MATHEMATICS OF COMPUTATION
Volume 77, Number 261, January 2008, Pages 335–352
S 0025-5718(07)01922-9
Article electronically published on September 17, 2007

ON MEINARDUS' EXAMPLES FOR THE CONJUGATE GRADIENT METHOD

REN-CANG LI

Abstract. The conjugate gradient (CG) method is widely used to solve a positive definite linear system $Ax = b$ of order N. It is well known that the relative residual of the kth approximate solution by CG (with the initial approximation $x_0 = 0$) is bounded above by

$$2\left[\Delta_\kappa^k + \Delta_\kappa^{-k}\right]^{-1} \quad \text{with} \quad \Delta_\kappa = \frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1},$$

where $\kappa \equiv \kappa(A) = \|A\|_2\|A^{-1}\|_2$ is A's spectral condition number. In 1963, Meinardus (*Numer. Math.*, 5 (1963), pp. 14–23) gave an example to achieve this bound for $k = N-1$ but without saying anything about all other $1 \le k < N-1$. This very example can be used to show that the bound is sharp for any given k by constructing examples to attain the bound, but such examples depend on k and for them the $(k+1)$th residual is exactly zero. Therefore it would be interesting to know if there is any example on which the CG relative residuals are comparable to the bound for all $1 \le k \le N-1$. There are two contributions in this paper:

(1) A closed formula for the CG residuals for all $1 \le k \le N-1$ on Meinardus' example is obtained, and in particular it implies that the bound is always within a factor of $\sqrt{2}$ of the actual residuals;
(2) A complete characterization of extreme positive linear systems for which the kth CG residual achieves the bound is also presented.

1. Introduction

The conjugate gradient (CG) method is widely used to solve a positive definite linear system $Ax = b$ (often with certain preconditioning). The basic idea is to seek approximate solutions from the so-called Krylov subspaces. While different implementation may render different numerical behavior, mathematically[1] the kth approximate solution x_k by CG is the optimal one in the sense that the kth approximation error $A^{-1}b - x_k$ satisfies [11, Theorem 6:1]

$$\|A^{-1}b - x_k\|_A = \min_{x \in \mathcal{K}_k} \|A^{-1}b - x\|_A, \tag{1.1}$$

Received by the editor September 20, 2005, Revised January 9, 2006.
2000 *Mathematics Subject Classification.* Primary 65F10.
Key words and phrases. Conjugate gradient method, Krylov subspace, rate of convergence, Vandermonde matrix, condition number.
This work was supported in part by the National Science Foundation CAREER award under Grant No. CCR-9875201 and by the National Science Foundation under Grant No. DMS-0510664.

[1]Without loss of generality, we assume that A is already preconditioned and the initial approximation $x_0 = 0$.

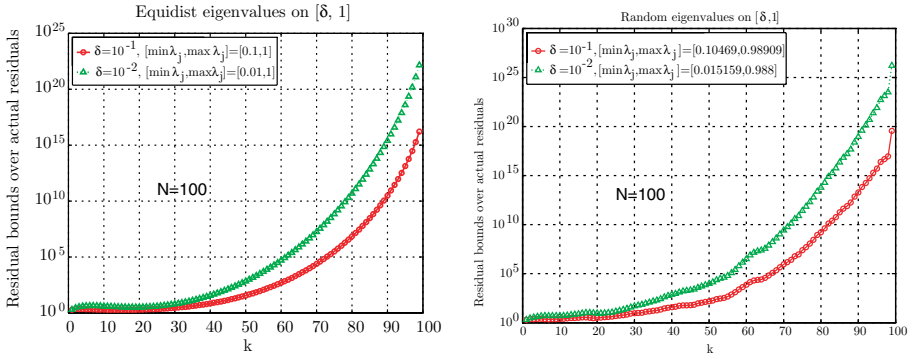

FIGURE 1.1. Conjugate gradient method for $Ax = b$ with $A = \mathsf{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ and b the vector of all ones—Ratios of (1.4) over the actual residuals for equidistant or random distributed λ_j's

or, equivalently, the kth residual $r_k = b - Ax_k$ satisfies

$$\|r_k\|_{A^{-1}} = \min_{x \in \mathcal{K}_k} \|b - Ax\|_{A^{-1}}, \tag{1.2}$$

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

$$\mathcal{K}_k \equiv \mathcal{K}_k(A, b) \stackrel{\text{def}}{=} \mathsf{span}\{b, Ab, \ldots, A^{k-1}b\}, \tag{1.3}$$

and M-vector norm $\|z\|_M \stackrel{\text{def}}{=} \sqrt{z^*Mz}$. Here the superscript "$\cdot^*$" takes conjugate transpose. In practice, x_k is computed recursively from x_{k-1} via short term recurrences [4, 7, 10, 19]. But exactly how it is computed, though extremely crucial in practice, is not important to our analysis here in this paper.

CG always converges for positive definite A. In fact, we have the following well known and frequently referenced error bound (see, e.g., [4, 10, 19, 22]):

$$\frac{\|r_k\|_{A^{-1}}}{\|r_0\|_{A^{-1}}} \equiv \frac{\|A^{-1}b - x_k\|_A}{\|A^{-1}b\|_A} \le 2\left[\Delta_\kappa^k + \Delta_\kappa^{-k}\right]^{-1}, \tag{1.4}$$

where $\kappa \equiv \kappa(A) = \|A\|_2\|A^{-1}\|_2$ is the spectral condition number, generic notation $\|\cdot\|_2$ is for either the spectral norm (the largest singular value) of a matrix or the euclidian length of a vector, and

$$\Delta_t \stackrel{\text{def}}{=} \frac{\sqrt{t}+1}{|\sqrt{t}-1|} \quad \text{for } t > 0 \tag{1.5}$$

that will be used frequently later for different t. The widely cited Kaniel [13] (1966) gave a proof of (1.4) while saying "This result is known" with a pointer to Meinardus [18] (1963). The same bound was proved to hold for Richardson-like processes [5, pp. 28–31], making it likely that (1.4) could be known before 1963 because their proofs do not differ much.

This paper is concerned with how sharp this well-known error bound is. Consider $A = \mathsf{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ and b the vector of all ones, where $\lambda_j \in [\delta, 1]$ is either randomly or equidistantly distributed on the interval. Figure 1.1 plots the ratios of the residual bounds by (1.4) over the actual residuals. What it shows is that initially for small k, bounds by (1.4) are good indications of actual residuals, but

as k becomes larger and larger, this bound overestimates the actual ones too much to be of any use. Are the phenomena in Figure 1.1 representative? Often this is what people observed [20], known as superlinear convergence.

In [18], Meinardus devised an $N \times N$ positive definite linear system $Ax = b$ for which he proved that

$$\frac{\|r_{N-1}\|_{A^{-1}}}{\|r_0\|_{A^{-1}}} = 2\left[\Delta_\kappa^{N-1} + \Delta_\kappa^{-(N-1)}\right]^{-1},$$

but without saying anything about all other $1 \le k < N-1$. This example of Meinardus' can be easily modified to give examples which achieve the error bound in (1.4) for any given $1 \le k < N-1$; e.g., trivially embed an example of Meinardus' of dimension $k+1$ with zero blocks to make it of dimension N. Greenbaum [10, p. 52] and [9] claimed that the error bound for a given k can be attained for A having eigenvalues at the extreme points of a translated Chebyshev polynomial of degree k and some particular b whose explicit form was not given, however. The same conclusion can be reached from [1, p. 561] if A's eigenvalues are chosen as Greenbaum's. This in a sense shows that the error bound in (1.4) is sharp and cannot be improved in general. But examples, i.e., A and b, constructed as such depend on the given step-index k and CG on any of these examples for k other than the example it was constructed for behaves much differently and in particular $r_{k+1} = 0$ exactly. So this only proves that the error bound is "*locally*" sharp: for given k,

$$\max_{\kappa(A)=\gamma} \frac{2\left[\Delta_\gamma^k + \Delta_\gamma^{-k}\right]^{-1}}{\|r_k\|_{A^{-1}}/\|r_0\|_{A^{-1}}} = 1. \tag{1.6}$$

What about its "*global*" sharpness? For example,

(1.7) *Is there any positive definite system $Ax = b$ for which relative residuals $\|r_k\|_{A^{-1}}/\|r_0\|_{A^{-1}}$ achieve the error bounds by* (1.4) *for all $1 \le k < N-1$?*

This question turns out to be too strong and the answer is "no" by Kaniel [13, Theorem 4.4], who showed that if r_k attains the bound, then it must be $r_{k+1} = 0$, (see also Theorem 2.2 below). So instead we ask

(1.8) *Is*

$$\sup_{\kappa(A)=\gamma} \max_{1\le k\le n-1} \frac{2\left[\Delta_\gamma^k + \Delta_\gamma^{-k}\right]^{-1}}{\|r_k\|_{A^{-1}}/\|r_0\|_{A^{-1}}}$$

modestly bounded?

