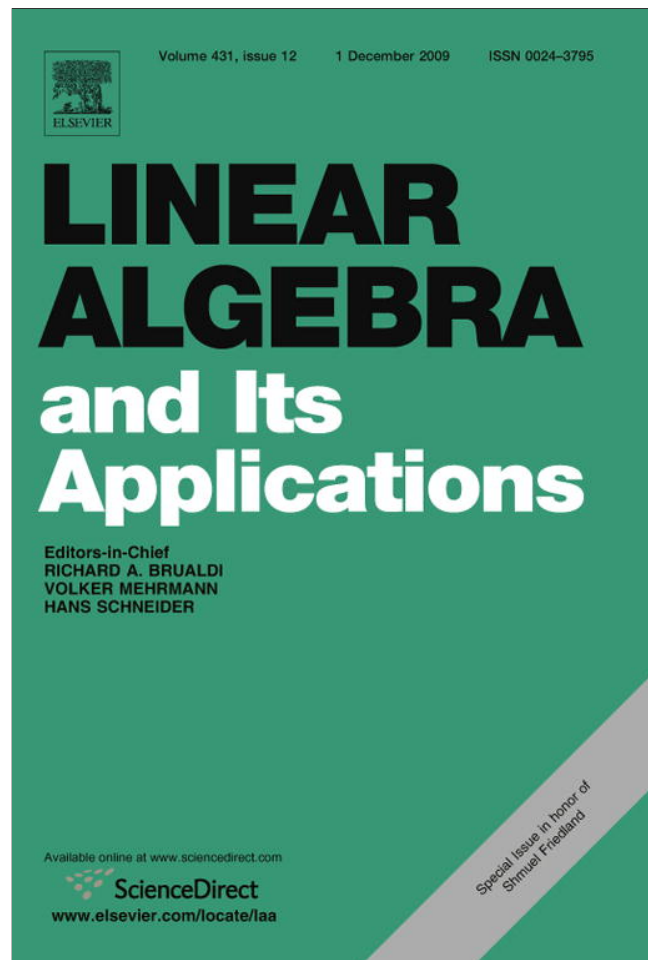


Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



Contents lists available at ScienceDirect

Linear Algebra and its Applications

journal homepage: www.elsevier.com/locate/laa

The rate of convergence of GMRES on a tridiagonal toeplitz linear system. II

Ren-Cang Li^{a,*}, Wei Zhang^{b,1,2}^a Department of Mathematics, University of Texas at Arlington, P.O. Box 19408, Arlington, TX 76019, United States^b Department of Mathematics, University of Kentucky, Lexington, KY 40506, United States

ARTICLE INFO

Article history:

Received 21 November 2008

Accepted 13 May 2009

Available online 23 June 2009

Submitted by F. Zhang

Dedicated to Professor Shmuel Friedland on the occasion of his 65th birthday

AMS classification:

65F10

Keywords:

GMRES

Rate of convergence

Tridiagonal Toeplitz matrix

Linear system

Asymptotic speed

ABSTRACT

This paper continues the recent work of the authors' [R.-C. Li, W. Zhang, The rate of convergence of GMRES on a tridiagonal Toeplitz linear system, Numer. Math. 112 (2009) 267–293 (electronically published on 19 December 2008)] on the rate of convergence of GMRES for a tridiagonal Toeplitz linear system $Ax = b$. Much simpler formulas than the earlier ones for GMRES residuals when b is the first or the last column of the identity matrix are established, and these formulas allow us to confirm the rate of convergence that was conjectured but only partially proven earlier. Simpler and sharper bounds than earlier ones when all b 's entries, except its first and last ones, are zeros are also obtained.

© 2009 Elsevier Inc. All rights reserved.

1. Introduction

This paper continues our recent work [9] and is concerned with the convergence analysis of the Generalized Minimal Residual method (GMRES) on linear system $Ax = b$ whose coefficient matrix A is a (nonsymmetric) tridiagonal Toeplitz coefficient matrix

* Corresponding author.

E-mail addresses: rcli@uta.edu (R.-C. Li), wzhang@ms.uky.edu, wz.zhang@siemens.com (W. Zhang).URL: <http://www.uta.edu/faculty/rcli> (R.-C. Li).¹ Supported in part by the National Science Foundation under Grant Nos. DMS-0702335 and DMS-0810506.² Present address: Siemens PLM Software Inc., 10824 Hope Street Cypress, CA 90630, United States.

$$A = \begin{pmatrix} \lambda & \mu & & & \\ & \ddots & \ddots & & \\ \nu & & & & \\ & \ddots & \ddots & & \\ & & & \mu & \\ & & & \nu & \lambda \end{pmatrix}, \tag{1.1}$$

where λ, μ, ν are assumed nonzero and possibly complex. When A is normal or symmetric positive definite, a collection of results were obtained in [1,14,15,6,7,12,13] for GMRES (or the conjugate gradient method). In general when A is nonsymmetric, convergence analysis for GMRES is a bit more complicated. The reader is referred to [9] and references therein.

The basic idea of GMRES is to seek approximate solutions from the so-called Krylov subspaces. Specifically, the k th approximation x_k is sought so that the k th residual $r_k = b - Ax_k$ satisfies [17] (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 k th Krylov subspace $\mathcal{K}_k \equiv \mathcal{K}_k(A, b)$ is defined as $\text{span}\{b, Ab, \dots, A^{k-1}b\}$, and generic norm $\|\cdot\|_2$ is the usual ℓ_2 norm of a vector or the spectral norm of a matrix.

This paper answers a couple of questions raised and remained unanswered in [9]. As in there, exact arithmetic is assumed, A in (1.1) is N -by- N , and k is GMRES iteration index. Since in exact arithmetic GMRES computes the exact solution in at most N steps, i.e., $r_N = 0$. For this reason, we restrict $k < N$ at all times.

