r_k\|_2}{\|r_0\|_2} \le \sqrt{k+1}\left[\frac{1}{2} + \Phi_{k+1}(\tau,\zeta)\right]^{-1/2}, \tag{2.28}$$

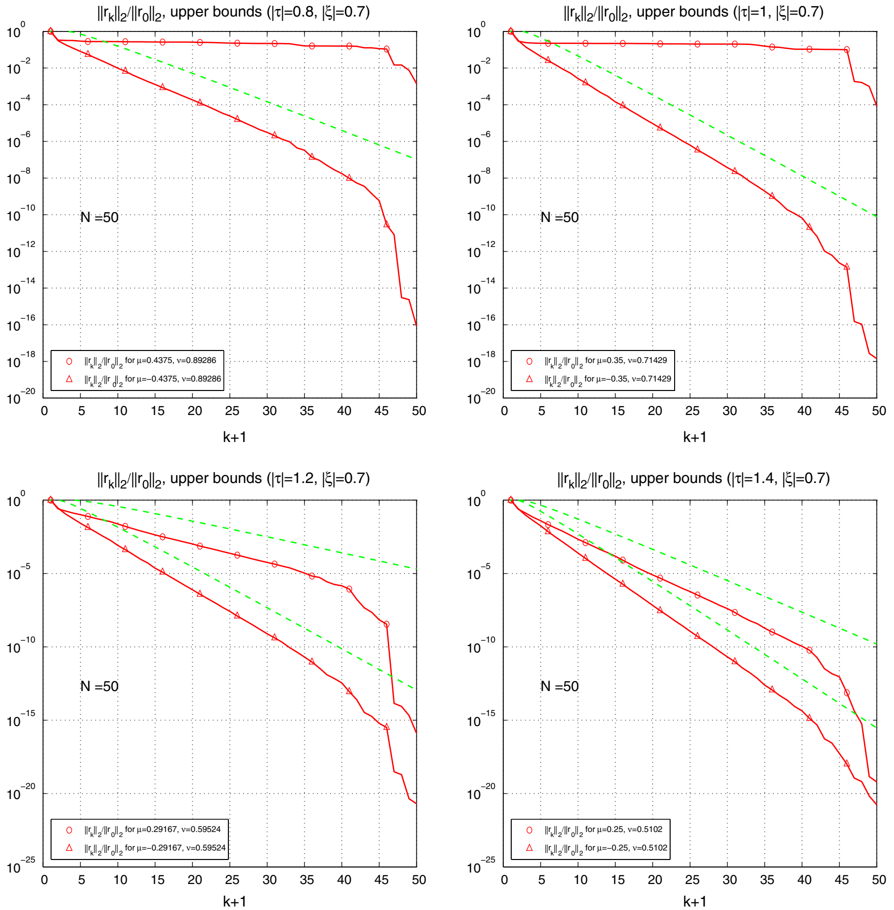

Fig. 1 GMRES residuals for random b uniformly in $[-1, 1]$, and their upper bounds (*dashed lines*) by (2.28). All indicate that our upper bounds are tight, except for the last few steps. Upper bounds for the case $\mu > 0$ in the top two plots are visually indistinguishable from the *horizonal line* 10^0, suggesting slow convergence. Only $|\xi| < 1$ is shown because of (2.27)

where

$$\Phi_{k+1}(t, s) \stackrel{\text{def}}{=} \sum_{j=0}^{k}{}' \, |s|^{2j} \left|T_j(t)\right|^2, \tag{2.29}$$

and $\sum_j'$ *means the first term is halved.*

Figure 1 plots residual histories for several examples of GMRES with each of b's entries being uniformly random in $[-1, 1]$. In each of the plots in Fig. 1, as well as in

Figs. 2 and 3, we fix $|\tau|$ and $|\xi|$, take $\lambda = 1$ always, and then take

$$|\mu| = \frac{|\xi|}{2|\tau|}, \quad \mu = \pm|\mu|, \quad \text{and} \quad \nu = |\nu| = \frac{1}{2|\tau\xi|}.$$

Thus $\mu, \nu \in \mathbb{R}$, and in fact $\nu > 0$ always. When $\mu > 0$, $\xi = -\sqrt{\mu/\nu} < 0$ and $\tau = 1/(2\sqrt{\mu\nu}) > 0$, but when $\mu < 0$, both $\xi = -\iota\sqrt{|\mu/\nu|}$ and $\tau = -\iota/(2\sqrt{|\mu\nu|})$ are imaginary, where $\iota = \sqrt{-1}$ is the imaginary unit. Figure 1 indicates that GMRES converges much faster for $\mu < 0$ than for $\mu > 0$ in each of the plots. There is a simple explanation for this: the eigenvalues of A are further away from the origin for a pure imaginary τ than for a real τ for any fixed $|\tau|$.

Our next main result given in Theorem 2.2 tells the worst asymptotic speed for $\|r_k\|_2$.

Theorem 2.2 *Under the conditions of* Theorem 2.1,

$$\lim_{k\to\infty} \inf_{N>k} \left[\sup_{r_0} \frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k} = \lim_{k\to\infty} \sup_{N>k} \left[\sup_{r_0} \frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k} = \min\{(\zeta\rho)^{-1}, 1\}. \tag{2.30}$$

Moreover, $\sup_{r_0}$ *among all possible* r_0 *can be replaced by* $\sup_{r_0\in\{e_1,e_N\}}$.

2.4 Special right-hand sides

We now consider three special right-hand sides: $b = e_1$ or e_N or $b_{(1)}e_1 + b_{(N)}e_N$. In particular they show that the upper bound in Theorem 2.1 is within a factor about at most $(k+1)^{3/2}$ of the true residual for $b = e_1$ or e_N, depending on whether $|\xi| \le 1$ or $|\xi| \ge 1$.

Because of the equivalence principle (2.27), only $b = e_1$ will be considered. Inequalities in Theorems 2.3 and 2.5 upon replacing $|\xi|$ by $|\xi|^{-1}$ lead to related results for the case $b = e_N$.

Theorem 2.3 *In* Theorem 2.1, *if* $b = e_1$, *then the kth GMRES residual* r_k *satisfies for* $1 \le k < N$

$$\frac{1}{2}\left[\sum_{j=0}^{\lceil\frac{k+1}{2}\rceil-1} |\xi|^{2j}\right]^{-1} \left[\Phi_{k+1}(\tau,\xi) - \frac{1}{4}\right]^{-1/2} \le \|r_k\|_2$$

$$\le \frac{1}{2}(1+|\xi|^2)\left[\Phi_{k+1}(\tau,\xi) - \frac{1}{4}\right]^{-1/2}. \tag{2.31}$$

In particular, for $|\xi| \le 1$

$$\frac{1}{2\lceil\frac{k+1}{2}\rceil}\left[\Phi_{k+1}(\tau,\xi) - \frac{1}{4}\right]^{-1/2} \le \|r_k\|_2 \le \left[\Phi_{k+1}(\tau,\xi) - \frac{1}{4}\right]^{-1/2}. \tag{2.32}$$