This question has been recently answered positively in Li [14], using A with eigenvalues being the translated zeros of Nth Chebyshev polynomial of the first kind. It is proved there that the ratio in (1.8) can be bounded from above by a bound that asymptotically approaches to $\sqrt{2}\,\Delta_\gamma/\sqrt{\Delta_\gamma^2 - 1}$ as k goes to infinity. It depends on $\kappa(A) = \gamma$ and, unfortunately, can be arbitrarily large as $\gamma \to 1^+$. A much stronger bound on the ratio, namely $\sqrt{2}$, is implied later in this paper.

In what follows, we shall compute the CG residuals on Meinardus' examples for all $1 \le k \le N-1$ and investigate extreme positive linear systems for which the kth CG residual achieves the error bound in (1.4). Before we set out to do so, let us look at some numerical examples. Figure 1.2 plots the ratios of the error bounds

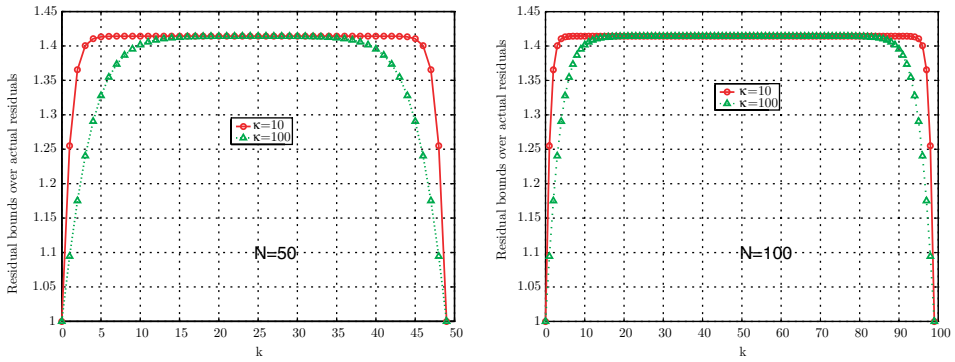

FIGURE 1.2. Ratios of the error bound in (1.4) over the exact CG residuals for Meinardus' example.

by (1.4) over the actual CG relative residuals, i.e., the right-hand side of (1.4) over its left-hand side, on a Meinardus' example, where the exact CG residuals were carefully computed within MAPLE[2] with a sufficiently high precision. While it is not surprising at all to see that the ratios are not smaller than 1, they seem not to be bigger than $\sqrt{2}$ as well. This in fact will be confirmed by one of our main results, which will also furnish another example for the global sharpness question (1.8), in addition to the one in [14].

The rest of this paper is organized as follows. Section 2 explains Meinardus' examples and gives our main results—the closed formula for CG residuals for a Meinardus example and a complete characterization of extreme positive linear systems for which the kth CG residual achieves the error bound in (1.4). Proofs for our main results are rather long and thus are given separately in Section 3 and Section 4. Concluding remarks are given in Section 5.

Notation. Throughout this paper, $\mathbb{C}^{n\times m}$ is the set of all $n\times m$ complex matrices, $\mathbb{C}^n = \mathbb{C}^{n\times 1}$, and $\mathbb{C} = \mathbb{C}^1$. Similarly define $\mathbb{R}^{n\times m}$, $\mathbb{R}^n$, and $\mathbb{R}$ except replacing the word *complex* by *real*. 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 "$.^T$" takes transpose only. We shall also adopt MATLAB-like convention to access the entries of vectors and matrices. $i:j$ is the set of integers from i to j inclusive. For vector u and matrix X, $u_{(j)}$ is u's jth entry, $X_{(i,j)}$ is X's (i,j)th entry, $\mathsf{diag}(u)$ is the diagonal matrix with $(\mathsf{diag}(u))_{(j,j)} = u_{(j)}$; X's submatrices $X_{(k:\ell,i:j)}$, $X_{(k:\ell,:)}$, and $X_{(:,i:j)}$ consist of intersections of row k to row ℓ and column i to column j, row k to row ℓ, and column i to column j, respectively.

2. MEINARDUS' EXAMPLES AND MAIN RESULTS

The mth Chebyshev polynomial of the first kind is

$$T_m(t) = \cos(m\arccos t) \quad \text{for } |t|\le 1, \tag{2.1}$$

$$= \frac{1}{2}\left(t+\sqrt{t^2-1}\right)^m + \frac{1}{2}\left(t-\sqrt{t^2-1}\right)^m \quad \text{for } |t|\ge 1. \tag{2.2}$$

It frequently shows up in numerical analysis and computations because of its numerous nice properties, for example $|T_m(t)|\le 1$ for $|t|\le 1$ and $|T_m(t)|$ grows extremely

[2] `http://www.maplesoft.com/`.

fast for $|t| > 1$. Later we will need

$$\left|T_m\left(\frac{1+t}{1-t}\right)\right| \equiv \left|T_m\left(\frac{t+1}{t-1}\right)\right| = \frac{1}{2}\left[\Delta_t^m + \Delta_t^{-m}\right] \quad \text{for } 1 \neq t > 0. \tag{2.3}$$

The first equality holds because $T_m(-t) = (-1)^m T_m(t)$. We shall prove the second equality for $0 < t < 1$ only and a proof for $t > 1$ is similar. For $0 < t < 1$, we have

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

which proves (2.3) for $0 < t < 1$. $T_m(t)$ has $m+1$ extreme points in $[-1, 1]$, so-called *the mth Chebyshev extreme nodes*:

$$\tau_{jm} = \cos\vartheta_{jm},\ \vartheta_{jm} = \frac{j}{m}\pi,\ 0 \le j \le m, \tag{2.4}$$

at which $|T_m(\tau_{jm})| = 1$. Given $\alpha < \beta$, set

$$\omega = \frac{\beta - \alpha}{2} > 0, \quad \tau = -\frac{\alpha+\beta}{\beta-\alpha}. \tag{2.5}$$

The linear transformation

$$t(z) = \frac{z}{\omega} + \tau = \frac{2}{\beta-\alpha}\left(z - \frac{\alpha+\beta}{2}\right) \tag{2.6}$$

maps $z \in [\alpha, \beta]$ one-to-one and onto $t \in [-1, 1]$. With its inverse transformation $x(t) = \omega(t - \tau)$, we define the so-called *mth translated Chebyshev extreme nodes* on $[\alpha, \beta]$:

$$\tau_{jm}^{\mathsf{tr}} = \omega(\tau_{jm} - \tau),\ 0 \le j \le m. \tag{2.7}$$

It can be verified that $\tau_{0m} = \beta$ and $\tau_{mm} = \alpha$.

Now we are ready to state Meinardus examples. Assume $0 < \alpha < \beta$. For the sake of presentation, set

$$n = N - 1.$$

Let Q be any $N \times N$ unitary matrix. A Meinardus' example is a positive definite system $Ax = b$ with

$$A = Q\Lambda Q^*, \quad b = Q\Lambda^{1/2} g, \tag{2.8}$$

where

$$\Lambda \stackrel{\text{def}}{=} \mathsf{diag}(\tau_{0n}^{\mathsf{tr}}, \tau_{1n}^{\mathsf{tr}}, \ldots, \tau_{nn}^{\mathsf{tr}}), \quad g_{(j+1)} \stackrel{\text{def}}{=} \begin{cases} \sqrt{1/\tau_{jn}^{\mathsf{tr}}}, & \text{for } j \in \{0, n\}, \\ \sqrt{2/\tau_{jn}^{\mathsf{tr}}}, & \text{for } 1 \le j \le n-1. \end{cases} \tag{2.9}$$

So an example of Meinardus' is any member of the family parameterized by unitary Q. Theorem 2.1 is one of the two main results of this paper.

Theorem 2.1. *Let $0 < \alpha < \beta$ and let A and b be given by* (2.8) *and* (2.9). *r_k is the kth CG residual with initially $r_0 = b$. Then*

$$\frac{\|r_k\|_{A^{-1}}}{\|r_0\|_{A^{-1}}} = \rho_k \times 2\left[\Delta_\kappa^k + \Delta_\kappa^{-k}\right]^{-1} \tag{2.10}$$

for $1 \le k \le n$, where $\kappa \equiv \kappa(A) = \beta/\alpha$ and

$$\frac{1}{2} < \frac{1}{2}\left(1 + \frac{2\Delta_\kappa^n}{\Delta_\kappa^{2n}+1}\right) \le \rho_k^2 = \frac{1}{2}\left(1 + \frac{\Delta_\kappa^{2k} + \Delta_\kappa^{2(n-k)}}{\Delta_\kappa^{2n}+1}\right) \le 1. \tag{2.11}$$

Remark 2.1. (1) As far as the equality is concerned, (2.10) is valid for $k = 0$ as well, which corresponds to the very beginning of CG.

(2) The factor ρ_k is symmetrical in k about $n/2$, i.e., $\rho_k = \rho_{n-k}$. This phenomenon certainly showed up in Figure 1.2 which equivalently plotted ρ_k^{-1}.

(3) $\rho_k \le 1$ with equality if and only if $k = 0$ or n.

