A NOTE ON CONJUGATE GRADIENT CONVERGENCE

AARON E. NAIMAN*, IVO M. BABUŠKA†, AND HOWARD C. ELMAN‡

Abstract. The one-dimensional discrete Poisson equation on a uniform grid with n points produces a linear system of equations with a symmetric, positive-definite coefficient matrix. Hence, the conjugate gradient method can be used, and standard analysis gives an upper bound of $O(n)$ on the number of iterations required for convergence. This paper introduces a systematically defined set of solutions dependent on a parameter β, and for several values of β, presents exact analytic expressions for the number of steps $k(\beta, \tau, n)$ needed to achieve accuracy τ. The asymptotic behavior of these expressions has the form $O(n^\alpha)$ as $n \to \infty$ and $O(\tau^\gamma)$ as $\tau \to 0$. In particular, two choices of β corresponding to nonsmooth solutions give $\alpha = 0$, i.e., iteration counts independent of n; this is in contrast to the standard bounds. The standard asymptotic convergence behavior, $\alpha = 1$, is seen for a relatively smooth solution. Numerical examples illustrate and supplement the analysis.

Key words. conjugate gradient, convergence rates

AMS subject classifications. 65F10, 65G99, 65L10, 65L12, 65N22

1. Introduction.

The conjugate gradient method is widely used for solving systems of linear equations stemming from the discretization of boundary value problems for elliptic differential equations. Considering the conjugate gradient as an iterative method, the required number of iterations depends in general on the distribution of eigenvalues of the coefficient matrix, and upper estimates are well-known. In particular, the standard analysis leads to the upper bound

$$2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^p \tag{1.1}$$

for the relative error in the energy norm after p steps (see, e.g., [2, p. 525]), where κ is the condition number. For the one-dimensional model problem (see Section 2) on a uniform grid with n points, the condition number is approximately $4n^2/\pi^2$, and this estimate leads to an upper bound of

$$k \approx \frac{\ln \frac{2}{\tau}}{\pi} n \tag{1.2}$$

for the number of iterations to make the relative error smaller than a specified accuracy τ (provided τ is not too small), see [1, p. 567].

These bounds are derived using the inequality

$$\left\|e^{(p)}\right\| \le \min_{P_k} \max_{\text{Eigenvalues } \lambda} |P_p(\lambda)| \left\|e^{(0)}\right\| \tag{1.3}$$

for the relative error in the energy norm, where the minimum is over all polynomials of degree p such that $P_p(0) = 1$. It is known that this result is sharp in the sense that

* Department of Applied Mathematics, Jerusalem College of Technology-Machon Lev, Jerusalem, Israel, e-mail: naiman@math.jct.ac.il.

† Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742. The work of this author was supported in part by the U. S. Office of Naval Research under grant N-0014-90-J1030 and by the National Science Foundation under grant DMS-9120877.

‡ Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, e-mail: elman@cs.umd.edu. The work of this author was supported in part by the U. S. Army Research Office under grant DAAL-0392-G-0016 and by the National Science Foundation under grants ASC-8958544 and DMS-9423133.

for any distribution of at least $p+1$ eigenvalues, there is an initial guess (depending on p) that produces equality in (1.3) [1, p. 561], [4]. This observation can be used to derive sharp upper bounds on the error that are stronger than (1.1) [4].

The conventional wisdom holds that (1.2) characterizes the behavior of the conjugate gradient method for the model problem in terms of both n and τ, except in the case where the initial error contains only a small number of eigenvectors. In this paper, we examine this issue. We introduce a systematically defined set of solutions dependent on a parameter β that controls smoothness, and we study the performance of the conjugate gradient method applied to the discrete one-dimensional model problem when these are the solutions. For several choices of β, we give explicit formulae for the exact number of iterations needed to achieve accuracy τ, as a function of n and τ. Two of these results, corresponding to nonsmooth solutions, show that the number of iterations is independent of n for large n. These results stand in contrast to the standard bounds: In each of the two examples, the initial error contains components in all (n) eigenvectors of the coefficient matrix, but the required iteration counts do not depend on n. Additional analysis together with empirical results for a wide choice of values of β suggest that this type of behavior is typical. We also show, in a third example, that there are members of this solution class for which upper bounds essentially of the form (1.2) are seen. This result differs from (1.2), though, in that the coefficient multiplying n decreases as τ decreases. We do not know whether there is a class of initial guesses whose convergence behavior depends on both n and τ in a manner predicted by (1.2).