Two main contributions of [9] are the upper bound

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

and the worst asymptotic speed

$$\liminf_{k \rightarrow \infty} \sup_{N > k} \left[\sup_{r_0} \frac{\|r_k\|_2}{\|r_0\|_2} \right]^{1/k} = \limsup_{k \rightarrow \infty} \sup_{N > k} \left[\sup_{r_0} \frac{\|r_k\|_2}{\|r_0\|_2} \right]^{1/k} = \min\{(\zeta\rho)^{-1}, 1\}, \tag{1.3}$$

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

$$\xi = -\frac{\sqrt{\mu\nu}}{\nu}, \quad \tau = \frac{\lambda}{2\sqrt{\mu\nu}}, \quad \zeta = \min\left\{|\xi|, \frac{1}{|\xi|}\right\}, \tag{1.4}$$

$$\rho = \max\left\{\left|\tau + \sqrt{\tau^2 - 1}\right|, \left|\tau - \sqrt{\tau^2 - 1}\right|\right\}. \tag{1.5}$$

Note $\rho \geq 1$ always because $(\tau + \sqrt{\tau^2 - 1})(\tau - \sqrt{\tau^2 - 1}) = 1$.

Also specifically for $b = e_1$ or e_N , in [9] we proved that when $|\xi| \leq 1$

$$\liminf_{k \rightarrow \infty} \sup_{N > k} \|r_k\|_2^{1/k} = \limsup_{k \rightarrow \infty} \sup_{N > k} \|r_k\|_2^{1/k} = \min\{(|\xi|\rho)^{-1}, 1\} \text{ for } b = e_1 \tag{1.6}$$

and that when $|\xi| \geq 1$

$$\liminf_{k \rightarrow \infty} \sup_{N > k} \|r_k\|_2^{1/k} = \limsup_{k \rightarrow \infty} \sup_{N > k} \|r_k\|_2^{1/k} = \min\{(|\xi|^{-1}\rho)^{-1}, 1\} \text{ for } b = e_N. \tag{1.7}$$

Whether or not both (1.6) and (1.7) would be true regardless of the magnitude of $|\xi|$ is the first unanswered question in [9] that will be confirmed positively later in this paper.

Numerical evidences given in [9] suggested that the upper bound by (1.2) is very accurate in revealing the convergence speed for the case when all of b 's entries except its first and last ones are zeros. A lower bound and an upper bound on the residual r_k were established to explain the numerical behavior. But it was only a partial success because the lower bound was only proven for $1 \leq k \leq N/2 - 1$, even

though numerically the lower bound appears to be good for all k . Thus it remains to be explained for the other k , i.e., $N/2 < k < N$. This is the second unanswered question in [9] to be addressed here.

A major difference in technicality between [9] and this paper is the use of Chebyshev polynomials of the first kind in the former and Chebyshev polynomials of the second kind in the latter. The outcome is simpler formulas and tighter bounds for residuals for the circumstances of the two unanswered questions to allow us to solve them, while the corresponding upper bound to (1.2) for a general right-hand side b is a bit more complicated however [20].

Toeplitz linear systems arise frequently in a variety of applications, such as image processing, numerical differential equations and integral equations, time series analysis, and control theory. There have been extensive studies on how to efficiently solve these systems by Krylov subspace methods. Interested readers are referred to [2,16]. In general, however, for Toeplitz matrices that are not tridiagonal, precise residuals and sharp bounds in terms of defining parameters are very difficult, if at all possible, to establish.

The rest of this paper is organized as follows. Main results are presented in Section 2 and their proofs are deferred to Section 3. Section 4 gives a couple of 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 j th column. The superscript “ \cdot^* ” takes conjugate transpose while “ \cdot^T ” takes transpose only.

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 j th entry, $X_{(ij)}$ is X 's (i, j) th entry, $\text{diag}(u)$ is the diagonal matrix with $(\text{diag}(u))_{(jj)} = u_{(j)}$; 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.

2. Main results

2.1. Setting the stage

This subsection is basically taken from [9, Section 2.1] with minor modifications. It is given here for completeness.

Let $N \times N$ tridiagonal Toeplitz A be given as in (1.1). Throughout the rest of this paper, ν , λ , and μ are reserved as the defining parameters of A . Set ξ , τ , ζ , and ρ as in (1.4) and (1.5). Matrix A is diagonalizable when $\mu \neq 0$ and $\nu \neq 0$. In fact [18, pp. 113–115] (see also [3,11]),

$$A = X\Lambda X^{-1}, \quad X = SZ, \quad \Lambda = \text{diag}(\lambda_1, \dots, \lambda_N), \tag{2.1}$$

$$\omega = -2\sqrt{\mu\nu}, \quad t_j = \cos \theta_j, \quad \theta_j = \frac{j\pi}{N+1}, \tag{2.2}$$

$$\lambda_j = \lambda - 2\sqrt{\mu\nu} t_j = \omega(t_j - \tau), \tag{2.3}$$

$$Z_{(:,j)} = \sqrt{\frac{2}{N+1}} (\sin j\theta_1, \dots, \sin j\theta_N)^T, \tag{2.4}$$

$$S = \text{diag}(1, \xi^{-1}, \dots, \xi^{-N+1}). \tag{2.5}$$

Any branch of $\sqrt{\mu\nu}$, once picked and fixed, is a valid choice in this paper. It can be verified that $Z^T Z = I_N$ and $Z^T = Z$.