(4) ρ_k is strictly decreasing for $k \le \lfloor n/2 \rfloor$ (the largest integer that is no bigger than $n/2$) and strictly increasing for $k \ge \lceil n/2 \rceil$ (the smallest integer that is no less than $n/2$), and

$$\frac{1}{\sqrt{2}} < \min_{0\le k\le n} \rho_k = \rho_{\lfloor n/2\rfloor} \to \frac{1}{\sqrt{2}} \quad \text{as } n \to \infty.$$

The fact that $\rho_n = 1$ has already been established by Meinardus [18]. With it, one can easily construct a positive definite linear system $Ax = b$ for which the kth CG residual achieves the error bound in (1.4). For example, A and b are given by (2.8) and (2.9), where $\Lambda = \mathsf{diag}(\tau_{0k}^{\mathsf{tr}}, \tau_{1k}^{\mathsf{tr}}, \ldots, \tau_{kk}^{\mathsf{tr}}, \ldots)$, i.e., $k+1$ of A's eigenvalues are $\tau_{0k}^{\mathsf{tr}}, \tau_{1k}^{\mathsf{tr}}, \ldots, \tau_{kk}^{\mathsf{tr}}$, and $g_{(j+1)}$ is $\sqrt{1/\tau_{jk}^{\mathsf{tr}}}$ for $j \in \{0, k\}$ and $\sqrt{2/\tau_{jk}^{\mathsf{tr}}}$ for $1 \le j \le k-1$ and zero for all other j, then $\|r_k\|_{A^{-1}}/\|r_0\|_{A^{-1}} = 2\left[\Delta_\kappa^k + \Delta_\kappa^{-k}\right]^{-1}$. For this example $r_{k+1} = 0$, i.e., convergence occurs at the $(k+1)$th step! This is not a coincidence, as it must be due to Kaniel [13, Theorem 4.4]. The following theorem characterizes all extreme linear systems as such.

Theorem 2.2. *Let $Ax = b \ne 0$ be a positive definite linear system of order N, and $1 \le k < N$. If the kth CG residual r_k (initially $r_0 = b$) achieves the error bound in* (1.4)*, i.e.,*

$$\frac{\|r_k\|_{A^{-1}}}{\|r_0\|_{A^{-1}}} = 2\left[\Delta_\kappa^k + \Delta_\kappa^{-k}\right]^{-1}, \tag{2.12}$$

where $\kappa \equiv \kappa(A) = \|A\|_2\|A^{-1}\|_2$, then the following statements hold.

(1) *$A = Q\Lambda Q^*$ and $b = Q\Lambda^{1/2}g$ for some unitary $Q \in \mathbb{C}^{N\times N}$,*

$$\Lambda = \mathsf{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$$

with $0 < \lambda_1 \le \lambda_2 \le \cdots \le \lambda_N$, and $g \in \mathbb{R}^N$ with all $g_{(j)} \ge 0$.

(2) *$\sum_{\lambda_j=\lambda_1} g_{(j)}^2 > 0$ and $\sum_{\lambda_j=\lambda_N} g_{(j)}^2 > 0$.*

(3) *Let $\alpha = \min_j \lambda_j$, and $\beta = \max_j \lambda_j$, and let τ_{jk}^{tr} be the translated Chebyshev extreme nodes on $[\alpha, \beta]$. The distinct λ_j's in $\{\lambda_j : g_{(j)} > 0\}$ consist of exactly τ_{jk}^{tr}, $0 \le j \le k$, i.e.,*

$$\{\tau_{jk}^{\mathsf{tr}},\ 0 \le j \le k\} \subset \{\lambda_j : g_{(j)} > 0\}, \text{ and}$$
$$\lambda_i \in \{\tau_{jk}^{\mathsf{tr}},\ 0 \le j \le k\} \text{ if } g_{(i)} > 0.$$

(4) [13, Theorem 4.4] *$r_{k+1} \equiv 0$.*

(5) *Let $\mathbb{J}_\ell = \{j : \lambda_j = \tau_{\ell k}^{\mathsf{tr}}, g_{(j)} > 0\}$. For some constant $\mu > 0$,*

$$\|g_{\mathbb{J}_\ell}\|_2 = \mu \begin{cases} \sqrt{1/\tau_{\ell k}^{\mathsf{tr}}}, & \text{for } \ell \in \{0, k\}, \\ \sqrt{2/\tau_{\ell k}^{\mathsf{tr}}}, & \text{for } 1 \le \ell \le k-1. \end{cases} \tag{2.13}$$

Remark 2.2. Any $Ax = b$ described by items (1), (2), (3), and (5) is essentially equivalent to an example of Meinardus', (2.8) and (2.9), with $N = k+1$. Therefore this theorem practically says that the kth CG residual r_k (initially $r_0 = b$) achieves the error bound in (1.4) *if and only if* $Ax = b$ is an example of Meinardus'. Thus

unless $N = 2$, there is no positive linear system whose kth CG residual achieves the error bound in (1.4) for all $1 \le k < N$.

3. Proof of Theorem 2.1

We will adopt in whole the notation introduced in Section 2 and assume $0 < \alpha < \beta$. Recall, in particular, $n = N - 1$ and A is $N \times N$.

Theorem 2.1 will be proved through a restatement. For A as in (2.8) and (2.9),

$$\begin{aligned} \min_{x \in \mathcal{K}_k} \|b - Ax\|_{A^{-1}} &= \min_{\phi_k(0)=1} \|\phi_k(A)b\|_{A^{-1}} \\ &= \min_{\phi_k(0)=1} \|\phi_k(\Lambda)\Lambda^{-1/2}Q^*b\|_2 \\ &= \min_{\phi_k(0)=1} \|\phi_k(\Lambda)g\|_2 \\ &= \min_{|u_{(1)}|=1} \|\mathsf{diag}(g)\, V_{k+1,n}^T u\|_2, \end{aligned} \tag{3.1}$$

where $\phi_k(t)$ is a polynomial of degree k, $u \in \mathbb{C}^{k+1}$, and with $\alpha_{j+1} = \tau_{jn}^{\mathsf{tr}}$ for $0 \le j \le n$,

$$V_{k+1,N} \stackrel{\text{def}}{=} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \alpha_1 & \alpha_2 & \cdots & \alpha_N \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1^k & \alpha_2^k & \cdots & \alpha_N^k \end{pmatrix}, \tag{3.2}$$

a $(k+1) \times N$ rectangular Vandermonde matrix. Note also that $\|r_0\|_{A^{-1}} = \|g\|_2$. Therefore, after substitution $k+1 \to k$, Theorem 2.1 can be equivalently stated as follows.

Theorem 3.1. *Let $0 < \alpha < \beta$, g as in (2.9), and $V_{k,N}$ as in (3.2) with $\alpha_{j+1} = \tau_{jn}^{\mathsf{tr}}$ for $0 \le j \le n$. Then*

$$\min_{|u_{(1)}|=1} \frac{\|\mathsf{diag}(g)\, V_{k,N}^T u\|_2}{\|g\|_2} = \rho_{k-1} \times 2 \left[\Delta_\kappa^{k-1} + \Delta_\kappa^{-(k-1)}\right]^{-1}. \tag{3.3}$$

for $1 \le k \le N = n+1$, where $\kappa = \beta/\alpha$.

The rest of this section is devoted to the proof of this theorem. Notice that $T_j(t(z)) \equiv T_j(z/\omega + \tau)$ is a polynomial of degree j in z; so we write

$$T_j(z/\omega + \tau) = a_{jj}z^j + a_{j-1\,j}z^{j-1} + \cdots + a_{1j}z + a_{0j},$$

where $a_{ij} \equiv a_{ij}(\omega, \tau)$ are functions of ω and τ in (2.5). Their explicit dependence on ω and τ is often suppressed for convenience. For integer $m \ge 1$, define upper triangular $R_m \in \mathbb{R}^{m \times m}$, a matrix-valued function in ω and τ, as

$$R_m \equiv R_m(\omega, \tau) \stackrel{\text{def}}{=} \begin{pmatrix} a_{00} & a_{01} & a_{02} & \cdots & a_{0\,m-1} \\ & a_{11} & a_{12} & \cdots & a_{1\,m-1} \\ & & a_{22} & \cdots & a_{2\,m-1} \\ & & & \ddots & \vdots \\ & & & & a_{m-1\,m-1} \end{pmatrix}, \tag{3.4}$$

i.e., the jth column consists of the coefficients of $T_{j-1}(z/\omega+\tau)$. Write $V_N = V_{N,N}$ for short and set

$$\boldsymbol{S}_N \stackrel{\text{def}}{=} \begin{pmatrix} T_0(\tau_{0n}) & T_0(\tau_{1n}) & \cdots & T_0(\tau_{nn}) \\ T_1(\tau_{0n}) & T_1(\tau_{1n}) & \cdots & T_1(\tau_{nn}) \\ \vdots & \vdots & & \vdots \\ T_n(\tau_{0n}) & T_n(\tau_{1n}) & \cdots & T_n(\tau_{nn}) \end{pmatrix}. \tag{3.5}$$

Then $V_N^T R_N = \boldsymbol{S}_N^T$. Since R_N is upper triangular, we have

$$V_{k,N}^T = \boldsymbol{S}_{k,N}^T R_k^{-1}, \tag{3.6}$$

a key decomposition of $V_{k,N}^T$ that will play a vital role later in our proofs, where $\boldsymbol{S}_{k,N} = (\boldsymbol{S}_N)_{(1:k,:)}$ is $\boldsymbol{S}_N$'s first k rows. Set

$$\Omega = \mathsf{diag}(2^{-1}, 1, 1, \ldots, 1, 2^{-1}) \in \mathbb{R}^{N\times N}, \quad \Upsilon \stackrel{\text{def}}{=} \boldsymbol{S}_N \Omega \boldsymbol{S}_N^T. \tag{3.7}$$

Lemma 3.2 below says Υ is diagonal. So essentially (3.6) gives a QR-like decomposition of $V_{k,N}^T$.