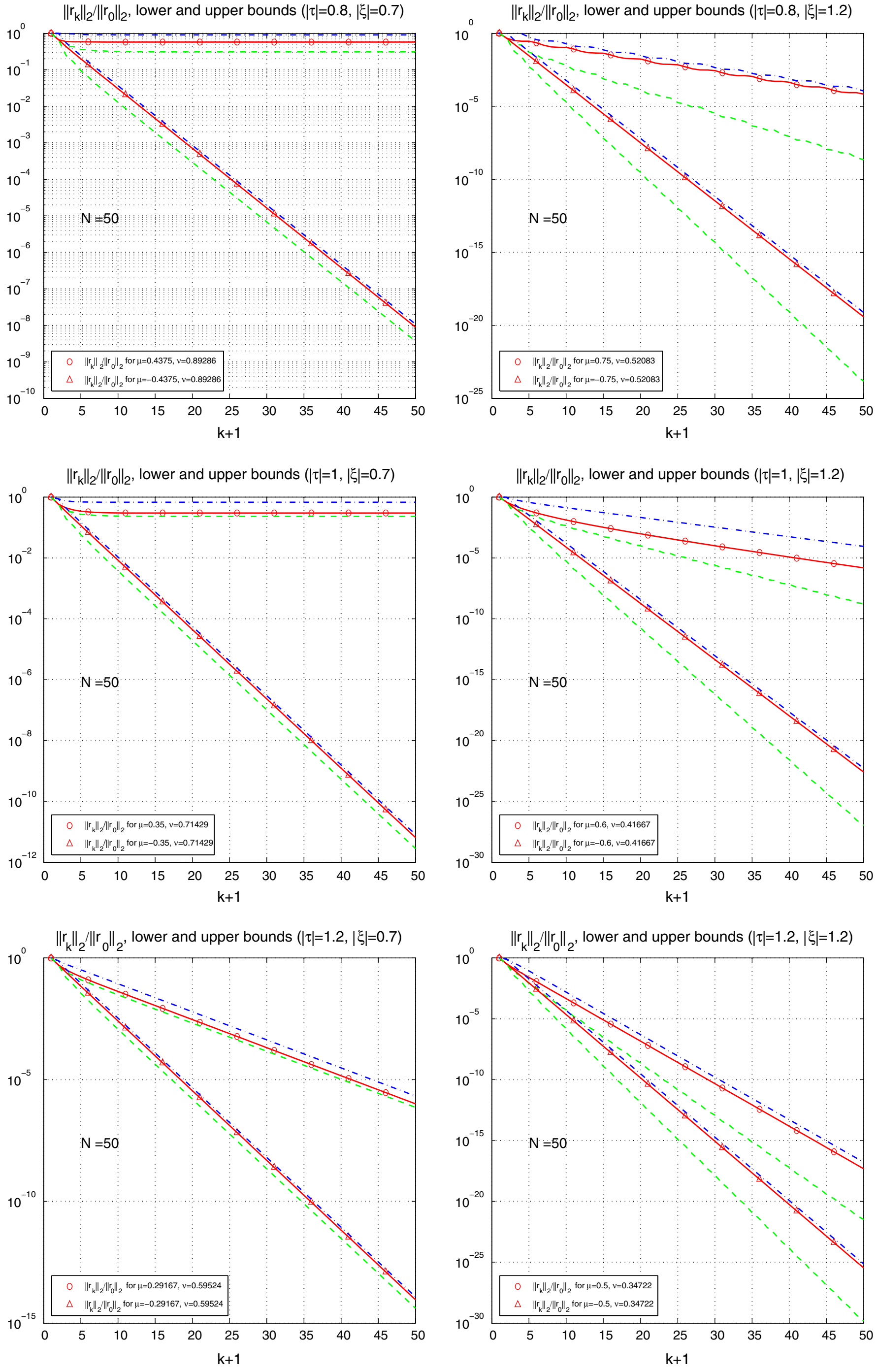

Fig. 2 GMRES residuals for $b = e_1$, sandwiched by their lower and upper bounds by (2.31). All lower and upper bounds are very good for $|\xi| \leq 1$ as expected, but only upper bounds are good when $|\xi| > 1$

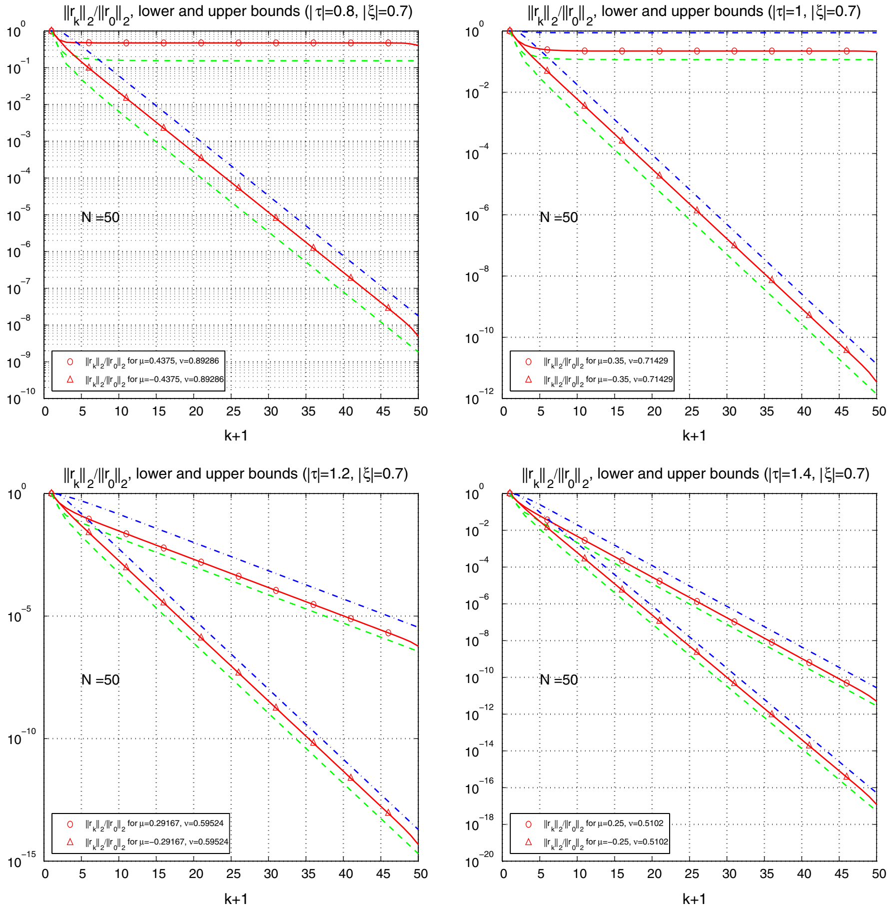

Fig. 3 GMRES residuals for $b = e_1 + e_N$, sandwiched by their lower and upper bounds by (2.33) and (2.34). Strictly speaking, (2.33) is only proved for $k \le N/2$, but it seems to be very good even for $k > N/2$ as well. We also ran GMRES for $b = e_1 - e_N$ and obtained residual history that is very much the same. Only $|\xi| < 1$ is shown because of (2.27)

The upper bound and the lower bound in (2.32) differ by a factor roughly $(k+1)$, and thus they are rather sharp; so are the bounds in (2.31) for $|\xi| \le 1$. Comparing them to (2.28), we conclude that the upper bound by (2.28) is fairly sharp for worst possible b.

But the bounds in (2.31) differ by a factor $\mathcal{O}(|\xi|^{k+1})$ for $|\xi| > 1$, and thus at least one of them (upper or lower bound) is bad. Our numerical examples indicate that the upper bounds are rather good regardless of the magnitude of $|\xi|$ for $b = e_1$ (see Fig. 2).

Given Theorem 2.3 and its implied result for $b = e_N$ by (2.27), it would not be unreasonable to expect that the upper bound in Theorem 2.1 would be sharp for very large or tiny $|\xi|$ within a factor possibly about at most $(k+1)^{3/2}$ for right-hand side b with $b_{(i)} = 0$ for $2 \le i \le N-1$ and $|b_{(1)}| = |b_{(N)}| > 0$. The following theorem indeed confirms this but only for $k \le N/2$. Our numerical examples even support that the lower bounds by (2.33) would be good for $k > N/2$ (see Fig. 3), too, but we do not have a way to mathematically justify it yet.

Theorem 2.4 *In* Theorem 2.1, *if* $b_{(i)} = 0$ *for* $2 \le i \le N-1$, *then the kth GMRES residual* r_k *satisfies for* $1 \le k \le N/2 - 1$

$$\frac{\|r_k\|_2}{\|r_0\|_2} \ge \frac{\min_{i\in\{1,N\}} |b_{(i)}|}{2\chi \, \|r_0\|_2} \left[\Phi_{k+1}(\tau,\zeta) - \frac{1}{4}\right]^{-1/2}, \tag{2.33}$$

and for $1 \le k < N$

$$\frac{\|r_k\|_2}{\|r_0\|_2} \le \sqrt{3}\left[\frac{1}{2} + \Phi_{k+1}(\tau,\zeta)\right]^{-1/2}, \tag{2.34}$$

where

$$1 < \chi = \sum_{j=0}^{\lceil\frac{k+1}{2}\rceil-1} \zeta^{2j} \le \left\lceil \frac{k+1}{2} \right\rceil.$$

Figures 2 and 3 plot residual histories for several examples of GMRES with $b = e_1$ and $b = e_1 + e_N$, respectively. Finally we have the following theorem about the asymptotic speeds of $\|r_k\|_2$ for $b = e_1$.