A key starting point for our analysis is [19, Lemma 2.1]

$$\|r_k\|_2 = \min_{u_{(1)}=1} \|(b, Ab, \dots, A^k b)u\|_2 = \min_{u_{(1)}=1} \|Y V_{k+1, N}^T u\|_2, \tag{2.6}$$

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

$$V_{k+1,N} \stackrel{\text{def}}{=} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \cdots & \lambda_N \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_1^k & \lambda_2^k & \cdots & \lambda_N^k \end{pmatrix} \text{ and } Y = X \text{diag}(X^{-1}b). \tag{2.7}$$

Recall Chebyshev polynomials of the second kind:

$$U_m(t) = \frac{\sin((m + 1) \arccos t)}{\sin \arccos t} \text{ for real } t \text{ and } |t| \leq 1, \tag{2.8}$$

$$= \frac{(t + \sqrt{t^2 - 1})^{m+1} - (t - \sqrt{t^2 - 1})^{m+1}}{2\sqrt{t^2 - 1}} \tag{2.9}$$

and define the m th Translated Chebyshev Polynomial of the second kind in z of degree m by

$$U_m(z; \omega, \tau) \stackrel{\text{def}}{=} U_m(z/\omega + \tau) \tag{2.10}$$

$$= a_{mm}z^m + a_{m-1m}z^{m-1} + \cdots + a_{1m}z + a_{0m}, \tag{2.11}$$

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

$$R_m \equiv R_m(\omega, \tau) \stackrel{\text{def}}{=} \begin{pmatrix} a_{00} & a_{01} & a_{02} & \cdots & a_{0m-1} \\ & a_{11} & a_{12} & \cdots & a_{1m-1} \\ & & a_{22} & \cdots & a_{2m-1} \\ & & & \ddots & \vdots \\ & & & & a_{m-1m-1} \end{pmatrix}, \tag{2.12}$$

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

$$U_N \stackrel{\text{def}}{=} \begin{pmatrix} U_0(t_1) & U_0(t_2) & \cdots & U_0(t_N) \\ U_1(t_1) & U_1(t_2) & \cdots & U_1(t_N) \\ \vdots & \vdots & \ddots & \vdots \\ U_{N-1}(t_1) & U_{N-1}(t_2) & \cdots & U_{N-1}(t_N) \end{pmatrix} \tag{2.13}$$

and $V_N = V_{N,N}$ for short, where t_j is defined in (2.2). Then

$$V_N^T R_N = U_N^T. \tag{2.14}$$

Eq. (2.14) yields $V_N^T = U_N^T R_N^{-1}$. Extracting the first $k + 1$ columns from both sides of $V_N^T = U_N^T R_N^{-1}$ yields

$$V_{k+1,N}^T = U_{k+1,N}^T R_{k+1}^{-1}, \tag{2.15}$$

where $U_{k+1,N} = (U_N)_{(1:k+1,:)}$. Since $X^{-1} = Z^T S^{-1} = Z S^{-1}$, we have

$$Y V_{k+1,N}^T = S Z \text{diag}(Z S^{-1}b) \left(U_N^T \right)_{(:,1:k+1)} R_{k+1}^{-1} = \tilde{M}_{(:,1:k+1)} R_{k+1}^{-1}, \tag{2.16}$$

where

$$\tilde{M} = Z \text{diag}(Z S^{-1}b) U_N^T. \tag{2.17}$$

The following equivalence principle [9] will be used later to simplify our analysis.

Any result on the rate of convergence of GMRES for $Ax = b$ leads to one for $A^T y = \Pi^T b$ after performing the following substitutions $\mu \leftarrow \nu, \nu \leftarrow \mu, \xi \leftarrow \xi^{-1}, b \leftarrow \Pi^T b.$	(2.18)
---	--------

2.2. Main results

We first consider two special right-hand sides: $b = e_1$ and e_N . Because of the equivalence principle (2.18), in principle only $b = e_1$ needs to be considered. For completeness, we shall explicitly present results for both $b = e_1$ and e_N but proofs will be given only for the case $b = e_1$ in Section 3.

Theorem 2.1. For $Ax = b$, where A is tridiagonal Toeplitz as in (1.1) with nonzero (real or complex) parameters ν, λ , and μ , the k th GMRES residual r_k satisfies for $1 \leq k < N$

$$\|r_k\|_2 = [\Psi_{k+1}(\tau, \xi)]^{-1/2} \quad \text{for } b = e_1, \tag{2.19}$$

$$\|r_k\|_2 = [\Psi_{k+1}(\tau, \xi^{-1})]^{-1/2} \quad \text{for } b = e_N, \tag{2.20}$$

where

$$\Psi_{k+1}(t, s) \stackrel{\text{def}}{=} \sum_{j=0}^k |s|^{2j} |U_j(t)|^2. \tag{2.21}$$

Remark 2.1. Exact expressions for $\|r_k\|_2$ for $b = e_1$ and e_N were also established in [9], too, but in terms of Chebyshev polynomials of the first kind. They are much more complicated than what we have here in terms of Chebyshev polynomials of the second kind. It is the simplicity of (2.19) and (2.20) that makes it possible for us to establish the asymptotic speeds conjectured in [9] and given below in Theorem 2.2.

Theorem 2.2. Assume the conditions of Theorem 2.1 hold. Then

$$\liminf_{k \rightarrow \infty, N > k} \|r_k\|_2^{1/k} = \limsup_{k \rightarrow \infty, N > k} \|r_k\|_2^{1/k} = \min\{(|\xi|\rho)^{-1}, 1\} \quad \text{for } b = e_1, \tag{2.22}$$

$$\liminf_{k \rightarrow \infty, N > k} \|r_k\|_2^{1/k} = \limsup_{k \rightarrow \infty, N > k} \|r_k\|_2^{1/k} = \min\{(|\xi|^{-1}\rho)^{-1}, 1\} \quad \text{for } b = e_N. \tag{2.23}$$

Remark 2.2. As we mention in Section 1, (2.22) for $|\xi| \leq 1$ and (2.23) for $|\xi| \geq 1$ have already been proved by the authors in [9].

Next theorem deals with the case $b = b_{(1)}e_1 + b_{(N)}e_N$ which was argued and numerically demonstrated in [9] to be the most difficult case for GMRES with a given A .