Lemma 3.1 below is probably well known but a precise reference is hard to find. However, it can be proved by using either Euler identity $\cos\theta = (e^{\iota k\theta} + e^{-\iota k\theta})/2$, where $\iota = \sqrt{-1}$, the imaginary unit, or the identities in [8, p. 30]. Detail is omitted.

Lemma 3.1. *Let $\vartheta_{kn} = \pi k/n$ as in* (2.4). *Then*

$$\sum_{k=0}^{n} \cos \ell\vartheta_{kn} = \begin{cases} N, & \text{if } \ell = 2mn \text{ for some integer } m, \\ 0, & \text{if } \ell \text{ is odd}, \\ 1, & \text{if } \ell \text{ is even, but } \ell \neq 2mn \text{ for any integer } m. \end{cases} \tag{3.8}$$

Lemma 3.2 is known to [2, p. 33], and probably long before that. A proof can be given with the help of Lemma 3.1, and again detail is omitted.

Lemma 3.2. *Let $\boldsymbol{S}_N$, Ω, and Υ be defined as in* (3.5) *and* (3.7). *Then $\Upsilon = \frac{n}{2}\Omega^{-1}$.*

Lemma 3.3. *Let $\Gamma = \mathsf{diag}(\mu+\nu\cos\vartheta_{0n}, \mu+\nu\cos\vartheta_{1n}, \ldots, \mu+\nu\cos\vartheta_{nn})$ and define $\Upsilon_{\mu,\nu} \stackrel{\text{def}}{=} \boldsymbol{S}_N\Omega\Gamma\boldsymbol{S}_N^T$, where $\mu, \nu \in \mathbb{C}$. We have*

$$\Upsilon_{\mu,\nu} = \frac{n}{4}\Omega^{-1}\left(2\mu\Omega + \nu H\right)\Omega^{-1}, \tag{3.9}$$

where

$$H = \begin{pmatrix} 0 & 1 & & & \\ 1 & 0 & 1 & & \\ & 1 & \ddots & \ddots & \\ & & \ddots & 0 & 1 \\ & & & 1 & 0 \end{pmatrix} \in \mathbb{R}^{N\times N}.$$

Proof. Notice that $\Upsilon_{\mu,\nu} = \mu\Upsilon_{1,0} + \nu\Upsilon_{0,1}$ and $\Upsilon_{1,0} = \boldsymbol{S}_N\Omega\boldsymbol{S}_N^T = \Upsilon = \frac{n}{2}\Omega^{-1}$ by Lemma 3.2. It is enough to calculate $\Upsilon_{0,1}$. Using $\sum''$ to mean the first and last

terms halved, we have for $0 \le i, j \le n$,

$$
\begin{aligned}
(3.10)\ (\boldsymbol{S}_N \Omega \Gamma \boldsymbol{S}_N^T)_{(i+1,j+1)} &= \sum_{k=0}^{n}{}'' (\boldsymbol{S}_N)_{(i+1,k)} \, (\mu + \nu \cos\vartheta_{kn}) (\boldsymbol{S}_N^T)_{(k,j+1)} \\
&= \sum_{k=0}^{n}{}'' T_i(\tau_{kn})(\mu + \nu \cos\vartheta_{kn}) T_j(\tau_{kn}) \\
&= \sum_{k=0}^{n}{}'' \cos i\vartheta_{kn} \, (\mu + \nu \cos\vartheta_{kn}) \, \cos j\vartheta_{kn} \\
&= \mu \sum_{k=0}^{n}{}'' \cos i\vartheta_{kn} \, \cos j\vartheta_{kn} \\
&\quad + \nu \sum_{k=0}^{n}{}'' \cos i\vartheta_{kn} \, \cos\vartheta_{kn} \, \cos j\vartheta_{kn}.
\end{aligned}
$$

So $(\Upsilon_{0,1})_{(i+1,j+1)} = \sum_{k=0}^{n}{}'' \cos i\vartheta_{kn} \, \cos\vartheta_{kn} \, \cos j\vartheta_{kn}$. Now

$$
\begin{aligned}
& 4 \sum_{k=0}^{n}{}'' \cos i\vartheta_{kn} \, \cos\vartheta_{kn} \, \cos j\vartheta_{kn} \\
&= \sum_{k=0}^{n}{}'' \cos(i+j+1)\vartheta_{kn} + \sum_{k=0}^{n}{}'' \cos(i+j-1)\vartheta_{kn} \\
&\quad + \sum_{k=0}^{n}{}'' \cos(i-j+1)\vartheta_{kn} + \sum_{k=0}^{n}{}'' \cos(i-j-1)\vartheta_{kn}.
\end{aligned}
$$

Apply Lemma 3.1 to conclude $\Upsilon_{0,1} = \frac{n}{4} \, \Omega^{-1} H \Omega^{-1}$ whose verification is straightforward, albeit tedious. □

Lemma 3.4. *Let $m \le n$ and $\xi \in \mathbb{C}$ such that $(-2\xi\Omega + H)_{(1:m,1:m)}$ is nonsingular. Then the first entry of the solution to $(-2\xi\Omega + H)_{(1:m,1:m)} \, y = e_1$ is*

$$
y_{(1)} = \frac{\gamma_-^m - \gamma_+^m}{\sqrt{\xi^2 - 1}(\gamma_-^m + \gamma_+^m)},
$$

where $\gamma_\pm = \xi \pm \sqrt{\xi^2 - 1}$.

Proof. Expand y to have a 0th entry $y_{(0)}$ and a $(m+1)$th entry $y_{(m+1)}$ satisfying

$$
y_{(0)} - \xi y_{(1)} = -1, \quad y_{(m+1)} = 0. \tag{3.11}
$$

Entry-wise, we have

$$
y_{(i-1)} - 2\xi y_{(i)} + y_{(i+1)} = 0, \quad \text{for } 1 \le i \le m.
$$

The general solution has form $y_{(i)} = c_+ \gamma_+^i + c_- \gamma_-^i$, where $\gamma_\pm$ are the two roots of $1 - 2\xi\gamma + \gamma^2 = 0$, i.e., $\gamma_\pm = \xi \pm \sqrt{\xi^2 - 1}$. We now determine c_+ and c_- by the edge conditions (3.11):

$$
\begin{aligned}
(1 - \xi\gamma_+) \, c_+ \quad + \quad (1 - \xi\gamma_-) \, c_- &= -1, \\
\gamma_+^{m+1} \, c_+ \quad + \quad \gamma_-^{m+1} \, c_- &= 0.
\end{aligned}
$$

Notice $\gamma_+\gamma_- = 1$ and

$$\begin{aligned}(1-\xi\gamma_+)\gamma_-^{m+1} - (1-\xi\gamma_-)\gamma_+^{m+1} &= (\gamma_- - \xi)\gamma_-^m - (\gamma_+ - \xi)\gamma_+^m \\ &= -\sqrt{\xi^2-1}(\gamma_-^m + \gamma_+^m)\end{aligned}$$

to get

$$c_+ = \frac{-\gamma_-^{m+1}}{-\sqrt{\xi^2-1}(\gamma_-^m+\gamma_+^m)}, \quad c_- = \frac{+\gamma_+^{m+1}}{-\sqrt{\xi^2-1}(\gamma_-^m+\gamma_+^m)}.$$

Finally $y_{(1)} = c_+\gamma_+ + c_-\gamma_-$. □

In its present general form, the next lemma was proved in [14]. It was also implied by the proof of [12, Theorem 2.1]. See also [16].