In the next section we define the model problem and present the class of solutions under consideration. In Section 3, we briefly review the properties of the conjugate gradient method used. In Section 4 we give, for $\beta = 0$, the expression for the error after p iterations and the required number of iterations, k, for achieving accuracy τ. In Section 5, we derive the analogous formulae for the case $\beta = 1$ and outline some qualitative differences from the results for $\beta = 0$. We present in Section 6 numerical computations for various values of β. Finally, in Section 7 we analyze the case $\beta = 2$ and show that in contrast to the previous two cases, where the iteration counts are independent of the size of the system (for large enough systems), here $k = O(n)$ as $n \to \infty$. The work presented in this paper represents part of the first author's thesis, and additional details can be found there [5].

2. The Model Problem and Solution Class.

Consider the problem

$$-u''(x) = f(x), \quad x \in (0,1), \tag{2.1}$$

with the Dirichlet boundary conditions

$$u(0) = u(1) = 0,$$

and finite difference or finite element approximation on a uniform mesh

$$M := \left\{ x_i = ih \,|\, i = 1, \cdots, n,\ h = \frac{1}{n+1} \right\},$$

i.e., all nodes are internal. The approximate solution $u^M = \left(u_1^M, \cdots, u_n^M\right)^T$ then satisfies the system

$$Au^M = f^M, \tag{2.2}$$

where $f^M = \left(f_1^M, \cdots, f_n^M\right)^T$, $f_i^M = f(x_i)h^2$, for finite differences (or lumped finite elements) and

$$A = \begin{pmatrix} 2 & -1 & 0 & \cdots & \cdots & 0 \\ -1 & 2 & -1 & \ddots & & \vdots \\ 0 & -1 & 2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & \cdots & 0 & -1 & 2 \end{pmatrix}_{n\times n} . \tag{2.3}$$

It is well known that the normalized eigenvectors, $v^{(s)}$, and eigenvalues, $\lambda^{(s)}$, of the matrix A are given by

$$v_i^{(s)} = \sqrt{\frac{2}{n+1}} \sin 2j_{si} \quad \text{and} \quad \lambda^{(s)} = 4\sin^2 j_s,$$

for $s, i = 1, \cdots, n$, where $j_a = \frac{a\pi}{2(n+1)}$. Hence we can write

$$u^M = \sum_{s=1}^{n} \xi^{(s)} v^{(s)},$$

as well as

$$\left|\left|u^M\right|\right|_2^2 = \sum_{s=1}^{n} \left(\xi^{(s)}\right)^2.$$

For any such u^M, the energy norm is

Definition 2.1. $\left|\left|u^M\right|\right|_E^2 \equiv \left(u^M\right)^T A\left(u^M\right) = 4\sum_{s=1}^{n} \left(\xi^{(s)}\right)^2 \sin^2 j_s$.

Instead of analyzing (2.2), we can analyze the number of iterations for the similar system

$$By = g, \tag{2.4}$$

where

$$B = \operatorname{diag}\left(4\sin^2 j_i\right), \quad i = 1, \cdots, n. \tag{2.5}$$

Let $y^\star = \left(y_1^\star, \cdots, y_n^\star\right)^T$, be the exact solution of (2.4) and let the initial guess $y^{(0)} = 0$. Letting $y^{(p)}$ denote the $p^{\underline{\text{th}}}$ conjugate gradient vector, we define the relative error

$$\Theta(y^\star, p) = \frac{\left|\left|y^\star - y^{(p)}\right|\right|_E}{\left|\left|y^\star\right|\right|_E}, \tag{2.6}$$

and we will study the required number of iterations $k(y^\star, \tau)$ needed to make $\Theta(y^\star, k(y^\star, \tau))$ equal to tolerance τ.

The character of $y^\star$ will depend on the smoothness of u^M. If the solution is smooth, then $|y_i^\star|$ will decay with increasing i. To understand the influence of smoothness, we will consider solutions to (2.5) of the form

$$|y_i^\star| := \left[\csc j_i\right]^\beta, \quad i = 1, \cdots, n \tag{2.7}$$

where β is a real parameter. The cases where β is large describe smoother solutions of (2.2), and when β is decreased, the solutions become less smooth. We will also generalize (2.7) further below.