Theorem 2.3. For $Ax = b$, where A is tridiagonal Toeplitz as in (1.1) with nonzero (real or complex) parameters ν, λ , and μ , and $b = b_{(1)}e_1 + b_{(N)}e_N$, the k th GMRES residual r_k satisfies for $1 \leq k < N$

$$\frac{\|r_k\|_2}{\|r_0\|_2} \leq \sqrt{2} [\Psi_{k+1}(\tau, \zeta)]^{-1/2} \tag{2.24}$$

and for $1 \leq k \leq N/2 - 1$

$$\frac{\|r_k\|_2}{\|r_0\|_2} \geq \frac{\left[|b_{(1)}|^2 (\Psi_{k+1}(\tau, \xi))^{-1} + |b_{(N)}|^2 (\Psi_{k+1}(\tau, \xi^{-1}))^{-1} \right]^{1/2}}{\|r_0\|_2} \tag{2.25}$$

and for $N/2 - 1 < k < N$

$$\frac{\|r_k\|_2}{\|r_0\|_2} \geq \frac{(|b_{(1)}| - |b_{(N)}| |\xi|^{2[N-(k+1)]})}{\|r_0\|_2} [\Psi_{k+1}(\tau, \xi)]^{-1/2} \quad \text{for } |\xi| \leq 1, \tag{2.26}$$

$$\frac{\|r_k\|_2}{\|r_0\|_2} \geq \frac{(|b_{(N)}| - |b_{(1)}| |\xi|^{-2[N-(k+1)]})}{\|r_0\|_2} [\Psi_{k+1}(\tau, \xi^{-1})]^{-1/2} \quad \text{for } |\xi| \geq 1. \tag{2.27}$$

Remark 2.3. For the case $b = b_{(1)}e_1 + b_{(N)}e_N$, a lower bound on $\|r_k\|_2$ was given only for $1 \leq k \leq N/2 - 1$ in [9]. Here we have lower bounds for all $1 \leq k < N$. Consider the situation $|b_{(1)}| = |b_{(N)}|$. Inequalities (2.24) and (2.25) lead to

$$\frac{1}{\sqrt{2}} [\Psi_{k+1}(\tau, \zeta)]^{-1/2} < \frac{\|r_k\|_2}{\|r_0\|_2} \leq \sqrt{2} [\Psi_{k+1}(\tau, \zeta)]^{-1/2} \tag{2.28}$$

and for $|\xi| \neq 1$, (2.24), (2.26), and (2.27) yield

$$\frac{1 - \zeta^{2[N-(k+1)]}}{\sqrt{2}} [\Psi_{k+1}(\tau, \zeta)]^{-1/2} \leq \frac{\|r_k\|_2}{\|r_0\|_2} \leq \sqrt{2} [\Psi_{k+1}(\tau, \zeta)]^{-1/2}. \tag{2.29}$$

Using the technique in the proof of Theorem 2.2 in the next section we conclude

$$\lim_{k \rightarrow \infty} [\Psi_{k+1}(\tau, \zeta)]^{-1/(2k)} = \min\{(\zeta\rho)^{-1}, 1\}$$

and thus it follows from (2.28) and (2.29) that

$$\lim_{k \rightarrow \infty} \inf_{N > k} \|r_k\|_2^{1/k} = \lim_{k \rightarrow \infty} \sup_{N > k} \|r_k\|_2^{1/k} = \min\{(\zeta\rho)^{-1}, 1\} \tag{2.30}$$

for (1) $1 \leq k \leq N/2 - 1$, and (2) $N/2 - 1 < k < N - 1$ and $|\xi| \neq 1$.

Remark 2.4. For $k = N - 1$, the leftmost quantity of (2.29) is zero; so are the right-hand sides of (2.26) and (2.27) when $|b_{(1)}| = |b_{(N)}|$. This makes $k = N - 1$ special and a finer analysis is called for in order to have some meaningful lower bound on $\|r_{N-1}\|_2$ to be established. One technique for this purpose is explained in Remark 3.2 following the proof of Theorem 2.3.

3. Proofs

Recall (2.17). We have

$$\begin{aligned} \tilde{M} &= \sum_{\ell=1}^N Z \text{diag}(ZS^{-1}b_{(\ell)}e_{\ell}) U_N^T = \sum_{\ell=1}^N b_{(\ell)}\xi^{\ell-1} Z \text{diag}(Ze_{\ell}) U_N^T \\ &= \sum_{\ell=1}^N b_{(\ell)}\xi^{\ell-1} Z \text{diag}(Z_{(c,\ell)}) U_N^T = \sum_{\ell=1}^N b_{(\ell)}\xi^{\ell-1} \tilde{M}_{\ell}, \end{aligned} \tag{3.1}$$

where $\tilde{M}_\ell = Z \text{diag}(Z_{(:,\ell)}) U_N^T$. Use $Z_{(:,\ell)} = \sqrt{\frac{2}{N+1}} (\sin \ell\theta_1, \dots, \sin \ell\theta_N)^T$ to get

$$\begin{aligned} \tilde{M}_\ell &= \frac{2}{N+1} \begin{pmatrix} \sin(\theta_1) & \sin(2\theta_1) & \cdots & \sin(N\theta_1) \\ \sin(\theta_2) & \sin(2\theta_2) & \cdots & \sin(N\theta_2) \\ \vdots & \vdots & \ddots & \vdots \\ \sin(\theta_N) & \sin(2\theta_N) & \cdots & \sin(N\theta_N) \end{pmatrix} \\ &\quad \times \begin{pmatrix} \sin(\ell\theta_1) & & & \\ & \sin(\ell\theta_2) & & \\ & & \ddots & \\ & & & \sin(\ell\theta_N) \end{pmatrix} \times \begin{pmatrix} \frac{\sin(\theta_1)}{\sin(\theta_1)} & \frac{\sin(2\theta_1)}{\sin(\theta_1)} & \cdots & \frac{\sin(N\theta_1)}{\sin(\theta_1)} \\ \frac{\sin(\theta_2)}{\sin(\theta_2)} & \frac{\sin(2\theta_2)}{\sin(\theta_2)} & \cdots & \frac{\sin(N\theta_2)}{\sin(\theta_2)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sin(\theta_N)}{\sin(\theta_N)} & \frac{\sin(2\theta_N)}{\sin(\theta_N)} & \cdots & \frac{\sin(N\theta_N)}{\sin(\theta_N)} \end{pmatrix} \\ &= Z \begin{pmatrix} \frac{\sin(\ell\theta_1)}{\sin(\theta_1)} & & & \\ & \frac{\sin(\ell\theta_2)}{\sin(\theta_2)} & & \\ & & \ddots & \\ & & & \frac{\sin(\ell\theta_N)}{\sin(\theta_N)} \end{pmatrix} Z. \end{aligned}$$