Theorem 2.5 *Assume the conditions of* Theorem 2.1 *hold. Let* $b = e_1$. *Then*

$$\begin{aligned}\min\{(|\xi|^2\rho)^{-1}, (|\xi|\rho)^{-1}, 1\} &\le \lim_{k\to\infty}\inf\ \inf_{N>k} \|r_k\|_2^{1/k} \\ &\le \lim_{k\to\infty}\sup\ \sup_{N>k} \|r_k\|_2^{1/k} \\ &\le \min\{(|\xi|\rho)^{-1}, 1\},\end{aligned} \tag{2.35}$$

and

$$\lim_{k\to\infty}\inf_{N>k} \|r_k\|_2^{1/k} = \lim_{k\to\infty}\sup_{N>k} \|r_k\|_2^{1/k} = \min\{(|\xi|\rho)^{-1}, 1\} \quad \textit{for } |\xi| \le 1. \tag{2.36}$$

Remark 1 As we commented before, our numerical examples indicate that the upper bounds in Theorem 2.3 is rather accurate regardless of the magnitude of $|\xi|$ for $b = e_1$ (see Fig. 2). This leads us to conjecture that the following equations

$$\lim_{k\to\infty}\inf_{N>k} \|r_k\|_2^{1/k} = \lim_{k\to\infty}\sup_{N>k} \|r_k\|_2^{1/k} = \min\{(|\xi|\rho)^{-1}, 1\} \quad \text{for } b = e_1, \tag{2.37}$$

$$\lim_{k\to\infty}\inf_{N>k} \|r_k\|_2^{1/k} = \lim_{k\to\infty}\sup_{N>k} \|r_k\|_2^{1/k} = \min\{(|\xi|^{-1}\rho)^{-1}, 1\} \quad \text{for } b = e_N, \tag{2.38}$$

would hold. As before if one of them is proven, the other one follows because of (2.27).

3 Proofs

In what follows, we will prove theorems in the previous section in the order of their appearance, except Theorem 2.2 whose proof requires Theorem 2.5 will be proved last.

In (2.24), there are two quantities to deal with

$$\min_{u_{(1)}=1} \|S_{k+1}R_{k+1}^{-1}u\|_2 \quad \text{and} \quad \|SM_{(:,1:k+1)}S_{k+1}^{-1}\|_2. \tag{3.1}$$

We shall now do so. In its present general form, the next lemma was proven in [13,16]. It was also implied by the proof of [12, Theorem 2.1]. See also [17].

Lemma 3.1 *If W has full column rank, then*

$$\min_{u_{(1)}=1} \|Wu\|_2 = \left[e_1^{\mathrm{T}}(W^*W)^{-1}e_1\right]^{-1/2}. \tag{3.2}$$

In particular if W is nonsingular, $\min_{u_{(1)}=1}\|Wu\|_2 = \|W^{-*}e_1\|_2^{-1}$.

By this lemma, we have (note $a_{00}=1$)

$$\min_{u_{(1)}=1} \|S_{k+1}R_{k+1}^{-1}u\|_2 = \|S_{k+1}^{-*}R_{k+1}^{*}e_1\|_2^{-1} = \left[\frac{1}{2} + \Phi_{k+1}(\tau,\xi)\right]^{-1/2}. \tag{3.3}$$

This gives the first quantity in (3.1). We now turn to the second one there. It can be seen that $SM_{(:,1:k+1)}S_{k+1}^{-1} = (SMS^{-1})_{(:,1:k+1)}$ since S is diagonal. To compute SMS^{-1}, we shall investigate M in (2.23) first.

$$\begin{aligned} M &= \sum_{\ell=1}^{N} Z\,\mathrm{diag}(ZS^{-1}b_{(\ell)}e_\ell)\,\boldsymbol{T}_N^{\mathrm{T}} \\ &= \sum_{\ell=1}^{N} b_{(\ell)}\xi^{\ell-1} Z\,\mathrm{diag}(Ze_\ell)\,\boldsymbol{T}_N^{\mathrm{T}} \\ &= \sum_{\ell=1}^{N} b_{(\ell)}\xi^{\ell-1} Z\,\mathrm{diag}(Z_{(:,\ell)})\,\boldsymbol{T}_N^{\mathrm{T}}. \end{aligned} \tag{3.4}$$

In Lemma 3.2 and in the proof of Lemma 3.3, without causing notational conflict, we will temporarily use k as a running index, as opposed to the rest of the paper where k is reserved for GMRES step index. The following lemma is probably well-known, but an exact source is hard to find. For a proof, see [15, Lemma 4.1].

Lemma 3.2 *For $\theta_j = \frac{j}{N+1}\pi$ and integer ℓ,*

$$\sum_{k=1}^{N} \cos \ell\theta_k = \begin{cases} N, & \text{if } \ell = 2m(N+1) \text{ for some integer } m, \\ 0, & \text{if } \ell \text{ is odd}, \\ -1, & \text{if } \ell \text{ is even, but } \ell \neq 2m(N+1) \text{ for any integer } m. \end{cases} \tag{3.5}$$

Lemma 3.3 *Let* $M_\ell \overset{\text{def}}{=} Z\,\mathrm{diag}(Z_{(:,\ell)})\boldsymbol{T}_N^{\mathrm{T}}$ *for* $1 \le \ell \le N$. *Then the entries of* M_ℓ *are zeros, except at those positions* (i, j)*, graphically forming four straight lines:*

$$\begin{aligned} &\text{(a) } i + j = \ell + 1, \\ &\text{(b) } i - j = \ell - 1, \\ &\text{(c) } j - i = \ell + 1, \\ &\text{(d) } i + j = 2(N+1) - \ell + 1. \end{aligned} \tag{3.6}$$

$(M_\ell)_{(i,j)} = 1/2$ *for* (a) *and* (b)*, except at their intersection* $(\ell, 1)$ *for which* $(M_\ell)_{(\ell,1)} = 1$. $(M_\ell)_{(i,j)} = -1/2$ *for* (c) *and* (d)*. Notice no valid entries for* (c) *if* $\ell \ge N - 1$ *and no valid entries for* (d) *if* $\ell \le 2$.

Proof For $1 \le i,\ j \le N$,

$$\begin{aligned} 2(N+1)\cdot(M_\ell)_{(i,j)} &= 4\sum_{k=1}^{N} \sin k\theta_i\ \sin \ell\theta_k\ \cos(j-1)\theta_k \\ &= 4\sum_{k=1}^{N} \sin i\theta_k\ \sin \ell\theta_k\ \cos(j-1)\theta_k \\ &= 2\sum_{k=1}^{N} [\cos(i-\ell)\theta_k - \cos(i+\ell)\theta_k]\cos(j-1)\theta_k \\ &= \sum_{k=1}^{N} [\cos(i+j-\ell-1)\theta_k + \cos(i-j-\ell+1)\theta_k \\ &\qquad - \cos(i+j+\ell-1)\theta_k - \cos(i-j+\ell+1)\theta_k]. \end{aligned}$$

Since all

$$\begin{aligned} i_1 &= i + j - \ell - 1, \\ i_2 &= i - j - \ell + 1, \\ i_3 &= i + j + \ell - 1, \\ i_4 &= i - j + \ell + 1 \end{aligned}$$

are either even or odd at the same time, Lemma 3.2 implies $(M_\ell)_{(i,j)} = 0$ unless one of them takes the form $2m(N+1)$ for some integer m. We now investigate all possible situations as such, keeping in mind that $1 \le i,\ j,\ \ell \le N$.

1. $i_1 = i + j - \ell - 1 = 2m(N+1)$. This happens if and only if $m = 0$, and thus $i + j = \ell + 1$. Then

   $$i_2 = -2j + 2, \quad i_3 = 2\ell, \quad i_4 = -2j + 2\ell + 2.$$

   They are all even. i_3 and i_4 do not take the form $2m(N+1)$ for some integers m. This is obvious for i_3, while $i_4 = 2m(N+1)$ implies $m = 0$ and $j = \ell + 1$, and thus $i = 0$ which cannot happen. However if $i_2 = 2m(N+1)$, then $m = 0$ and $j = 1$, and thus $i = \ell$.