3. Bounds for the Conjugate Gradient Method. We will analyze the conjugate gradient method for the diagonal system (2.5). As in the previous section, let

$$y^\star = (y_1^\star, \cdots, y_n^\star)^T,$$

be the exact solution of (2.4). Then the residual with zero initial guess is

$$r^{(0)} = g - By^{(0)} = g = By^\star,$$

so that by (2.4) we have

$$r_i^{(0)} = y_i^\star \bigl(4\sin^2 j_i\bigr), \;\; i = 1, \cdots, n.$$

The $p^{\underline{\text{th}}}$ iterate computed by the conjugate gradient method is the member of the Krylov sequence

$$\begin{aligned} &\operatorname{span}\{r^{(0)}, Br^{(0)}, B^2r^{(0)}, \cdots, B^{p-1}r^{(0)}\} = \\ &\quad \operatorname{span}\Bigl\{\bigl[y_i^\star\bigl(4\sin^2 j_i\bigr)\bigr], \Bigl[y_i^\star\bigl(4\sin^2 j_i\bigr)^2\Bigr], \Bigl[y_i^\star\bigl(4\sin^2 j_i\bigr)^3\Bigr], \cdots, \Bigl[y_i^\star\bigl(4\sin^2 j_i\bigr)^p\Bigr]\Bigr\}, \\ &\quad i = 1, \cdots, n \end{aligned}$$

that minimizes the energy norm. Here, the notation $[v_i]$ refers to the vector v whose i'th entry is v_i. The i'th component of the error satisfies

$$\left|y_i^\star - y_i^{(p)}\right| = \left|y_i^\star - \sum_{k=1}^{p} a_k y_i^\star \bigl(\sin^2 j_i\bigr)^k\right|, \;\; i = 1, \cdots, n$$

where the factor 4^k has been incorporated into the coefficient a_k. Now, since the iterates are mutually conjugate orthogonal, we can write the energy norm squared of the error of the $p^{\underline{\text{th}}}$ conjugate gradient iterate as

$$\begin{aligned} &\min_{a_k} \sum_{i=1}^{n} |y_i^\star|^2 \sin^2 j_i \left(1 - \sum_{k=1}^{p} a_k \sin^{2k} j_i\right)^2 = \\ (3.1) \qquad &\min_{a_k} \sum_{i=1}^{n} \left[|y_i^\star| \sin j_i \left(1 - \sum_{k=1}^{p} a_k \sin^{2k} j_i\right)\right]^2 . \end{aligned}$$

REMARK 3.1. Note that the 4 associated with the square of the energy norm is and will be left out of all of the minimization equations.

In the following sections, we will derive closed form expressions of (3.1) for several choices of $y^\star$.

4. Error for $\beta = 0$ and Analogous Problems. Before we proceed with the main theorem of this section, let us first state some identities, the first of which is an orthogonality property of the sine function[1].

IDENTITY 4.1.

$$\left[\sum_{i=1}^{n} \sin{(2k+1)j_i} \sin{(2\ell+1)j_i}\right] + \frac{1}{2}\sin\frac{\pi(2k+1)}{2}\sin\frac{\pi(2\ell+1)}{2} = \frac{n+1}{2}\delta_{k\ell},$$

[1] This identity is proven in [5, Appendix 3.B]. Note: the second term is the same as the summand of the first, with $i := n+1$.

for $0 \leq k, \ell \leq n$, *where* $\delta_{k\ell} = \begin{cases} 1, & k = \ell, \\ 0, & k \neq \ell, \end{cases}$ *is the Kronecker delta.*

IDENTITY 4.2.

$$\sum_{k=0}^{p} (2k+1)^2 = \frac{1}{3}(p+1) \left[4(p+1)^2 - 1\right]$$

IDENTITY 4.3.

$$\sum_{k=0}^{p} (-1)^k (2k+1) = (-1)^p (p+1)$$

IDENTITY 4.4.

$$\sum_{i=1}^{n} \sin^2 j_i = \frac{n}{2}$$

Proof. Invoke [3, (1.351.1)]. □

THEOREM 4.5. *For large enough* n*, the relative error and the required number of iterations* $k(y^\star, \tau)$ *for the solution vector* $y^\star$*, given by (2.7) with* $\beta = 0$*, are independent of* n*, with*

$$k(y^\star, \tau) \approx \frac{\frac{3}{4}^{\frac{1}{3}}}{\tau^{\frac{2}{3}}}. \tag{4.1}$$