Lemma 3.5. *If Z has full column rank, then*

$$\min_{|u_{(1)}|=1} \|Zu\|_2 = \left[e_1^T (Z^*Z)^{-1} e_1\right]^{-1/2}. \tag{3.12}$$

Proof. Set $v = Zu$. Since Z has full column rank, its Moore-Penrose pseudo-inverse is $Z^\dagger = (Z^*Z)^{-1}Z^*$ [21] and thus $u = Z^\dagger v$. This gives a one-one and onto mapping between $u \in \mathbb{C}^m$ and the column space $v \in \mathsf{span}(Z)$. Now

$$\min_{|u_{(1)}|=1} \|Zu\|_2 = \min_u \frac{\|Zu\|_2}{|u_{(1)}|} = \min_{v\in\mathsf{span}(Z)} \frac{\|v\|_2}{|e_1^T Z^\dagger v|} \ge \min_v \frac{\|v\|_2}{|e_1^T Z^\dagger v|} = \|e_1^T Z^\dagger\|_2^{-1}, \tag{3.13}$$

where the last min is achieved at

$$v_{\rm opt} = \left(e_1^T Z^\dagger\right)^* = Z(Z^*Z)^{-1}e_1 \in \mathsf{span}(Z),$$

which implies the "$\ge$" in (3.13) is actually an equality, and $u_{\rm opt} = Z^\dagger v_{\rm opt}/e_1^T Z^\dagger v_{\rm opt}$. Finally

$$\|e_1^T Z^\dagger\|_2 = \sqrt{e_1^T Z^\dagger (Z^\dagger)^* e_1} = \sqrt{e_1^T (Z^*Z)^{-1} e_1}.$$

This completes the proof. □

Proof of Theorem 3.1. By Lemma 3.5,

$$\min_{|u_{(1)}|=1} \frac{\|\mathsf{diag}(g) V_{k,N}^T u\|_2}{\|g\|_2} = \frac{\left[e_1^T \left(V_{k,N}[\mathsf{diag}(g)]^2 V_{k,N}^T\right)^{-1} e_1\right]^{-1/2}}{\|g\|_2}. \tag{3.14}$$

Let $\Gamma = \mathsf{diag}(\tau_{0n}^{\mathsf{tr}}, \tau_{1n}^{\mathsf{tr}}, \ldots, \tau_{nn}^{\mathsf{tr}}) \equiv \mathsf{diag}(\mu+\nu\cos\vartheta_{00}, \mu+\nu\cos\vartheta_{01}, \ldots, \mu+\nu\cos\vartheta_{nn})$, where $\mu = -\omega\tau$ and $\nu = \omega$ as in (2.5). Then

$$\begin{aligned} V_{k,N}[\mathsf{diag}(g)]^2 V_{k,N}^T &= 2V_{k,N}\Gamma^{-1}\Omega V_{k,N}^T \\ &= 2\begin{pmatrix} e^T \\ V_{k-1,N}\Gamma\end{pmatrix}\Gamma^{-1}\Omega\begin{pmatrix} e & \Gamma V_{k-1,N}^T\end{pmatrix} \\ &= 2\begin{pmatrix} e^T\Gamma^{-1}\Omega e & e^T\Omega V_{k-1,N}^T \\ V_{k-1,N}\Omega e & V_{k-1,N}\Gamma\Omega V_{k-1,N}^T\end{pmatrix}, \end{aligned} \tag{3.15}$$

where $e = (1, 1, \ldots, 1)^T$. Notice $V_{k-1,N}^T = \boldsymbol{S}_{k-1,N}^T R_{k-1}^{-1}$ by (3.6) to get

$$
\begin{aligned}
V_{k-1,N}\Omega e &= V_{k-1,N}\Omega V_{k-1,N}^T e_1 \\
&= R_{k-1}^{-T}\left(\Upsilon_{1,0}\right)_{(1:k-1,1:k-1)} R_{k-1}^{-1} e_1 \\
&= R_{k-1}^{-T}\left(\Upsilon_{1,0}\right)_{(1:k-1,1:k-1)} e_1, \qquad (3.16)\\
V_{k-1,N}\Gamma\Omega V_{k-1,N}^T &= R_{k-1}^{-T}\boldsymbol{S}_{k-1,N}\Gamma\Omega\boldsymbol{S}_{k-1,N}^T R_{k-1}^{-1} \\
&= R_{k-1}^{-T}\left(\Upsilon_{\mu,\nu}\right)_{(1:k-1,1:k-1)} R_{k-1}^{-1}, \qquad (3.17)
\end{aligned}
$$

in the notation introduced in Lemma 3.3. Recall[3]

$$
\begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}^{-1} = \begin{pmatrix} C_{11}^{-1} & -C_{11}^{-1}B_{12}B_{22}^{-1} \\ -B_{22}^{-1}B_{21}C_{11}^{-1} & B_{22}^{-1} + B_{22}^{-1}B_{21}C_{11}^{-1}B_{12}B_{22}^{-1} \end{pmatrix},
$$

assuming all inversions exist, where $C_{11} = B_{11} - B_{12}B_{22}^{-1}B_{21}$. We have from (3.15)

$$
\begin{aligned}
& e_1^T\left(V_{k,N}[\mathsf{diag}(g)]^2 V_{k,N}^T\right)^{-1} e_1 \qquad (3.18)\\
&\quad = \frac{1}{2}\left[\zeta - e^T\Omega V_{k-1,N}^T\left(V_{k-1,N}\Gamma\Omega V_{k-1,N}^T\right)^{-1} V_{k-1,N}\Omega e\right]^{-1},
\end{aligned}
$$

where $\zeta = e^T\Gamma^{-1}\Omega e$. But, from (3.16) and (3.17),

$$
\begin{aligned}
& e^T\Omega V_{k-1,N}^T\left(V_{k-1,N}\Gamma\Omega V_{k-1,N}^T\right)^{-1} V_{k-1,N}\Omega e \qquad (3.19)\\
&\quad = e_1^T\left(\Upsilon_{1,0}\right)_{(1:k-1,1:k-1)}\left[\left(\Upsilon_{\mu,\nu}\right)_{(1:k-1,1:k-1)}\right]^{-1}\left(\Upsilon_{1,0}\right)_{(1:k-1,1:k-1)} e_1 \\
&\quad = n^2 e_1^T\left[\left(\Upsilon_{\mu,\nu}\right)_{(1:k-1,1:k-1)}\right]^{-1} e_1,
\end{aligned}
$$

and for $k \le N$, by Lemma 3.4 with $m = k-1$ and $\xi = \tau$,

$$
\begin{aligned}
& e_1^T\left[\left(\Upsilon_{\mu,\nu}\right)_{(1:k-1,1:k-1)}\right]^{-1} e_1 \qquad (3.20)\\
&\quad = n^{-1}\, e_1^T\left[\left(2\mu\Omega + \nu H\right)_{(1:k-1,1:k-1)}\right]^{-1} e_1 \\
&\quad = \frac{1}{n\omega} e_1^T\left[\left(-2\tau\Omega + H\right)_{(1:k-1,1:k-1)}\right]^{-1} e_1 \\
&\quad = \frac{1}{n\omega}\frac{\gamma_-^{k-1} - \gamma_+^{k-1}}{\sqrt{\tau^2-1}\left(\gamma_-^{k-1} + \gamma_+^{k-1}\right)},
\end{aligned}
$$

where $\gamma_\pm = \tau \pm \sqrt{\tau^2 - 1}$. The conditions of Lemma 3.4 are fulfilled because $|\tau| > 1$ and $-2\tau\Omega + H$ is diagonally dominant and thus nonsingular. Since $2\zeta = \|g\|_2^2$, we have by (3.14) and (3.18)–(3.20)

$$
\min_{|u_{(1)}|=1}\frac{\|\mathsf{diag}(g)V_{k,N}^T u\|_2}{\|g\|_2} = \left[1 - \frac{n}{\omega\zeta\sqrt{\tau^2-1}}\frac{\gamma_-^{k-1} - \gamma_+^{k-1}}{\gamma_-^{k-1} + \gamma_+^{k-1}}\right]^{1/2}. \qquad (3.21)
$$

We now compute $\omega\zeta\sqrt{\tau^2-1}$. Let $f(z) \stackrel{\text{def}}{=} \prod_{j=0}^n (z - \tau_{jn}^{\mathsf{tr}})$. Then

$$
f(z) = \eta(z - \tau_{0n}^{\mathsf{tr}})(z - \tau_{nn}^{\mathsf{tr}})\, U_{n-1}(z/\omega + \tau)
$$

[3]This is well known. See, e.g., [6, pp. 102-103], [23, p. 23].

for some constant η, where $U_{n-1}(t)$ is the $(n-1)$th Chebyshev polynomial of the second kind. This is because the zeros of $U_{n-1}(z/\omega+\tau)$ are precisely $\tau^{\mathsf{tr}}_{jn} = \omega(\tau_{jn}-\tau)$, $j=1,2,\ldots,n-1$. Then, upon noticing $\tau^{\mathsf{tr}}_{0n}=\beta$ and $\tau^{\mathsf{tr}}_{nn}=\alpha$,

$$\begin{aligned} 2\zeta &= \sum_{j=0}^{n}{}'' \frac{2}{\tau^{\mathsf{tr}}_{jn}} = -\frac{1}{\tau^{\mathsf{tr}}_{0n}} + 2\sum_{j=0}^{n}\frac{1}{\tau^{\mathsf{tr}}_{jn}} - \frac{1}{\tau^{\mathsf{tr}}_{nn}} = -\left(\frac{1}{\alpha}+\frac{1}{\beta}\right) - 2\frac{f'(0)}{f(0)} \\ &= -\frac{\alpha+\beta}{\alpha\beta} - 2\frac{-(\alpha+\beta)\,U_{n-1}(\tau)+\alpha+\beta\,U'_{n-1}(\tau)/\omega}{\alpha\beta\,U_{n-1}(\tau)} \\ &= \frac{\alpha+\beta}{\alpha\beta} - \frac{2}{\omega}\frac{U'_{n-1}(\tau)}{U_{n-1}(\tau)}. \end{aligned}$$

Recall [3, p. 37]

$$\begin{aligned} 2U_{n-1}(t) &= \frac{(t+\sqrt{t^2-1})^n - (t-\sqrt{t^2-1})^n}{\sqrt{t^2-1}}, \\ 2U'_{n-1}(t) &= n\frac{(t+\sqrt{t^2-1})^n + (t-\sqrt{t^2-1})^n}{t^2-1} \\ &\quad - \frac{t\left[(t+\sqrt{t^2-1})^n - (t-\sqrt{t^2-1})^n\right]}{(t^2-1)\sqrt{t^2-1}}. \end{aligned} \tag{3.22}$$