So Lemma 3.2 implies $(M_\ell)_{(i,j)} = 1/2$ for $i + j = \ell + 1$ and $i \neq \ell$, while $(M_\ell)_{(\ell,1)} = 1$.

2. $i_2 = i - j - \ell + 1 = 2m(N+1)$. This happens if and only if $m = 0$, and thus $i - j = \ell - 1$. Then

$$i_1 = 2j - 2, \quad i_3 = 2j + 2\ell - 2, \quad i_4 = 2\ell.$$

They are all even. i_3 and i_4 do not take the form $2m(N+1)$ for some integers m. This is obvious for i_4, while $i_3 = 2m(N+1)$ implies $m = 1$ and $j = N + 2 - \ell$, and thus $i = N + 1$ which cannot happen. However if $i_1 = 2m(N+1)$, then $m = 0$ and thus $j = 1$ and $i = \ell$ which has already been considered in Item 1.
So Lemma 3.2 implies $(M_\ell)_{(i,j)} = 1/2$ for $i - j = \ell - 1$ and $i \neq \ell$, while $(M_\ell)_{(\ell,1)} = 1$.

3. $i_3 = i + j + \ell - 1 = 2m(N+1)$. This happens if and only if $m = 1$, and thus $i + j = 2(N+1) - \ell + 1$. Then

$$i_1 = 2(N+1) - 2\ell, \quad i_2 = 2(N+1) - 2j - 2\ell + 2, \quad i_4 = 2(N+1) - 2j + 2.$$

They are all even. i_1 and i_2 do not take the form $2m(N+1)$ for some integers m. This is obvious for i_1, while $i_2 = 2m(N+1)$ implies $m = 0$ and $j = N + 2 - \ell$, and thus $i = N + 1$ which cannot happen. However if $i_4 = 2m(N+1)$, then $m = 1$ and thus $j = 1$ and $i = 2(N+1) - \ell$ which is bigger than $N + 2$ and not possible.
So Lemma 3.2 implies $(M_\ell)_{(i,j)} = -1/2$ for $i + j = 2(N+1) - \ell + 1$.

4. $i_4 = i - j + \ell + 1 = 2m(N+1)$. This happens if and only if $m = 0$, and thus $j - i = \ell + 1$. Then

$$i_1 = 2j - 2\ell - 2, \quad i_2 = -2\ell, \quad i_3 = 2j - 2.$$

They are all even, and do not take the form $2m(N+1)$ for some integers m. This is obvious for i_2. $i_1 = 2m(N+1)$ implies $m = 0$ and $j = \ell + 1$, and thus $i = 0$ which cannot happen. $i_3 = 2m(N+1)$ implies $m = 0$ and thus $j = 1$ and $i = -\ell$ which cannot happen either.
So Lemma 3.2 implies $(M_\ell)_{(i,j)} = -1/2$ for $j - i = \ell + 1$.

This completes the proof. □

Now we know M_ℓ. We still need to find out SMS^{-1}. Let us examine it for $N = 5$ in order to get some idea about what it may look like. SMS^{-1} for $N = 5$ is

$$\begin{pmatrix}
b_{(1)} & \frac{1}{2}\xi^2 b_{(2)} & -\frac{1}{2}\xi^2 b_{(1)} + \frac{1}{2}\xi^4 b_{(3)} & -\frac{1}{2}\xi^4 b_{(2)} + \frac{1}{2}\xi^6 b_{(4)} & -\frac{1}{2}\xi^6 b_{(3)} + \frac{1}{2}\xi^8 b_{(5)} \\
b_{(2)} & \frac{1}{2} b_{(1)} + \frac{1}{2} b_{(3)}\xi^2 & \frac{1}{2} b_{(4)}\xi^4 & -\frac{1}{2}\xi^2 b_{(1)} + \frac{1}{2}\xi^6 b_{(5)} & -\frac{1}{2}\xi^4 b_{(2)} \\
b_{(3)} & \frac{1}{2} b_{(2)} + \frac{1}{2}\xi^2 b_{(4)} & \frac{1}{2} b_{(1)} + \frac{1}{2} b_{(5)}\xi^4 & 0 & -\frac{1}{2}\xi^2 b_{(1)} - \frac{1}{2}\xi^6 b_{(5)} \\
b_{(4)} & \frac{1}{2} b_{(3)} + \frac{1}{2} b_{(5)}\xi^2 & \frac{1}{2} b_{(2)} & \frac{1}{2} b_{(1)} - \frac{1}{2} b_{(5)}\xi^4 & -\frac{1}{2} b_{(4)}\xi^4 \\
b_{(5)} & \frac{1}{2} b_{(4)} & \frac{1}{2} b_{(3)} - \frac{1}{2} b_{(5)}\xi^2 & \frac{1}{2} b_{(2)} - \frac{1}{2}\xi^2 b_{(4)} & \frac{1}{2} b_{(1)} - \frac{1}{2} b_{(3)}\xi^2
\end{pmatrix}.$$

We observe that for $N = 5$, the entries of SMS^{-1} are polynomials in ξ with at most two terms. This turns out to be true for all N.

Lemma 3.4 *The following statements hold.*

1. *The first column of SMS^{-1} is b. Entries in every other columns taking one of the three forms: $(b_{(n_1)}\xi^{m_1} + b_{(n_2)}\xi^{m_2})/2$ with $n_1 \neq n_2$, $b_{(n_1)}\xi^{m_1}/2$, and 0, where $1 \leq n_1,\ n_2 \leq N$ and $m_i \geq 0$ are non-negative integer.*
2. *In each given column of SMS^{-1}, any particular entry of b appears at most twice.*

As the consequence, we have $\|SM_{(:,1:k+1)}S_{k+1}^{-1}\|_2 \leq \sqrt{k+1}\|b\|_2$ if $|\xi| \leq 1$.

Proof Notice $M = \sum_{\ell=1}^{N} b_{(\ell)}\xi^{\ell-1}M_\ell$ and consider M's $(i, j)th$ entry which comes from the contributions from all M_ℓ. But not all of M_ℓ contribute as most of them are zero at the position. Precisely, with the help of Lemma 3.3, those M_ℓ that contribute nontrivially to the $(i, j)th$ position are the following ones subject to satisfying the given inequalities.

(a) if $1 \leq i + j - 1 \leq N$ or equivalently $i + j \leq N + 1$, M_{i+j-1} gives a $1/2$.
(b) if $1 \leq i - j + 1 \leq N$ or equivalently $i \geq j$, M_{i-j+1} gives a $1/2$.
(c) if $1 \leq j - i - 1 \leq N$ or equivalently $j \geq i + 2$, M_{j-i-1} gives a $-1/2$.
(d) if $1 \leq 2(N+1)-(i+j)+1 \leq N$ or equivalently $i+j \geq N+3$, $M_{2(N+1)-(i+j)+1}$ gives a $-1/2$.

These inequalities, effectively 4 of them, divided entries of M into nine possible regions as detailed in Fig. 4. We shall examine each region one by one. Recall

$$(SMS^{-1})_{(i,j)} = \xi^{-i+1}M_{(i,j)}\xi^{j-1} = \xi^{j-i}M_{(i,j)},$$

and let

$$\begin{aligned}
\gamma_a &= \frac{1}{2}b_{(i+j-1)}\xi^{2j-2},\\
\gamma_b &= \frac{1}{2}b_{(i-j+1)},\\
\gamma_c &= -\frac{1}{2}b_{(j-i-1)}\xi^{2(j-i-1)},\\
\gamma_d &= -\frac{1}{2}b_{(2(N+1)-(i+j)+1)}\xi^{2(N+1)-(i+j)}.
\end{aligned}$$

Each entry in the nine possible regions in the rightmost plot of Fig. 4 is as follows.