Since $Z = Z^T$, we can also write

$$\tilde{M}_\ell = Z^T D_\ell Z, \tag{3.2}$$

where $D = \text{diag} \left(\frac{\sin(\ell\theta_1)}{\sin(\theta_1)}, \frac{\sin(\ell\theta_2)}{\sin(\theta_2)}, \dots, \frac{\sin(\ell\theta_N)}{\sin(\theta_N)} \right)$. This in particular leads to

$$\tilde{M}_1 = I_N, \quad \tilde{M}_N = (e_N, e_{N-1}, \dots, e_2, e_1). \tag{3.3}$$

$\tilde{M}_1 = I_N$ is easy to see. For \tilde{M}_N , we notice that $N\theta_j = j\pi - \theta_j$ and therefore $D_N = \text{diag}(1, -1, 1, -1, \dots, (-1)^{N-1})$, and that $D_N Z_{(:,j)} = Z_{(:,N-j+1)}$.

Remark 3.1. It turns out that all \tilde{M}_ℓ can be completely described. In fact, their entries are either 1s or 0s and follow very regular patterns. In his PhD thesis [20], Zhang used these patterns to arrive at an upper bound similar to (1.2) for any general right-hand side b . But the bound is not simpler than (1.2); so we decide not to reproduce it in this paper. The interested reader is referred to Zhang's PhD thesis [20].

In its present general form, the next lemma was proven in [5,8]. It was also implied by the proof of [4, Theorem 2.1]. See also [10].

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

$$\min_{u_{(1)}=1} \|Wu\|_2 = \left[e_1^T (W^*W)^{-1} e_1 \right]^{-1/2}. \tag{3.4}$$

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

Proof of Theorem 2.1. For $b = e_1$, $\tilde{M} = \sum_{\ell=1}^N b_{(\ell)} \xi^{\ell-1} \tilde{M}_\ell = \tilde{M}_1 = I_N$. Let $S_{k+1} = \text{diag}(1, \xi^{-1}, \dots, \xi^{-k})$, the $(k+1)$ th leading principle submatrix of S . The k th GMRES residual is, by (2.6) and (2.16),

$$\begin{aligned} \|r_k\|_2 &= \min_{u_{(1)}=1} \|\tilde{S} \tilde{M}_{(:,1:k+1)} R_{k+1}^{-1} u\|_2 \\ &= \min_{u_{(1)}=1} \|S_{k+1} R_{k+1}^{-1} u\|_2. \end{aligned}$$

Apply Lemma 3.1 to get $\|r_k\|_2 = \left\| S_{k+1}^{-*} R_{k+1}^* e_1 \right\|_2^{-1} = [\Psi_{k+1}(\tau, \xi)]^{-1/2}$, as expected. \square

Proof of Theorem 2.2. We shall divide the proof of (2.22) into two cases: $\rho = 1$, and $\rho > 1$.

Recall (2.8) and (2.9). Consider first the case $\rho = 1$. Then $\tau + \sqrt{\tau^2 - 1} = e^{i\theta}$ for some $0 \leq \theta \leq \pi$, where $i = \sqrt{-1}$ is the imaginary unit. Thus

$$\tau = \frac{(\tau + \sqrt{\tau^2 - 1}) + (\tau - \sqrt{\tau^2 - 1})}{2} = \cos \theta \in [-1, 1], \quad U_j(\tau) = \frac{\sin(j+1)\theta}{\sin \theta}.$$

Since

$$\frac{\sin(j+1)\theta}{\sin \theta} = \frac{\sin j\theta \cos \theta + \cos j\theta \sin \theta}{\sin \theta} = \frac{\sin j\theta}{\sin \theta} \cos \theta + \cos j\theta,$$

we have

$$|U_j(\tau)| \leq |U_{j-1}(\tau)| + 1 \leq j + 1.$$

Now if $|\xi| \leq 1$, then by (2.19)

$$\left[\frac{1}{6}(2k+3)(k+2)(k+1) \right]^{-1/2} = \left[\sum_{j=0}^k (j+1)^2 \right]^{-1/2} \leq \|r_k\|_2 \leq 1,$$

which implies

$$\liminf_{k \rightarrow \infty} \inf_{N > k} \|r_k\|_2^{1/k} = \limsup_{k \rightarrow \infty} \sup_{N > k} \|r_k\|_2^{1/k} = 1 = \min\{(|\xi|/\rho)^{-1}, 1\}. \tag{3.5}$$

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

$$\alpha |\xi|^{2(k-1)} \leq \sum_{j=0}^k |\xi|^{2j} |U_j(\tau)|^2 \leq \sum_{j=0}^k |\xi|^{2j} (j+1)^2 \leq (k+1)^3 |\xi|^{2k} \tag{3.6}$$

for some constant $\alpha > 0$, independent of θ and k . The second and third inequalities in (3.6) are due to the fact that $|U_j(\tau)| \leq j + 1$ and $|\xi| \geq 1$. To see the first inequality, we notice

$$\sum_{j=k-1}^k |\xi|^{2j} |U_j(\tau)|^{2j} \geq |\xi|^{2(k-1)} \frac{\sin^2(k+1)\theta + \sin^2 k\theta}{\sin^2 \theta}.$$

It suffices to show that there is a positive constant α , independent of k and θ , such that

$$\frac{\sin^2(k+1)\theta + \sin^2 k\theta}{\sin^2 \theta} \geq \alpha.$$

Assume to the contrary that

$$\inf_{k, \theta} \frac{\sin^2(k+1)\theta + \sin^2 k\theta}{\sin^2 \theta} = 0, \tag{3.7}$$

which means there are sequences $\{k_i\}$ and $\{\theta_i\}$ such that

$$\frac{\sin^2(k_i+1)\theta_i + \sin^2 k_i\theta_i}{\sin^2 \theta_i} \rightarrow 0 \text{ as } i \rightarrow \infty. \tag{3.8}$$