They yield

$$2U_{n-1}(\tau) = \frac{\gamma_+^n-\gamma_-^n}{\sqrt{\tau^2-1}}, \quad 2U'_{n-1}(\tau) = n\frac{\gamma_+^n+\gamma_-^n}{\tau^2-1} - \frac{\tau(\gamma_+^n-\gamma_-^n)}{(\tau^2-1)\sqrt{\tau^2-1}}.$$

Therefore, upon noticing $\omega=(\alpha+\beta)/2$ and $\tau=-(\beta+\alpha)/(\beta-\alpha)$,

$$\begin{aligned} 2\zeta &= \frac{\alpha+\beta}{\alpha\beta} + \frac{2}{\omega}\frac{n}{\sqrt{\tau^2-1}}\frac{\gamma_-^n+\gamma_+^n}{\gamma_-^n-\gamma_+^n} + \frac{2}{\omega}\frac{\tau}{\tau^2-1} \\ &= \frac{2}{\omega}\frac{n}{\sqrt{\tau^2-1}}\frac{\gamma_-^n+\gamma_+^n}{\gamma_-^n-\gamma_+^n}, \end{aligned} \tag{3.23}$$

$$\omega\zeta\sqrt{\tau^2-1} = n\frac{\gamma_-^n+\gamma_+^n}{\gamma_-^n-\gamma_+^n}. \tag{3.24}$$

Equation (3.21) and (3.24) imply

$$\min_{|u_{(1)}|=1}\frac{\|\mathsf{diag}(g)V_{k,N}^T u\|_2}{\|g\|_2} = \left[1-\frac{\gamma_-^n-\gamma_+^n}{\gamma_-^n+\gamma_+^n}\frac{\gamma_-^{k-1}-\gamma_+^{k-1}}{\gamma_-^{k-1}+\gamma_+^{k-1}}\right]^{1/2}. \tag{3.25}$$

Because $\tau=-(\kappa+1)/(\kappa-1)$,

$$\gamma_-^{k-1} = (-1)^{k-1}\Delta_\kappa^{k-1}, \quad \gamma_+^{k-1} = (-1)^{k-1}\Delta_\kappa^{-(k-1)},$$

and therefore

$$\min_{|u_{(1)}|=1}\frac{\|\mathsf{diag}(g)V_{k,N}^T u\|_2}{\|g\|_2} = \left[1-\frac{\Delta_\kappa^n-\Delta_\kappa^{-n}}{\Delta_\kappa^n+\Delta_\kappa^{-n}}\frac{\Delta_\kappa^{k-1}-\Delta_\kappa^{-(k-1)}}{\Delta_\kappa^{k-1}+\Delta_\kappa^{-(k-1)}}\right]^{1/2}. \tag{3.26}$$

For $k=N\equiv n+1$, the right-hand side of (3.26) is $2\left[\Delta_\kappa^n+\Delta_\kappa^{-n}\right]^{-1}$, as was shown by Meinardus [18]. For any other k, we have

$$\min_{|u_{(1)}|=1}\frac{\|\mathsf{diag}(g)V_{k,N}^T u\|_2}{\|g\|_2} = \rho_{k-1}\times 2\left[\Delta_\kappa^{k-1}+\Delta_\kappa^{-(k-1)}\right]^{-1}. \tag{3.27}$$

where

$$\begin{aligned}
\rho_{k-1} &\stackrel{\text{def}}{=} \frac{\text{Right-hand side of (3.26)}}{2\left[\Delta_\kappa^{k-1}+\Delta_\kappa^{-(k-1)}\right]^{-1}} \\
&= \left[\frac{\left(\Delta_\kappa^{k-1}+\Delta_\kappa^{-(k-1)}\right)^2}{4} - \frac{\left(\Delta_\kappa^{n}-\Delta_\kappa^{-n}\right)\left(\Delta_\kappa^{2(k-1)}-\Delta_\kappa^{-2(k-1)}\right)}{4\left(\Delta_\kappa^{n}+\Delta_\kappa^{-n}\right)}\right]^{1/2} \\
&= \left[\frac{1}{2}\frac{\left(\Delta_\kappa^{k-1}+\Delta_\kappa^{-(k-1)}\right)\left(\Delta_\kappa^{n-(k-1)}+\Delta_\kappa^{-[n-(k-1)]}\right)}{\Delta_\kappa^{n}+\Delta_\kappa^{-n}}\right]^{1/2} \\
&= \left[\frac{1}{2}\frac{\left(\Delta_\kappa^{2(k-1)}+1\right)\left(\Delta_\kappa^{2[n-(k-1)]}+1\right)}{\Delta_\kappa^{2n}+1}\right]^{1/2},
\end{aligned}$$

which yields (2.11). □

4. Proof of Theorem 2.2

We first prove two general lemmas for the Vandermonde matrix $V_N \equiv V_{N,N}$ as defined in (3.2) with arbitrary, possibly complex, nodes α_j.

Lemma 4.1. *Assume one or more of* 1) *there are fewer than n distinct α_j,* 2) *some $\alpha_j = 0$, and* 3) *some $g_{(j)} = 0$ occur. Then*

$$\min_{|u_{(1)}|=1} \|\mathsf{diag}(g)V_N^T u\|_2 = \begin{cases} 0, & \text{if all } \alpha_j \neq 0; \\ \sqrt{\sum_{\alpha_j=0}|g_{(j)}|^2}, & \text{otherwise.} \end{cases}$$

Proof. If all $\alpha_j \neq 0$, only cases 1) and 3) are possible. Let ℓ be the number of distinct α_j's, exclude those corresponding to $g_{(j)} = 0$. Then $\ell < n$. By permuting the rows of $\mathsf{diag}(g)V_N^T$, we may assume that $\alpha_1, \alpha_2, \ldots, \alpha_\ell$ are distinct and for α_j $(j > \ell)$ either it is equal to some α_i $(i \leq \ell)$ or corresponding $g_{(j)} = 0$. Set $v \in \mathbb{C}^N$ whose $v_{(j)}$ is the coefficient of z^{j-1} in the polynomial $\phi(z) = \prod_{j=1}^{\ell}(z-\alpha_j)$. Then $v_{(1)} = \prod_{j=1}^{\ell}(-\alpha_j) \neq 0$, and

$$\min_{|u_{(1)}|=1} \|\mathsf{diag}(g)V_N^T u\|_2 \leq \|\mathsf{diag}(g)V_n^T(v/v_{(1)})\|_2 = 0,$$

as expected.

If some $\alpha_j = 0$, then since $\|\mathsf{diag}(g)V_N^T u\|_2 \geq \sqrt{\sum_{\alpha_j=0}|g_{(j)}|^2}$ always for any vector u with $|u_{(1)}| = 1$, it suffices to find a vector u to annihilate all other rows corresponding to $\alpha_j \neq 0$. Such u can be constructed similarly to what we just did. □

The next lemma is essentially [17, Theorems 2.1 and 3.1], but stated differently. The proof below has a slightly different flavor.

Lemma 4.2 ([17])**.** *Let $V_N \equiv V_{N,N}$ be as defined in* (3.2) *with all nodes α_j (possibly complex) distinct, and let $f(z) = \prod_{j=1}^{N}(z-\alpha_j)$.*

(1) *If all* $g_{(j)} \neq 0$, *then*

$$\min_{|u_{(1)}|=1} \frac{\|\mathsf{diag}(g)V_N^T u\|_2}{\|g\|_2} = \left[\sum_{j=1}^N \left(\frac{|f(0)|}{|\alpha_j|\,|f'(\alpha_j)|}\right)^2 |g_{(j)}|^{-2} \sum_{j=1}^N |g_{(j)}|^2\right]^{-1/2}. \tag{4.1}$$

(2)

$$\max_g \min_{|u_{(1)}|=1} \frac{\|\mathsf{diag}(g)V_N^T u\|_2}{\|g\|_2} = \left[\sum_{j=1}^N \frac{|f(0)|}{|\alpha_j|\,|f'(\alpha_j)|}\right]^{-1/2}, \tag{4.2}$$

where the maximum is achieved if and only if for some constant $\mu > 0$,

$$|g_{(j)}| = \mu \left[\frac{|f(0)|}{|\alpha_j|\,|f'(\alpha_j)|}\right]^{1/2} \quad \textit{for } 1 \le j \le N. \tag{4.3}$$