1. (a) and (b): $(SMS^{-1})_{(i,j)} = \gamma_a + \gamma_b$.
2. (a) and (c): $(SMS^{-1})_{(i,j)} = \gamma_a + \gamma_c$.
3. (b) and (d): $(SMS^{-1})_{(i,j)} = \gamma_b + \gamma_d$.
4. (c) and (d): $(SMS^{-1})_{(i,j)} = \gamma_c + \gamma_d$.
5. (a) and $i - j = -1$: $(SMS^{-1})_{(i,j)} = \gamma_a$.
6. (b) and $i + j = N + 2$: $(SMS^{-1})_{(i,j)} = \gamma_b$.
7. (c) and $i + j = N + 2$: $(SMS^{-1})_{(i,j)} = \gamma_c$.

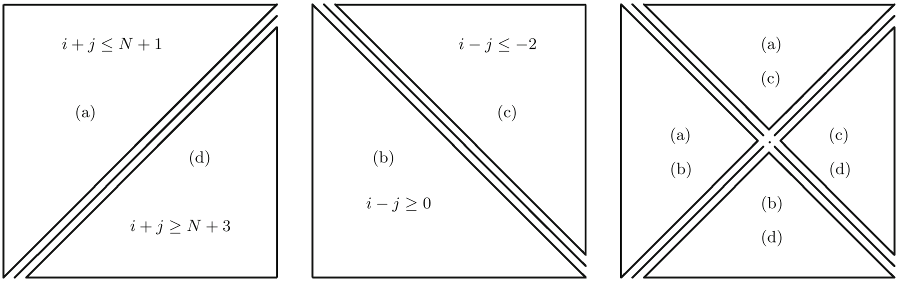

Fig. 4 Computation of $M_{(i,j)}$. *Left:* Regions of entries as divided by inequalities in (**a**) and (**d**); *Middle:* Regions of entries as divided by inequalities in (**b**) and (**c**); *Right:* Regions of entries as divided by all inequalities in (**a**), (**b**), (**c**), and (**d**)

8. (d) and $i - j = -1$: $(SMS^{-1})_{(i,j)} = \gamma_d$.
9. $i - j = -1$ and $i + j = N + 2$: $(SMS^{-1})_{(i,j)} = 0$. In this case, $i = (N+1)/2$ and $j = (N+3)/2$. So there is only one such entry when N is odd, and none when N is even.

With this profile on the entries of SMS^{-1}, we have Item 1 of the lemma immediately. Item 2 is the consequence of $M = \sum_{\ell=1}^{N} b_{(\ell)}\xi^{\ell-1}M_\ell$ and Lemma 3.3 which implies that there are at most two nonzero entries in each column of M_ℓ.

As the consequence of Item 1 and Item 2, each column of SMS^{-1} can be expressed as the sum of two vectors w and v such that $\|w\|_2,\ \|v\|_2 \le \|b\|_2/2$ when $|\xi| \le 1$, and thus $\|(SMS^{-1})_{(:,j)}\|_2 \le \|b\|_2$ for all $1 \le j \le N$. Therefore

$$\|SM_{(:,1:k+1)}S_{k+1}^{-1}\|_2 \le \sqrt{\sum_{j=1}^{k+1} \|(SMS^{-1})_{(:,j)}\|_2^2} \le \sqrt{k+1}\,\|b\|_2,$$

as expected. □

Proof of Theorem 2.1 We shall only prove

$$\|r_k\|_2 \le \|b\|_2\sqrt{k+1}\left[\frac{1}{2} + \Phi_{k+1}(\tau,\xi)\right]^{-1/2} \quad \text{for } |\xi| \le 1 \tag{3.7}$$

since for the other case when $|\xi| > 1$, we have, by (2.27) and (3.7) assuming proven true,

$$\|r_k\|_2 \le \|\Pi^{\mathrm{T}} b\|_2\sqrt{k+1}\left[\frac{1}{2} + \Phi_{k+1}(\tau,\xi^{-1})\right]^{-1/2}.$$

The conclusion of the theorem now follows from the observation that $\zeta = \min\{|\xi|, |\xi|^{-1}\}$.

Assume $|\xi| \le 1$. Inequality (3.7) is the consequence of (2.24), (3.3), and Lemma 3.4. □

Remark 2 The leftmost inequality in (2.24) gives a lower bound on $\|r_k\|_2$ in terms of $\sigma_{\min}(SM_{(:,1:k+1)}S_{k+1}^{-1})$ which, however, is hard to bound from below because it can be as small as zero, unless we know more about b such as $b = e_1$ as in Theorem 2.3.

Proof of Theorem 2.3 If $b = e_1$, then $M = M_1$ is upper triangular. More specifically

$$M = M_1 = \begin{pmatrix} 1 & 0 & -1/2 & & \\ & 1/2 & 0 & \ddots & \\ & & 1/2 & & -1/2 \\ & & & \ddots & 0 \\ & & & & 1/2 \end{pmatrix} \tag{3.8}$$

and, by (2.21),

$$Y V_{k+1,N}^{\mathrm{T}} = \begin{pmatrix} S_{k+1}\widetilde{M}R_{k+1}^{-1} \\ 0 \end{pmatrix} = \begin{pmatrix} S_{k+1}\widetilde{M}S_{k+1}^{-1}D^{-1} \times DS_{k+1}R_{k+1}^{-1} \\ 0 \end{pmatrix},$$

where $\widetilde{M} = M_{(1:k+1,1:k+1)}$ and $D = \mathrm{diag}(2, 1,1, \dots, 1)$. Therefore

$$\begin{aligned} \sigma_{\min}(S_{k+1}\widetilde{M}S_{k+1}^{-1}D^{-1}) &\le \frac{\min_{u_{(1)}=1} \|Y V_{k+1,N}^{\mathrm{T}} u\|_2}{\min_{u_{(1)}=1} \|DS_{k+1}R_{k+1}^{-1}u\|_2} \\ &\le \|S_{k+1}\widetilde{M}S_{k+1}^{-1}D^{-1}\|_2. \end{aligned} \tag{3.9}$$

Recall $\|r_k\|_2 = \min_{u_{(1)}=1} \|Y V_{k+1,N}^{\mathrm{T}} u\|_2$. Let

$$P_{k+1} = (e_1, e_3, \dots, e_2, e_4, \dots) \in \mathbb{R}^{(k+1)\times(k+1)}.$$

It can be seen that

$$P_{k+1}^{\mathrm{T}}(S_{k+1}\widetilde{M}S_{k+1}^{-1}D^{-1})P_{k+1} = \frac{1}{2}\begin{pmatrix} E_1 & \\ & E_2 \end{pmatrix},$$

where $E_1 \in \mathbb{C}^{\lceil \frac{k+1}{2} \rceil \times \lceil \frac{k+1}{2} \rceil}$, $E_2 \in \mathbb{C}^{\lfloor \frac{k+1}{2} \rfloor \times \lfloor \frac{k+1}{2} \rfloor}$, and

$$E_i = \begin{pmatrix} 1 & -\xi^2 & & \\ & 1 & \ddots & \\ & & \ddots & -\xi^2 \\ & & & 1 \end{pmatrix}, \quad E_i^{-1} = \begin{pmatrix} 1 & \xi^2 & \cdots & \xi^{2(m-1)} \\ & 1 & \ddots & \vdots \\ & & \ddots & \xi^2 \\ & & & 1 \end{pmatrix}.$$

Hence $\|E_i\|_2 \le \sqrt{\|E_i\|_1\|E_i\|_\infty} = 1 + |\xi|^2$. Therefore

$$\|S_{k+1}\widetilde{M}S_{k+1}^{-1}D^{-1}\|_2 = \frac{1}{2}\max\{\|E_1\|_2, \|E_2\|_2\} \le \frac{1}{2}(1 + |\xi|^2).$$

Similarly use $\|E_i^{-1}\|_2 \le \sqrt{\|E_i^{-1}\|_1 \|E_i^{-1}\|_\infty}$ to get

$$\|E_1^{-1}\|_2 \le \sum_{j=0}^{\lceil\frac{k+1}{2}\rceil-1} |\xi|^{2j}, \quad \|E_2^{-1}\|_2 \le \sum_{j=0}^{\lfloor\frac{k+1}{2}\rfloor-1} |\xi|^{2j}.$$