Notice

$$\sin^2(k_i+1)\theta_i + \sin^2 k_i\theta_i \leq \frac{\sin^2(k_i+1)\theta_i + \sin^2 k_i\theta_i}{\sin^2 \theta_i},$$

to conclude that $\sin^2(k_i+1)\theta_i + \sin^2 k_i\theta_i \rightarrow 0$. For that to happen, because

$$\begin{aligned} \sin^2(k_i+1)\theta_i + \sin^2 k_i\theta_i &= 1 - \cos 2(k_i+1)\theta_i + 1 - \cos 2k_i\theta_i \\ &= 2[1 - \cos(2k_i+1)\theta_i \cos \theta_i] \end{aligned}$$

must $\cos \theta_i$ comes arbitrarily close to 1 or -1 , or equivalently θ_i comes arbitrarily close to an integer multiple of π , and at the same time both $\cos 2(k_i + 1)\theta_i$ and $\cos 2k_i\theta_i$ come arbitrarily close to 1, or equivalently $(k_i + 1)\theta_i$ and $k_i\theta_i$ come³ arbitrarily close to an integer multiple of π . Consequently for i large enough $[\sin^2(k_i + 1)\theta_i + \sin^2 k_i\theta_i] / \sin^2 \theta_i$ is approximately $(k_i + 1)^2 + k_i^2 \geq 1$, a contradiction to (3.8). Therefore (3.7) cannot hold. This proves (3.6) which yields

$$|\xi|^{-1} \leq \liminf_{k \rightarrow \infty, N > k} \|r_k\|_2^{1/k} \leq \limsup_{k \rightarrow \infty, N > k} \|r_k\|_2^{1/k} \leq |\xi|^{-1} = \min\{(|\xi|^{-1}), 1\}. \tag{3.9}$$

It remains to prove the theorem for the case $\rho > 1$. Suppose $\rho > 1$. Then by (2.9)

$$|U_j(\tau)| \sim \frac{\rho^{j+1}}{2|\sqrt{\tau^2 - 1}|}, \quad \|r_k\|_2 \sim \left[1 + \sum_{j=1}^k (|\xi|\rho)^{2j} \frac{\rho^2}{4|\tau^2 - 1|} \right]^{-1/2}. \tag{3.10}$$

Now if $|\xi|\rho \leq 1$, then

$$1 \leq 1 + \sum_{j=1}^k (|\xi|\rho)^{2j} \frac{\rho^2}{4|\tau^2 - 1|} \leq 1 + \frac{k\rho^2}{4|\tau^2 - 1|},$$

which, together with (3.10), lead to

$$\liminf_{k \rightarrow \infty, N > k} \|r_k\|_2^{1/k} = \limsup_{k \rightarrow \infty, N > k} \|r_k\|_2^{1/k} = 1 = \min\{(|\xi|\rho)^{-1}, 1\}.$$

If $|\xi|\rho > 1$, then from (3.10) we have

$$\|r_k\|_2 \sim \left[\frac{(|\xi|\rho)^{2(k+1)} - 1}{(|\xi|\rho)^2 - 1} \cdot \frac{\rho^2}{4|\tau^2 - 1|} \right]^{-1/2},$$

which yields

$$\liminf_{k \rightarrow \infty, N > k} \|r_k\|_2^{1/k} = \limsup_{k \rightarrow \infty, N > k} \|r_k\|_2^{1/k} = (|\xi|\rho)^{-1} = \min\{(|\xi|\rho)^{-1}, 1\}.$$

The proof is completed. \square

Proof of Theorem 2.3. For $b = b_{(1)}e_1 + b_{(N)}e_N$, we have $\tilde{M} = b_{(1)}\tilde{M}_1 + \xi^{N-1}b_{(N)}\tilde{M}_N$, and

$$\tilde{S}\tilde{M} = b_{(1)} \begin{pmatrix} 1 & & & \\ & \xi^{-1} & & \\ & & \ddots & \\ & & & \xi^{-N+1} \end{pmatrix} + b_{(N)} \begin{pmatrix} & & & \xi^{N-1} \\ & & & \\ & & & \\ 1 & \xi & & \end{pmatrix}. \tag{3.11}$$

Let $S_{k+1} = S_{(1:k+1, 1:k+1)}$ as in the proof of Theorem 2.1. For $|\xi| \leq 1$, we have by (2.6) and (2.16)

$$\|r_k\|_2 = \min_{u_{(1)}=1} \|\tilde{S}\tilde{M}_{(:,1:k+1)}S_{k+1}^{-1}S_{k+1}R_{k+1}^{-1}u\|_2 \tag{3.12}$$

$$\leq \|\tilde{S}\tilde{M}_{(:,1:k+1)}S_{k+1}^{-1}\|_2 \times \min_{u_{(1)}=1} \|S_{k+1}R_{k+1}^{-1}u\|_2 \tag{3.13}$$

and

$$\|\tilde{S}\tilde{M}_{(:,1:k+1)}S_{k+1}^{-1}\|_2 \leq |b_{(1)}| + |b_{(N)}| \leq \sqrt{2}\|b\|_2, \tag{3.14}$$

$$\min_{u_{(1)}=1} \|S_{k+1}R_{k+1}^{-1}u\|_2 = \|S_{k+1}^{-*}R_{k+1}^*e_1\|_2 = [\Psi_{k+1}(\tau, \xi)]^{-1/2}. \tag{3.15}$$

³ This is important because, for example, that θ_i comes arbitrarily close to 0 does not automatically imply $(k_i + 1)\theta_i$ and $k_i\theta_i$ come arbitrarily close to 0 at the same time and thus L'hospital's rule may not be applicable to see what $[\sin^2(k_i + 1)\theta_i + \sin^2 k_i\theta_i] / \sin^2 \theta_i$ is approaching to.