Proof. In Lemma 3.5, take $Z = \mathsf{diag}(g)V_N^T$. The assumptions make this Z nonsingular. Therefore

$$\begin{aligned}\left[\min_{|u_{(1)}|=1} \|\mathsf{diag}(g)V_N^T u\|_2\right]^{-2} &= e_1^T(\bar{V}_N \Phi V_N^T)^{-1} e_1 \\ &= e_1^T(V_N \Phi V_N^*)^{-1} e_1 \\ &= (V_N^{-1}e_1)^* \Phi^{-1} (V_N^{-1} e_1),\end{aligned}$$

where $\Phi = [\mathsf{diag}(g)]^* \,\mathsf{diag}(g)$ and $\bar{V}_N$ is the complex conjugate of V_N. Let $y = V_N^{-1}e_1$, the first column of V_N^{-1} which consists of the constant terms of the Lagrangian basis functions,

$$\ell_j(z) = \prod_{i \neq j} \frac{z - \alpha_i}{\alpha_i - \alpha_j}, \quad 1 \le j \le N,$$

since $\ell_j(\alpha_i) = 1$ for $i = j$ and 0 otherwise, which means the jth row of V_N^{-1} consists of the coefficients of $\ell_j(z)$. Therefore

$$\begin{aligned} e_1^T(\bar{V}_N \Phi V_N^T)^{-1} e_1 &= \sum_{j=1}^N \left(\frac{|f(0)|}{|\alpha_j|\,|f'(\alpha_j)|}\right)^2 |g_{(j)}|^{-2}, \\ \min_{|u_{(1)}|=1} \frac{\|\mathsf{diag}(g)V_N^T u\|_2}{\|g\|_2} &= \left[\sum_{j=1}^N \left(\frac{|f(0)|}{|\alpha_j|\,|f'(\alpha_j)|}\right)^2 |g_{(j)}|^{-2} \sum_{j=1}^N |g_{(j)}|^2\right]^{-1/2} \\ &\le \left[\sum_{j=1}^N \frac{|f(0)|}{|\alpha_j|\,|f'(\alpha_j)|}\right]^{-1/2}, \end{aligned} \tag{4.4}$$

where it is an equality if and only if $|g_{(j)}|$ are given by (4.3). □

Remark 4.1. This lemma closely relates to a result of Greenbaum [9, (2.2) and Theorem 1] which in our notation essentially proved that if all nodes $\alpha_j > 0$, there exist k of α_j's: $\alpha_{j_1}, \ldots, \alpha_{j_k}$ such that

$$\max_g \min_{|u_{(1)}|=1} \frac{\|\mathsf{diag}(g)V_{k,N}^T u\|_2}{\|g\|_2} = \max_h \min_{|u_{(1)}|=1} \frac{\|\mathsf{diag}(h)V_k^T u\|_2}{\|h\|_2},$$

where V_k is the $k \times k$ Vandermonde matrix with nodes α_{j_i} $(1 \le i \le k)$. Notice the difference in conditions: Lemma 4.3 only covers $k = N$, while this result of Greenbaum's is for all $1 \le k \le N$ but requires all $\alpha_j > 0$. Greenbaum [9, Theorem 1] also obtained an expression for the optimal h but a bit of more complicated than we get from applying Lemma 4.3. It is not clear how to find out the most relevant nodes α_{j_i}.

Lemma 4.3. *Let $\omega, \tau \in \mathbb{C}$ (not necessarily associated with any interval $[\alpha, \beta]$ as previously required), and let $n = N - 1$ and τ_{jn}^{tr} as in (2.7) with any given ω and τ. Suppose the Vandermonde matrix V_N has nodes $\alpha_{j+1} = \tau_{jn}^{\mathsf{tr}}$ for $0 \le j \le n$.*

(1) *If all $g_{(j)} \neq 0$, then*

(4.5)
$$\min_{|u_{(1)}|=1} \frac{\|\mathsf{diag}(g)\, V_N^T u\|_2}{\|g\|_2} = \frac{n\omega}{|\tau_{0n}^{\mathsf{tr}} \tau_{nn}^{\mathsf{tr}} U_{n-1}(\tau)|} \cdot \left[\left(\sum_{j=1}^{N} |g_{(j)}|^2 \right) \times \left(\frac{1}{(2\tau_{0n}^{\mathsf{tr}})^2} |g_{(1)}|^{-2} + \sum_{j=2}^{N-1} \frac{1}{(\tau_{jn}^{\mathsf{tr}})^2} |g_{(j+1)}|^{-2} + \frac{1}{(2\tau_{nn}^{\mathsf{tr}})^2} |g_{(N)}|^{-2} \right) \right]^{-1/2},$$

where $U_{n-1}(t)$ is the $(n-1)$th Chebyshev polynomial of the second kind as in (3.22).

(2)

(4.6)
$$\max_{g} \min_{|u_{(1)}|=1} \frac{\|\mathsf{diag}(g)\, V_N^T u\|_2}{\|g\|_2} = \frac{n\omega}{|\tau_{0n}^{\mathsf{tr}} \tau_{nn}^{\mathsf{tr}} U_{n-1}(\tau)|} \left[\frac{1}{|2\tau_{0n}^{\mathsf{tr}}|} + \sum_{j=1}^{n-1} \frac{1}{|\tau_{jn}^{\mathsf{tr}}|} + \frac{1}{|2\tau_{nn}^{\mathsf{tr}}|} \right]^{-1},$$

where the maximum is achieved if and only if for some $\mu > 0$

(4.7)
$$|g_{(j+1)}| = \begin{cases} \mu \sqrt{1/|\tau_{jn}^{\mathsf{tr}}|}, & \text{for } j \in \{0, n\}, \\ \mu \sqrt{2/|\tau_{jn}^{\mathsf{tr}}|}, & \text{for } 1 \le j \le n-1. \end{cases}$$

Proof. $f(z) = \prod_{j=1}^{N} (z - \alpha_j)$ admits

$$f(z) = \eta\, (z - \tau_{0n}^{\mathsf{tr}})(z - \tau_{nn}^{\mathsf{tr}}) U_{n-1}(z/\omega + \tau),$$

where η^{-1} is the coefficient of z^{n-1} in $U_{n-1}(z/\omega + \tau)$. We have

$$\begin{aligned} f(0) &= \eta\, \tau_{0n}^{\mathsf{tr}} \tau_{nn}^{\mathsf{tr}}\, U_{n-1}(\tau), \\ f'(\tau_{0n}^{\mathsf{tr}}) &= -\eta\, (\tau_{0n}^{\mathsf{tr}} - \tau_{nn}^{\mathsf{tr}})\, U_{n-1}(1) \\ &= -\eta\, (\tau_{0n}^{\mathsf{tr}} - \tau_{nn}^{\mathsf{tr}})\, n \\ &= -\eta\, 2n\omega, \\ f'(\tau_{nn}^{\mathsf{tr}}) &= -\eta\, (\tau_{nn}^{\mathsf{tr}} - \tau_{0n}^{\mathsf{tr}})\, U_{n-1}(-1) \\ &= (-1)^n \eta\, (\tau_{nn}^{\mathsf{tr}} - \tau_{0n}^{\mathsf{tr}})\, n \\ &= -(-1)^n \eta\, 2n\omega, \end{aligned}$$

and for $1 \le j \le n-1$,

$$\begin{aligned} f'(\tau_{jn}^{\mathsf{tr}}) &= \eta\,(\tau_{jn}^{\mathsf{tr}} - \tau_{0n}^{\mathsf{tr}})(\tau_{jn}^{\mathsf{tr}} - \tau_{nn}^{\mathsf{tr}})U'_{n-1}(\tau_{jn})/\omega \\ &= \eta\,(\tau_{jn}^{\mathsf{tr}} - \tau_{0n}^{\mathsf{tr}})(\tau_{jn}^{\mathsf{tr}} - \tau_{nn}^{\mathsf{tr}})n/[\omega(1-\tau_{jn}^2)] \\ &= -\eta n\omega. \end{aligned}$$

Therefore by Lemma 4.2, we have (4.5) and (4.6). □

Remark 4.2. As a corollary to (4.6) and the error bound in (1.4), we deduce that the right-hand side of (4.6) is equal to $|T_n(\tau)| = 2\left[\Delta_\kappa^n + \Delta_\kappa^{-n}\right]^{-1}$.

Proof of Theorem 2.2. Item (1) is always true for any given positive definite system $Ax = b$. In fact let $A = \widetilde{Q}\Lambda\widetilde{Q}^*$ be its eigendecomposition, where $\widetilde{Q}$ is unitary, and Λ as in the theorem since A is positive definite. Set $\tilde{g} = \Lambda^{-1/2}\widetilde{Q}^*b$. Define $g = (|\tilde{g}_{(1)}|, |\tilde{g}_{(2)}|, \ldots, |\tilde{g}_{(N)}|)^T \in \mathbb{R}^N$. Then $\tilde{g} = Dg$ for some diagonal D with $|D_{(j,j)}| = 1$. Finally $A = Q\Lambda Q^*$ and $b = Q\Lambda^{1/2}g$ with $Q = \widetilde{Q}D$ still unitary.

Next we notice that

$$\begin{aligned} \|r_k\|_{A^{-1}} &= \min_{x\in\mathcal{K}_k} \|b - Ax\|_{A^{-1}} \\ &= \min_{p_k(0)=1} \|p_k(\Lambda)g\|_2 \\ &= \min_{p_k(0)=1} \sqrt{\sum_{j=1}^{N} |p_k(\lambda_j)|^2 g_{(j)}^2}, \end{aligned} \tag{4.8}$$

where $p_k(z)$ denotes a polynomial of degree no more than k. If either inequality in item (2) is violated, the effective condition number $\kappa' < \kappa(A)$ as far as CG is concerned and the error bound in (1.4) gives

$$\frac{\|r_k\|_{A^{-1}}}{\|r_0\|_{A^{-1}}} \le 2\left[\Delta_{\kappa'}^k + \Delta_{\kappa'}^{-k}\right]^{-1} < 2\left[\Delta_\kappa^k + \Delta_\kappa^{-k}\right]^{-1},$$

contradicting (2.12). This proves item (2).