Therefore

$$\begin{aligned}\sigma_{\min}(S_{k+1}\widetilde{M}S_{k+1}^{-1}D^{-1}) &= \frac{1}{2}\min\{\sigma_{\min}(E_1), \sigma_{\min}(E_2)\} \\ &= \frac{1}{2}\min\{\|E_1^{-1}\|_2^{-1}, \|E_2^{-1}\|_2^{-1}\} \\ &\ge \frac{1}{2}\left[\sum_{j=0}^{\lceil\frac{k+1}{2}\rceil-1} |\xi|^{2j}\right]^{-1}.\end{aligned}$$

Finally, by Lemma 3.1, we have

$$\min_{u_{(1)}=1} \|DS_{k+1}R_{k+1}^{-1}u\|_2 = \|D^{-*}S_{k+1}^{-*}R_{k+1}^{*}e_1\|_2^{-1} = \left[\Phi_{k+1}(\tau,\xi) - \frac{1}{4}\right]^{-1/2}.$$

This, together with (3.9), lead to (2.31). □

Proof of Theorem 2.4 Now $b = b_{(1)}e_1 + b_{(N)}e_N$. Notice the form of M_1 in (3.8), and that M_N is M_1 after its rows reordered from the last to the first. For the case $M = b_{(1)}M_1 + \xi^{N-1}b_{(N)}M_N$, and also Lemma 3.4 implies that only positive powers of ξ appear in the entries of SMS^{-1}. Therefore when $|\xi| \le 1$,

$$\begin{aligned}\|SM_{(:,1:k+1)}S_{k+1}^{-1}\|_2 &\le \|SMS^{-1}\|_2 \\ &\le |b_{(1)}|\,\|\,|M_1|\,\|_2 + |b_{(N)}|\,\|\,|M_N|\,\|_2 \\ &\le |b_{(1)}|\sqrt{3/2} + |b_{(N)}|\sqrt{3/2} \\ &\le \sqrt{3}\|b\|_2, \end{aligned} \tag{3.10}$$

where $|M_\ell|$ takes entrywise absolute value, and we have used

$$\|\,|M_N|\,\|_2 = \|\,|M_1|\,\|_2 \le \sqrt{\|M_1\|_1\|M_1\|_\infty} = \sqrt{3/2}.$$

Inequality (2.34) for $|\xi| \le 1$ is the consequence of (2.24), (3.3), and (3.10). Inequality (2.34) for $|\xi| \ge 1$ follows from itself for $|\xi| \le 1$ applied to the permuted system (2.25).

To prove (2.33), we use the lines of arguments in the proof for Theorem 2.3 and notice that for $1 \le k \le N/2 - 1$

$$Y V_{k+1,N}^{\mathrm{T}} = \begin{matrix} k+1 \\ N-2(k+1) \\ k+1 \end{matrix} \overset{k+1}{\begin{pmatrix} W_1 \\ 0 \\ W_2 \end{pmatrix}}$$

It can be seen from the proof for Theorem 2.3 that

$$\min_{u_{(1)}=1} \|W_1 u\|_2 \ge \frac{|b_{(1)}|}{2} \left[\sum_{j=0}^{\lceil \frac{k+1}{2} \rceil - 1} |\xi|^{2j} \right]^{-1} \left[\Phi_{k+1}(\tau, \xi) - \frac{1}{4} \right]^{-1/2},$$

$$\min_{u_{(1)}=1} \|W_2 u\|_2 \ge \frac{|b_{(N)}|}{2} \left[\sum_{j=0}^{\lceil \frac{k+1}{2} \rceil - 1} |\xi|^{-2j} \right]^{-1} \left[\Phi_{k+1}(\tau, \xi^{-1}) - \frac{1}{4} \right]^{-1/2}.$$

Finally use

$$\min_{u_{(1)}=1} \|Y V_{k+1,N}^{\mathrm{T}} u\|_2 \ge \max \left\{ \min_{u_{(1)}=1} \|W_1 u\|_2, \min_{u_{(1)}=1} \|W_2 u\|_2 \right\}$$

to complete the proof. □

Proof of Theorem 2.5 We note that

$$\limsup_{k\to\infty} \sup_{N>k} \|r_k\|_2^{1/k} \le 1$$

because $\|r_k\|_2$ is nonincreasing. Suppose $b = e_1$. Consider first $\rho > 1$. Then $|T_j(\tau)| \sim \frac{1}{2}\rho^j$, and thus

$$\Phi_{k+1}(\tau, \xi) - \frac{1}{4} \sim \frac{1}{4} \sum_{j=0}^{k} (|\xi|\rho)^{2j} = \frac{1}{4} \cdot \frac{(|\xi|\rho)^{2(k+1)} - 1}{(|\xi|\rho)^2 - 1}. \tag{3.11}$$

If $|\xi|\rho > 1$, then (3.11) and Theorem 2.3 imply

$$\begin{aligned} \limsup_{k\to\infty} \sup_{N>k} \|r_k\|_2^{1/k} &\le \lim_{k\to\infty} \left[\frac{1}{2}(1 + |\xi|^2) \right]^{1/k} \\ &\quad \times \left[\Phi_{k+1}(\tau, \xi) - \frac{1}{4} \right]^{-1/(2k)} \\ &= (|\xi|\rho)^{-1}, \end{aligned} \tag{3.12}$$

$$\liminf_{k\to\infty} \inf_{N>k} \|r_k\|_2^{1/k} \geq \lim_{k\to\infty} \frac{1}{2^{1/k}} \left[\sum_{j=0}^{\lceil \frac{k+1}{2} \rceil -1} |\xi|^{2j} \right]^{-1/k} \times \left[\Phi_{k+1}(\tau, \xi) - \frac{1}{4} \right]^{-1/(2k)} = \begin{cases} (|\xi|\rho)^{-1}, & \text{for } |\xi| \leq 1, \\ (|\xi|^2\rho)^{-1}, & \text{for } |\xi| > 1. \end{cases} \tag{3.13}$$

They together give (2.35) for the case $|\xi|\rho > 1$. If $|\xi|\rho \leq 1$, then must $|\xi| < 1$ and $\min\{(|\xi|\rho)^{-1}, 1\} = 1$, $\min\{(|\xi|^2\rho)^{-1}, (|\xi|\rho)^{-1}, 1\} = 1$, and

$$\liminf_{k\to\infty} \inf_{N>k} \|r_k\|_2^{1/k} \geq 1$$

by (3.13) because $\Phi_{k+1}(\tau, \xi) - \frac{1}{4}$ is approximately bounded by $(k+1)/4$ by (3.11). So (2.35) holds for the case $|\xi|\rho \leq 1$, too.

Now consider $\rho = 1$. Then $\tau + \sqrt{\tau^2 - 1} = e^{\iota\theta}$ for some $0 \leq \theta \leq \pi$, where $\iota = \sqrt{-1}$ is the imaginary unit. Thus $\tau \in [-1, 1]$ and in fact

$$2\tau = (\tau + \sqrt{\tau^2 - 1}) + (\tau - \sqrt{\tau^2 - 1}) = 2\cos\theta, \quad T_j(\tau) = \cos j\theta.$$

Therefore $\Phi_{k+1}(\tau, \xi) - \frac{1}{4} \sim \frac{1}{4} + \sum_{j=1}^{k} |\xi|^{2j}(\cos j\theta)^2$. Instead of (3.12) and (3.13), we have

$$\limsup_{k\to\infty} \sup_{N>k} \|r_k\|_2^{1/k} \leq \eta_2, \quad \liminf_{k\to\infty} \inf_{N>k} \|r_k\|_2^{1/k} \geq |\xi|^{-1}\eta_1, \tag{3.14}$$

where

$$\eta_1 = \liminf_{k\to\infty} \left[\frac{1}{4} + \sum_{j=1}^{k} |\xi|^{2j}(\cos j\theta)^2 \right]^{-1/(2k)},$$

$$\eta_2 = \limsup_{k\to\infty} \left[\frac{1}{4} + \sum_{j=1}^{k} |\xi|^{2j}(\cos j\theta)^2 \right]^{-1/(2k)}.$$