Inequality (2.24) for $|\xi| \leq 1$ is a consequence of (3.13), (3.14), (3.15). Inequality (2.24) for $|\xi| > 1$ is implied by itself for $|\xi| \leq 1$ and the equivalence principle (2.18).

We now prove the lower bounds for $\|r_k\|_2$. Notice (3.11); write $S\tilde{M} = b_{(1)}S + b_{(N)}T$; and let $T_{k+1} = T_{(N-k:N,1:k+1)}$. First if $k \leq N/2 - 1$,

$$YV_{k+1,N}^T = S\tilde{M}_{(:,1:k+1)}R_{k+1}^{-1} = \begin{matrix} k+1 \\ N-2(k+1) \\ k+1 \end{matrix} \begin{pmatrix} b_{(1)}S_{k+1}R_{k+1}^{-1} \\ 0 \\ b_{(N)}T_{k+1}R_{k+1}^{-1} \end{pmatrix}$$

By Lemma 3.1, we have

$$\begin{aligned} \min_{u_{(1)}=1} \|b_{(1)}S_{k+1}R_{k+1}^{-1}u\|_2 &= |b_{(1)}| \|S_{k+1}^{-*}R_{k+1}^*e_1\|_2 = |b_{(1)}| [\Psi_{k+1}(\tau, \xi)]^{-1/2}, \\ \min_{u_{(1)}=1} \|b_{(N)}T_{k+1}R_{k+1}^{-1}u\|_2 &= |b_{(N)}| \|T_{k+1}^{-*}R_{k+1}^*e_1\|_2 = |b_{(N)}| [\Psi_{k+1}(\tau, \xi^{-1})]^{-1/2}. \end{aligned}$$

Finally use

$$\min_{u_{(1)}=1} \|YV_{k+1,N}^T u\|_2 \geq \left[\min_{u_{(1)}=1} \|b_{(1)}S_{k+1}R_{k+1}^{-1}u\|_2^2 + \min_{u_{(1)}=1} \|b_{(N)}T_{k+1}R_{k+1}^{-1}u\|_2^2 \right]^{1/2}$$

to complete the proof of (2.25).

It remains to investigate the case $N > k > N/2 - 1$. Suppose $|\xi| \leq 1$ and $|b_{(1)}| > |b_{(N)}| |\xi|^{2[N-(k+1)]}$. It follows from (3.12) that

$$\|r_k\|_2 \geq \sigma_{\min}(S\tilde{M}_{(:,1:k+1)}S_{k+1}^{-1}) \times \min_{u_{(1)}=1} \|S_{k+1}R_{k+1}^{-1}u\|_2,$$

where $\sigma_{\min}(\cdot)$ is the smallest singular values of a matrix. It can be seen that the first $k+1$ rows of $S\tilde{M}_{(:,1:k+1)}S_{k+1}^{-1}$ are $b_{(1)}I_{k+1} + b_{(N)}W$ for some W satisfying $\|W\|_2 \leq |\xi|^{2[N-(k+1)]}$. Therefore

$$\sigma_{\min}(S\tilde{M}_{(:,1:k+1)}S_{k+1}^{-1}) \geq |b_{(1)}| - |b_{(N)}| |\xi|^{2[N-(k+1)]},$$

and thus

$$\|r_k\|_2 \geq (|b_{(1)}| - |b_{(N)}| |\xi|^{2[N-(k+1)]}) [\Psi_{k+1}(\tau, \xi)]^{-1/2}. \tag{3.16}$$

By the equivalence principle (2.18), we have for $|\xi| \geq 1$, if

$$|b_{(N)}| > |b_{(1)}| |\xi|^{-2[N-(k+1)]},$$

then

$$\|r_k\|_2 \geq (|b_{(N)}| - |b_{(1)}| |\xi|^{-2[N-(k+1)]}) [\Psi_{k+1}(\tau, \xi^{-1})]^{-1/2}. \tag{3.17}$$

This completes the proof. \square

Remark 3.2. For $|b_{(1)}| = |b_{(N)}|$, both (3.16) and (3.17) become trivial inequalities when $k = N - 1$. We shall explain a different argument for $k = N - 1$ that will lead to nontrivial lower bounds even if $|b_{(1)}| = |b_{(N)}|$. Suppose $k = N - 1$ and $|\xi| \leq 1$. We have

$$S\tilde{M}S^{-1} = b_{(1)}I_N + b_{(N)}e_N e_1^T + b_{(N)}W$$

for some W satisfying $\|W\|_2 \leq |\xi|^2$. Let $\delta = |b_{(N)}|/|b_{(1)}|$. Then the smallest singular value of $b_{(1)}I_N + b_{(N)}e_N e_1^T$ is the same as that of $b_{(1)} \begin{pmatrix} 1 & 0 \\ \delta & 1 \end{pmatrix}$ which is

$$|b_{(1)}| \left[\frac{2}{2 + \delta^2 + \delta\sqrt{2 + \delta^2}} \right]^{1/2}.$$

Thus

$$\begin{aligned} \sigma_{\min}(S\tilde{M}S^{-1}) &\geq \sigma_{\min}(b_{(1)}I_N + b_N e_N e_1^T) - \|b_{(N)}W\|_2 \\ &\geq |b_{(1)}| \left[\frac{2}{2 + \delta^2 + \delta\sqrt{2 + \delta^2}} \right]^{1/2} - |b_{(N)}| |\xi|^2. \end{aligned}$$

So if $|\xi| \leq 1$ and

$$|b_{(1)}| \left[\frac{2}{2 + \delta^2 + \delta\sqrt{2 + \delta^2}} \right]^{1/2} > |b_{(N)}| |\xi|^2,$$

or equivalently

$$\frac{1}{\delta} \left[\frac{2}{2 + \delta^2 + \delta\sqrt{2 + \delta^2}} \right]^{1/2} > |\xi|^2, \tag{3.18}$$

then

$$\|r_{N-1}\|_2 \geq |b_{(1)}| \left(\frac{1}{\delta} \left[\frac{2}{2 + \delta^2 + \delta\sqrt{2 + \delta^2}} \right]^{1/2} - |\xi|^2 \right) [\Psi_{k+1}(\tau, \xi)]^{-1/2}. \tag{3.19}$$