For item (3), we first claim that λ_j for which $g_{(j)} > 0$ is in $\{\tau_{jk}^{\mathsf{tr}},\ 0 \le j \le k\}$. Otherwise if there was a j_0 such that $g_{(j_0)} > 0$ and $\lambda_{j_0} \notin \{\tau_{jk}^{\mathsf{tr}},\ 0 \le j \le k\}$, then $|T_k(\lambda_{j_0}/\omega + \tau)| < 1$, where ω and τ are given by (2.5). Now take $p_k(z) = q_k(z)$ in (4.8), where $q_k(z) = T_k(z/\omega + \tau)/T_k(\tau)$, to get

$$\begin{aligned} \|r_k\|_{A^{-1}} &\le \sqrt{|q_k(\lambda_{j_0})|^2 g_{(j_0)}^2 + \sum_{j\ne j_0} |q_k(\lambda_j)|^2 g_{(j)}^2} \\ &< |T_k(\tau)|^{-1} \sqrt{g_{(j_0)}^2 + \sum_{j\ne j_0} g_{(j)}^2} \\ &= |T_k(\tau)|^{-1} \|r_0\|_{A^{-1}}, \end{aligned}$$

contradicting (2.12). This proves the claim. On the other hand, since $r_k \ne 0$, there are at least $k+1$ distinct values in $\{\lambda_j : g_{(j)} > 0\}$ and therefore $\{\lambda_j : g_{(j)} > 0\} \supset \{\tau_{jk}^{\mathsf{tr}},\ 0 \le j \le k\}$. Item (3) is proved.

Item (3) says effectively A has $k+1$ distinct eigenvalues as far as CG is concerned and thus $r_{k+1} = 0$. This is item (4). Kaniel [13] gave a different proof of this fact.

Define $\hat{g} \in \mathbb{R}^{k+1}$ by $\hat{g}_{(\ell+1)} = \|g_{\mathbb{J}_\ell}\|_2$. (4.8) gives

$$\|r_k\|_{A^{-1}} = \min_{p_k(0)=1} \sqrt{\sum_{j=0}^{k} |p_k(\tau_{jk}^{\mathsf{tr}})|^2 \hat{g}_{(j+1)}^2} = \min_{|u_{(1)}|=1} \|\mathsf{diag}(\hat{g}) V_{k+1}^T u\|_2,$$

where $V_{k+1} \equiv V_{k+1,k+1}$ is the $(k+1) \times (k+1)$ Vandermonde matrix as defined in (3.2) with nodes $\alpha_{j+1} = \tau_{jk}^{\mathsf{tr}}$ for $0 \le j \le k$. The condition (2.12) and the error bound in (1.4) imply that for $\hat{g}$

$$\min_{|u_{(1)}|=1} \|\mathsf{diag}(\hat{g}) V_{k+1}^T u\|_2 = \max_{h} \min_{|u_{(1)}|=1} \|\mathsf{diag}(h) V_{k+1}^T u\|_2.$$

Lemma 4.3 shows $\hat{g}_{(\ell+1)} = \|g_{\mathbb{J}_\ell}\|_2$ must take the form of (2.13). □

5. Concluding remarks

We have found a closed formula for the CG residuals for Meinardus' examples. These residuals may deviate from the well-known error bounds in (1.4) by a factor no bigger than $1/\sqrt{2}$, indicating the error bounds by (1.4) governing the CG convergence rate is very tight in general. Three key technical components that made our computations possible are as follows:

(1) transforming CG residual computations as minimization problems involving rectangular Vandermonde matrices,
(2) the QR-like decomposition $V_N^T = \boldsymbol{S}_N^T R_N^{-1}$, and
(3) the solution to $\min_{|u_{(1)}|=1} \|Zu\|_2$.

It turns out that QR-like decompositions exist for quite a few Vandermonde matrices, and the combination of the three technical components have been used in [14, 15] for arriving at the asymptotically optimally conditioned real Vandermonde matrices, analyzing the sharpness of existing error bounds for CG and the symmetric Lanczos method for eigenvalue problems.

We completely characterized the extreme positive linear systems for which the kth CG residual achieves the error bound in (1.4). Roughly speaking, as far as CG is concerned, these extreme examples are nothing but one of Meinardus' examples of order $k+1$. As a consequence, unless $N = 2$ there is no positive linear system whose kth CG residual achieves the error bound in (1.4) for all $1 \le k < N$.

Acknowledgment

The author wishes to acknowledge the anonymous referee to draw his attention to Chapter II in [5] by H. Rutishauser.

References

[1] O. Axelsson, *Iterative solution methods*, Cambridge University Press, New York, 1996. MR1276069 (95f:65005)
[2] B. Beckermann, *On the numerical condition of polynomial bases: Estimates for the condition number of Vandermonde, Krylov and Hankel matrices, Habilitationsschrift, Universität Hannover, April 1996; see* `http://math.univ-lille1.fr/~bbecker/abstract/Habilitationsschrift_Beckermann.pdf`.
[3] P. Borwein and T. Erdélyi, *Polynomials and polynomial inequalities*, Graduate Texts in Mathematics, vol. 161, Springer, New York, 1995. MR1367960 (97e:41001)
[4] J. Demmel, *Applied numerical linear algebra*, SIAM, Philadelphia, 1997. MR1463942 (98m:65001)

[5] M. Engeli, Th. Ginsburg, H. Rutishauser, and E. Stiefel, *Refined iterative methods for computation of the solution and the eigenvalues of self-adjoint boundary value problems*, Birkhäuser Verlag, Basel/Stuttgart, 1959. MR0145689 (26:3218)
[6] V. N. Faddeeva, *Computational methods of linear algebra*, Dover Publications, New York, 1959, Translated from the Russia by Curtis D. Benster. MR0400669 (53:4500)
[7] G. H. Golub and C. F. Van Loan, *Matrix computations*, 3rd ed., Johns Hopkins University Press, Baltimore, Maryland, 1996. MR1417720 (97g:65006)
[8] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*, Academic Press, New York, 1980, Corrected and Enlarged Edition prepared by A. Jeffrey, incorporated the fourth edition prepared by Yu. V. Geronimus and M. Yu. Tseytlin, translaed from the Russian by Scripta Technica, Inc. MR1773820 (2001c:00002)
[9] A. Greenbaum, *Comparision of splittings used with the conjugate gradient algorithm*, Numer. Math. **33** (1979), 181–194. MR0549448 (80k:65035)
[10] ———, *Iterative methods for solving linear systems*, SIAM, Philadelphia, 1997. MR1474725 (98j:65023)
[11] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Standards **49** (1952), 409–436. MR0060307 (15,651a)
[12] I. C. F. Ipsen, *Expressions and bounds for the GMRES residual*, BIT **40** (2000), no. 3, 524–535. MR1780406 (2001g:65032)
[13] S. Kaniel, *Estimates for some computational techniques in linear algebra*, Math. Comp. **20** (1966), no. 95, 369–378. MR0234618 (38:2934)
[14] 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, Available at `http://www.ms.uky.edu/~math/MAreport/`.
[15] ———, *Vandermonde matrices with Chebyshev nodes*, Technical Report 2005-02, Department of Mathematics, University of Kentucky, 2005, Available at `http://www.ms.uky.edu/~math/MAreport/`, Lin. Alg. Appl., to appear.
[16] J. Liesen, M. Rozložník, and Z. Strakos, *Least squares residuals and minimal residual methods*, SIAM J. Sci. Comput. **23** (2002), no. 5, 1503–1525. MR1885072 (2003a:65033)
[17] J. Liesen and P. Tichý, *The worst-case GMRES for normal matrices*, BIT **44** (2004), no. 1, 79–98. MR2057363 (2005d:65046)
[18] G. Meinardus, *Über eine Verallgemeinerung einer Ungleichung von L. V. Kantorowitsch*, Numer. Math. **5** (1963), 14–23. MR0160311 (28:3525)
[19] Y. Saad, *Iterative methods for sparse linear systems*, 2nd ed., SIAM, Philadelphia, 2003. MR1990645 (2004h:65002)
[20] G. L. G. Sleijpen and A. van der Sluis, *Further results on the convergence behavior of conjugate-gradients and Ritz values*, Linear Algebra Appl. **246** (1996), 233–378. MR1407670 (97j:65067)
[21] G. W. Stewart and J.-G. Sun, *Matrix perturbation theory*, Academic Press, Boston, 1990. MR1061154 (92a:65017)
[22] L. N. Trefethen and D. Bau, III, *Numerical linear algebra*, SIAM, Philadelphia, 1997. MR1444820 (98k:65002)
[23] K. Zhou, J. C. Doyle, and K. Glover, *Robust and optimal control*, Prentice Hall, 1995.

Department of Mathematics, University of Texas at Arlington, P.O. Box 19408, Arlington, Texas 76019-0408

E-mail address: `rcli@.uta.edu`

URL: `http://www.uta.edu/faculty/rcli`