Now if $|\xi| \leq 1$, then $\frac{1}{4} \leq \frac{1}{4} + \sum_{j=1}^{k} |\xi|^{2j}(\cos j\theta)^2 \leq \frac{1}{4} + k$ and thus

$$\eta_1 = \eta_2 = 1 \quad \text{if } |\xi| \leq 1. \tag{3.15}$$

If, however, $|\xi| > 1$, we claim that

$$\alpha|\xi|^{2(k-1)} \leq \frac{1}{4} + \sum_{j=1}^{k} |\xi|^{2j}(\cos j\theta)^2 \leq (k+1)|\xi|^{2k} \tag{3.16}$$

for some constant $\alpha > 0$, independent of θ and k. The second inequality in (3.16) is quite obvious. To see the first inequality, we notice

$$\begin{aligned}\sum_{j=k-1}^{k} |\xi|^{2j}(\cos j\theta)^2 &\geq |\xi|^{2(k-1)}[\cos^2 k\theta + \cos^2(k-1)\theta] \\ &= |\xi|^{2(k-1)}\left[\frac{1+\cos 2k\theta}{2} + \frac{1+\cos 2(k-1)\theta}{2}\right] \\ &= |\xi|^{2(k-1)}\,[1+\cos(2k-1)\theta\ \cos\theta]\,.\end{aligned}$$

It suffices to show that there is positive constant α, independent of k and θ, such that $1+\cos(2k-1)\theta\ \cos\theta \geq \alpha$. Assume to the contrary that

$$\inf_{k,\theta}\,[1+\cos(2k-1)\theta\ \cos\theta] = 0, \tag{3.17}$$

which means there are sequences $\{k_i\}$ and θ_i such that $1+\cos(2k_i-1)\theta_i\ \cos\theta_i \to 0$ as $i \to \infty$. Since $|\cos(\cdot)| \leq 1$, this implies as $i \to \infty$ either $\cos(2k_i-1)\theta_i \to 1$ and $\cos\theta_i \to -1$ or $\cos(2k_i-1)\theta_i \to -1$ and $\cos\theta_i \to 1$. But none of which is possible. Therefore (3.17) cannot hold. This proves (3.16) which yields

$$\eta_1 = \eta_2 = |\xi|^{-1} \quad \text{if } |\xi| > 1. \tag{3.18}$$

(2.35) for $\rho = 1$ is the consequence of (3.14), (3.15), and (3.18).

Finally if $|\xi| \leq 1$, then both the leftmost side and rightmost side of (2.35) are equal to $\min\{(|\xi|\rho)^{-1}, 1\}$. This proves (2.36). □

Proof of Theorem 2.2 Note that

$$\limsup_{k\to\infty}\,\sup_{N>k}\left[\sup_{r_0}\frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k} \leq 1.$$

Because of the equivalence principle (2.27), it suffices to consider only $|\xi| \leq 1$. Then $\zeta = |\xi|$.

First we prove

$$\limsup_{k\to\infty}\,\sup_{N>k}\left[\sup_{r_0}\frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k} \leq \min\{(|\xi|\rho)^{-1}, 1\}. \tag{3.19}$$

If $\rho = 1$, then $\min\{(|\xi|\rho)^{-1}, 1\} = 1$ because $|\xi|^{-1} \geq 1$; no proof is needed. If $\rho > 1$, then $|T_j(\tau)| \sim \frac{1}{2}\rho^j$, and thus (3.11). Now if $|\xi|\rho > 1$, then (3.11) and Theorem 2.1 imply

$$\limsup_{k\to\infty} \sup_{N>k} \left[\sup_{r_0} \frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k} \leq (|\xi|\rho)^{-1}$$

which also holds if $|\xi|\rho \leq 1$ because then $(|\xi|\rho)^{-1} \geq 1$.

Next we notice

$$\begin{aligned} \lim_{k\to\infty} \inf_{N>k} \left[\left.\frac{\|r_k\|_2}{\|r_0\|_2}\right|_{r_0=e_1} \right]^{1/k} &\leq \liminf_{k\to\infty} \inf_{N>k} \left[\sup_{r_0} \frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k} \\ &\leq \limsup_{k\to\infty} \sup_{N>k} \left[\sup_{r_0} \frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k}, \end{aligned} \tag{3.20}$$

where the first limit has been proven to exist and equal to $\min\{(|\xi|\rho)^{-1}, 1\}$ in Theorem 2.5. The proof is now completed by combining (3.19), (3.20), and (2.36). □

4 Exact residual norms for $b = e_1$ and e_N

In this section we present two theorems in which exact formulas for $\|r_k\|_2$ for $b = e_1$ and $b = e_N$ are established. Let S and S_{k+1} have their assignments as in Sect. 3.

Theorem 4.6 *In* Theorem 2.1, *if* $b = e_1$, *then the kth GMRES residual* r_k *satisfies for* $1 \leq k < N$

$$\|r_k\|_2 = \|2S_{k+1}^{-*} y_{(1:k+1)}\|_2^{-1}, \tag{4.1}$$

where $y \in \mathbb{C}^{2\lceil N/2\rceil}$ *is defined as*

$$y_{(2j-1)} = \sum_{i=1}^{j}{}' \, \bar{T}_{2i-2}(\tau), \quad y_{(2j)} = \sum_{i=1}^{j} \bar{T}_{2i-1}(\tau) \quad \textit{for } j = 1,2,\ldots,\lceil N/2\rceil,$$

and $\bar{T}_j(\tau)$ *is the complex conjugate of* $T_j(\tau)$.

Proof We still have (3.8), and $Y V_{k+1,N}^{\mathrm{T}} = \begin{pmatrix} S_{k+1}\widetilde{M} R_{k+1}^{-1} \\ 0 \end{pmatrix}$, where $\widetilde{M} = M_{(1:k+1,1:k+1)}$ as in the proof of Theorem 2.3. Let $D = \operatorname{diag}(2, 1, 1, \ldots, 1)$. Noticing $S_{k+1}\widetilde{M} R_{k+1}^{-1} = S_{k+1}\widetilde{M} D^{-1} \times D R_{k+1}^{-1}$ is nonsingular, we have by Lemma 3.1

$$\min_{u_{(1)}=1} \|Y V_{k+1,N}^{\mathrm{T}} u\|_2 = \|w\|_2^{-1},$$

where $w = S_{k+1}^{-*}(\widetilde{M}D^{-1})^{-\mathrm{T}} D^{-\mathrm{T}} R_{k+1}^{*} e_1$, or equivalently

$$(\widetilde{M}D^{-1})^{\mathrm{T}} S_{k+1}^{*} w = D^{-\mathrm{T}} R_{k+1}^{*} e_1.$$

We shall now solve it for w. Let

$$P_{k+1} = (e_1, e_3, \ldots, e_2, e_4, \ldots) \in \mathbb{R}^{(k+1)\times(k+1)}.$$

It can be verified that

$$P_{k+1}^{\mathrm{T}}(\widetilde{M}D^{-1})P_{k+1} = \frac{1}{2}\begin{pmatrix} G_1 & \\ & G_2 \end{pmatrix},$$

where $G_1 \in \mathbb{R}^{\lceil \frac{k+1}{2} \rceil \times \lceil \frac{k+1}{2} \rceil}$, $G_2 \in \mathbb{R}^{\lfloor \frac{k+1}{2} \rfloor \times \lfloor \frac{k+1}{2} \rfloor}$, and

$$G_i = \begin{pmatrix} 1 & -1 & & \\ & 1 & \ddots & \\ & & \ddots & -1 \\ & & & 1 \end{pmatrix}, \quad G_i^{-1} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ & 1 & \ddots & \vdots \\ & & \ddots & 1 \\ & & & 1 \end{pmatrix}.$$

Solve $(P_{k+1}^{\mathrm{T}} \widetilde{M} D^{-1} P_{k+1})^{\mathrm{T}} P_{k+1}^{\mathrm{T}} S_{k+1}^{*} w = P_{k+1}^{\mathrm{T}} D^{-\mathrm{T}} R_{k+1}^{*} e_1 \equiv P_{k+1}^{\mathrm{T}} z$ for w to get

$$w = 2 S_{k+1}^{-*} P_{k+1} \begin{pmatrix} G_1^{-\mathrm{T}} & \\ & G_2^{-\mathrm{T}} \end{pmatrix} P_{k+1}^{\mathrm{T}} z,$$

where $z = (\frac{1}{2}T_0(\tau), T_1(\tau), T_2(\tau), \ldots, T_k(\tau))^*$. Finally notice $w = 2S_{k+1}^{-*} y_{(1:k+1)}$ to complete the proof. □

Remark 3 For $b = e_N = \Pi^{\mathrm{T}} e_1$, apply Theorem 4.6 to the permuted system (2.25) to get the kth GMRES residual r_k satisfies for $1 \le k < N$

$$\|r_k\|_2 = \|2 S_{k+1}^{*} y_{(1:k+1)}\|_2^{-1}, \tag{4.2}$$

where $y \in \mathbb{C}^{2\lceil N/2 \rceil}$ is the same as the one in Theorem 4.6.