Similarly by the equivalence principle (2.18), we have if $|\xi| \geq 1$ and

$$|b_{(N)}| \left[\frac{2}{2 + \delta^{-2} + \delta^{-1}\sqrt{2 + \delta^{-2}}} \right]^{1/2} > |b_{(1)}| |\xi|^{-2},$$

or equivalently

$$\frac{1}{\delta^{-1}} \left[\frac{2}{2 + \delta^{-2} + \delta^{-1}\sqrt{2 + \delta^{-2}}} \right]^{1/2} > |\xi|^{-2}, \tag{3.20}$$

then

$$\|r_{N-1}\|_2 \geq |b_{(N)}| \left(\frac{1}{\delta^{-1}} \left[\frac{2}{2 + \delta^{-2} + \delta^{-1}\sqrt{2 + \delta^{-2}}} \right]^{1/2} - |\xi|^{-2} \right) [\Psi_{k+1}(\tau, \xi^{-1})]^{-1/2}. \tag{3.21}$$

To summarize, we have proved⁴

$$(3.19) \text{ holds if } |\xi| \leq 1; \quad (3.21) \text{ holds if } |\xi| \geq 1.$$

Inequalities (3.19) and (3.21) can provide useful estimates when $|b_{(1)}| = |b_{(N)}|$, while (3.16) and (3.17) cannot at $k = N - 1$.

4. Concluding remarks

Different from [9] which used Chebyshev polynomials of the first kind, this paper uses Chebyshev polynomials of the second kind to express or bound GMRES residuals for a tridiagonal Toeplitz linear system $Ax = b$. It results in much simpler formulas and sharper bounds when b is e_1, e_N , or $b_{(1)}e_1 + b_{(N)}e_N$. The simplicity of the formulas enables us to confirm (1.6) and (1.7), conjectured but only partially proven in [9]. Our results for the case $b_{(1)}e_1 + b_{(N)}e_N$ improve the corresponding ones

⁴ Even if (3.18) is violated, (3.19) is still valid, but just a trivial one because then its right-hand side is negative. A similar statement applies to (3.21).

in [9], too, not only the results are simpler and sharper but also the lower bounds cover all $1 \leq k < N$ while previously only $1 \leq k \leq N/2 - 1$.

Despite all these improvements with the use of Chebyshev polynomials of the second kind, the result for general right-hand side b so obtained is not better than with the first kind and in fact a little bit more complicated. We decide not to include it here. The interested reader is referred to [20] for more detail.

References

- [1] B. Beckermann, A.B.J. Kuijlaars, Superlinear CG convergence for special right-hand sides, *Electron. Trans. Numer. Anal.* 14 (2002) 1–19.
- [2] R.H. Chan, X.-Q. Jin, *An Introduction to Iterative Toeplitz Solvers*, SIAM, Philadelphia, PA, 2007.
- [3] O.G. Ernst, Residual-minimizing Krylov subspace methods for stabilized discretizations of convection–diffusion equations, *SIAM J. Matrix Anal. Appl.* 21 (4) (2000) 1079–1101.
- [4] I.C.F. Ipsen, Expressions and bounds for the GMRES residual, *BIT* 40 (3) (2000) 524–535.
- [5] R.-C. Li, 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/>>.
- [6] R.-C. Li, Hard cases for conjugate gradient method, *Int. J. Inform. Syst. Sci.* 4 (1) (2008) 15–29.
- [7] R.-C. Li, Convergence of CG and GMRES on a tridiagonal Toeplitz linear system, *BIT* 47 (3) (2007) 577–599.
- [8] R.-C. Li, On Meinardus' examples for the conjugate gradient method, *Math. Comput.* 77 (261) (2008) 335–352. (electronically published on September 17, 2007).
- [9] R.-C. Li, W. Zhang, The rate of convergence of GMRES on a tridiagonal Toeplitz linear system, *Numer. Math.* 112 (2009) 267–293. (electronically published on 19 December 2008).
- [10] J. Liesen, M. Rozložník, Z. Strakoš, Least squares residuals and minimal residual methods, *SIAM J. Sci. Comput.* 23 (5) (2002) 1503–1525.
- [11] J. Liesen, Z. Strakoš, Convergence of GMRES for tridiagonal Toeplitz matrices, *SIAM J. Matrix Anal. Appl.* 26 (1) (2004) 233–251.
- [12] J. Liesen, P. Tichý, The worst-case GMRES for normal matrices, *BIT* 44 (1) (2004) 79–98.
- [13] J. Liesen, P. Tichý, On the worst-case convergence of MR and CG for symmetric positive definite tridiagonal Toeplitz matrices, *Electron. Trans. Numer. Anal.* 20 (2005) 180–197.
- [14] A.E. Naiman, I.M. Babuška, H.C. Elman, A note on conjugate gradient convergence, *Numer. Math.* 76 (2) (1997) 209–230.
- [15] A.E. Naiman, S. Engelberg, A note on conjugate gradient convergence – Part II, Part III, *Numer. Math.* 85 (4) (2000) 665–683, 685–696.
- [16] M.K. Ng, *Iterative Methods for Toeplitz Systems*, Oxford University Press, New York, 2004.
- [17] Y. Saad, M. Schultz, GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems, *SIAM J. Sci. Stat. Comput.* 7 (1986) 856–869.
- [18] G.D. Smith, *Numerical Solution of Partial Differential Equations*, second ed., Clarendon Press, Oxford, UK, 1978.
- [19] I. Zavorin, D.P. O'Leary, H. Elman, Complete stagnation of GMRES, *Linear Algebra Appl.* 367 (2003) 165–183.
- [20] W. Zhang, GMRES on a tridiagonal Toeplitz linear system, Ph.D. Thesis, University of Kentucky, Lexington, KY, July 2007.