5 Concluding remarks

There are a few GMRES error bounds with simplicity comparable to the well-known bound for the conjugate gradient method [3,10,20,24]. In [6, Section 6], Eiermann and Ernst proved

$$\frac{\|r_k\|_2}{\|r_0\|_2} \le \left[1 - \gamma(A)\gamma(A^{-1})\right]^{k/2}, \tag{5.1}$$

where $\gamma(A) = \inf\{|z^* A z| : \|z\|_2 = 1\}$ is the distance from the origin to A's field of values. When A's Hermitian part, $H = (A + A^*)/2$, is positive definite, it yields a bound by Elman [8] (see also [7])

$$\frac{\|r_k\|_2}{\|r_0\|_2} \le \left[1 - \left(\frac{1}{\|H^{-1}\|_2 \|A\|_2}\right)^2\right]^{k/2}. \tag{5.2}$$

As observed in [1], this bound of Elman can be easily extended to cover the case when only $\gamma(A) > 0$

$$\frac{\|r_k\|_2}{\|r_0\|_2} \le (\sin\theta)^k, \quad \theta = \arccos\frac{\gamma(A)}{\|A\|_2}. \tag{5.3}$$

Recently Beckermann, Goreinov, and Tyrtyshnikov [1] improved (5.3) to

$$\frac{\|r_k\|_2}{\|r_0\|_2} \le (2 + 2/\sqrt{3})(2 + \delta)\delta^k, \quad \delta = 2\sin\frac{\theta}{4 - 2\theta/\pi}. \tag{5.4}$$

All three bounds (5.1), (5.3), and (5.4) yield meaningful estimates only when $\gamma(A) > 0$, i.e., A's field of values does not contain the origin.

However in general, there is not much concrete quantitative results for the convergence rate of GMRES, based on limited information on A and/or b. In part, it is a very difficult problem, and such a result most likely does not exist, thanks to the negative result of Greenbaum, Pták, and Strakoš [11] which says that "*Any Nonincreasing Convergence Curve is Possible for GMRES*". A commonly used approach, as a step towards getting a feel of how fast GMRES may be, is through assuming that A is diagonalizable to arrive at (2.9):

$$\|r_k\|_2/\|r_0\|_2 \le \kappa(X) \min_{\phi_k(0)=1} \max_i |\phi_k(\lambda_i)|, \tag{5.5}$$

and then putting aside the effect of $\kappa(X)$ to study only the effect in the factor of the associated minimization problem. This approach does not always yield satisfactory results, especially when $\kappa(X) \gg 1$ which occurs when A is highly non-normal. Getting a fairly accurate quantitative estimate for the convergence rate of GMRES for a highly non-normal case is likely to be very difficult. Trefethen [23] established residual bounds based on pseudospectra, which sometimes is more realistic than (5.5) but is very expensive to compute. In [4], Driscoll, Toh, and Trefethen provided an nice explanation on this matter.

Our analysis here on tridiagonal Toeplitz A represents one of few diagonalizable cases where one can analyze r_k directly to arrive at simple quantitative results such as (5.1)–(5.4). Previous other results except those in Ernst [9], while helpful in explaining and understanding various convergence behaviors, are more qualitative than quantitative.

Two conjectures are made in Remark 1. There is also an unsolved question whether (2.33) or a slightly modified version of it holds for $N/2 < k < N$ (see Fig. 3).

Acknowledgments The authors were supported in part by the National Science Foundation under Grant No. DMS-0510664 and DMS-0702335. The authors wish to thank an anonymous referee for his constructive suggestions that improve the presentation considerably.

References

1. Beckermann, B., Goreinov, S.A., Tyrtyshnikov, E.E.: Some remarks on the Elman estimate for GMRES. SIAM J. Matrix Anal. Appl. **27**, 772–778 (2006)
2. Beckermann, B., Kuijlaars, A.B.J.: Superlinear CG convergence for special right-hand sides. Electron. Trans. Numer. Anal. **14**, 1–19 (2002)
3. Demmel, J.: Applied Numerical Linear Algebra. SIAM, Philadelphia (1997)
4. Driscoll, T.A., Toh, K.-C., Trefethen, L.N.: From potential theory to matrix iterations in six steps. SIAM Rev. **40**, 547–578 (1998)
5. Eiermann, M.: Fields of values and iterative methods. Linear Algebra Appl. **180**, 167–197 (1993)
6. Eiermann, M., Ernst, O.G.: Geometric aspects in the theory of Krylov subspace methods. Acta Numer. **10**, 251–312 (2001)
7. Eisenstat, S.C., Elman, H.C., Schultz, M.H.: Variational iterative methods for nonsymmetric systems of linear equations. SIAM J. Numer. Anal. **20**, 345–357 (1983)
8. Elman, H.C.: Iterative Methods for Large, Sparse Nonsymmetric Systems of Linear Equations. Ph.D. thesis, Department of Computer Science, Yale University (1982)
9. Ernst, O.G.: Residual-minimizing Krylov subspace methods for stabilized discretizations of convection-diffusion equations. SIAM J. Matrix Anal. Appl. **21**, 1079–1101 (2000)
10. Greenbaum, A.: Iterative Methods for Solving Linear Systems. SIAM, Philadelphia (1997)
11. Greenbaum, A., Pták, V., Strakoš, Z.: Any nonincreasing convergence curve is possible for GMRES. SIAM J. Matrix Anal. Appl. **17**, 465–469 (1996)
12. Ipsen, I.C.F.: Expressions and bounds for the GMRES residual. BIT **40**, 524–535 (2000)
13. Li, R.-C.: Sharpness in rates of convergence for CG and symmetric Lanczos methods, Technical Report 2005-01 (Department of Mathematics, University of Kentucky, 2005). http://www.ms.uky.edu/~math/MAreport/
14. Li, R.-C.: Convergence of CG and GMRES on a tridiagonal Toeplitz linear system. BIT **47**, 577–599 (2007)
15. Li, R.-C.: Hard cases for conjugate gradient method. Int. J. Info. Sys. Sci. **4**, 15–29 (2008)
16. Li, R.-C.: On Meinardus' examples for the conjugate gradient method. Math. Comp. **77**, 335–352 (2008) (Electronically published on September 17, 2007)
17. Liesen, J., Rozložník, M., Strakoš, Z.: Least squares residuals and minimal residual methods. SIAM J. Sci. Comput. **23**, 1503–1525 (2002)
18. Liesen, J., Strakoš, Z.: Convergence of GMRES for tridiagonal Toeplitz matrices. SIAM J. Matrix Anal. Appl. **26**, 233–251 (2004)
19. Liesen, J., Strakoš, Z.: GMRES convergence analysis for a convection-diffusion model problem. SIAM J. Sci. Comput. **26**, 1989–2009 (2005)
20. Saad, Y.: Iterative Methods for Sparse Linear Systems, 2nd edn. SIAM, Philadelphia (2003)
21. Saad, Y., Schultz, M.: GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. SIAM J. Sci. Stat. Comput. **7**, 856–869 (1986)
22. Smith, G.D.: Numerical Solution of Partial Differential Equations, 2nd edn. Clarendon Press, Oxford (1978)
23. Trefethen, L.N., Pseudospectra of matrices. In: Griffiths, D.F., Watson, G.A. (eds.) Numerical Analysis 1991: Proceedings of the 14th Dundee Conference, June, 1991. Research Notes in Mathematics Series. Longman Press, London (1992)
24. Trefethen, L.N., Bau, D. III.: Numerical Linear Algebra. SIAM, Philadelphia (1997)
25. Zavorin, I., O'Leary, D.P., Elman, H.: Complete stagnation of GMRES. Linear Algebra Appl. **367**, 165–183 (2003)