Proof. In this case, we have to minimize

$$\sum_{i=1}^{n} \left(\sin j_i - \sum_{k=1}^{p} a_k \sin^{2k+1} j_i \right)^2,$$

which is equivalent to minimizing (using the multiple angle formulae, see [3, (1.320.3)])

$$K \stackrel{\text{def}}{=} \sum_{i=1}^{n} \left(\sin j_i - \sum_{k=0}^{p} c_k \sin{(2k+1)j_i} \right)^2, \tag{4.2}$$

subject to the constraint

$$\sum_{k=0}^{p} c_k (2k+1) = 0.$$

This constrained minimization problem can be solved by the Lagrange multiplier technique. To this end, we introduce the Lagrange multiplier λ and seek the stationary point of the function

$$\sum_{i=1}^{n} \left(\sin j_i - \sum_{k=0}^{p} c_k \sin{(2k+1)j_i} \right)^2 + \lambda \sum_{k=0}^{p} c_k (2k+1) \stackrel{\text{def}}{=} \mathcal{J}.$$

With Identity 4.1 in mind, we rewrite $\mathcal{J}$ as

$$\begin{align} \mathcal{J} \;=\; & \sum_{i=1}^{n}\left(\sin j_i - \sum_{k=0}^{p} c_k \sin{(2k+1)j_i}\right)^2 + \notag \\ & \underbrace{\frac{1}{2}\left(\sin\frac{\pi}{2} - \sum_{k=0}^{p} c_k \sin\frac{(2k+1)\pi}{2}\right)^2}_{\equiv\mathcal{A}} + \tag{4.3} \\ & \lambda\sum_{k=0}^{p} c_k(2k+1) - \frac{1}{2}\mathcal{A}^2. \tag{4.4} \end{align}$$

Differentiating, we get

$$\begin{aligned} \frac{\partial\mathcal{J}}{\partial c_\ell} &= -2\sum_{i=1}^{n}\left(\sin j_i - \sum_{k=0}^{p} c_k \sin{(2k+1)j_i}\right)\sin{(2\ell+1)j_i} - \\ &\quad \left(\sin\frac{\pi}{2} - \sum_{k=0}^{p} c_k \sin\frac{(2k+1)\pi}{2}\right)\sin\frac{(2\ell+1)\pi}{2} + \\ &\quad \lambda(2\ell+1) + \sin\frac{(2\ell+1)\pi}{2}\mathcal{A} \\ &= -2\left\{\left[\sum_{i=1}^{n}\sin j_i \sin{(2\ell+1)j_i}\right] + \frac{1}{2}\sin\frac{\pi}{2}\sin\frac{(2\ell+1)\pi}{2}\right\} + \\ &\quad 2\sum_{k=0}^{p} c_k\left\{\left[\sum_{i=1}^{n}\sin{(2k+1)j_i}\sin{(2\ell+1)j_i}\right] + \right. \\ &\quad \left.\frac{1}{2}\sin\frac{(2k+1)\pi}{2}\sin\frac{(2\ell+1)\pi}{2}\right\} + \lambda(2\ell+1) + (-1)^{\ell}\mathcal{A}. \end{aligned}$$

Now, when $\ell = 0$, we have, invoking Identity 4.1,

$$\begin{aligned} \frac{\partial\mathcal{J}}{\partial c_0} &= -2\left(\frac{n+1}{2}\right) + 2\sum_{k=0}^{p} c_k\delta_{k\,0}\frac{n+1}{2} + \lambda + \mathcal{A} \\ &= -(n+1) + c_0(n+1) + \lambda + \mathcal{A} \\ &= -(n+1)\,(1-c_0) + \lambda + \mathcal{A}. \end{aligned}$$

For $\ell \neq 0$, by orthogonality the first term equals zero, and using Identity 4.1, we are left with

$$\begin{aligned} \frac{\partial\mathcal{J}}{\partial c_\ell} &= 2\sum_{k=0}^{p} c_k\delta_{k\,\ell}\frac{n+1}{2} + \lambda(2\ell+1) + (-1)^{\ell}\mathcal{A} \\ &= c_\ell(n+1) + \lambda(2\ell+1) + (-1)^{\ell}\mathcal{A}. \end{aligned}$$

Setting these equations equal to zero, we obtain the following system of equations

$$c_0 = 1 - \frac{\lambda+\mathcal{A}}{n+1}, \text{ and } c_\ell = -\frac{\lambda(2\ell+1) + (-1)^{\ell}\mathcal{A}}{n+1},\ \ \ell = 1,\cdots,p. \tag{4.5}$$

Between the constraint equation and the definition of $\mathcal{A}$ we have two equations with two unknowns, λ and $\mathcal{A}$. From the constraint equation we have

$$0 = 1 - \frac{\mathcal{A}+\lambda}{1+n} + \frac{3\,\mathcal{A} - 3\,(-1)^p\,\mathcal{A} - 3\,(-1)^p\,\mathcal{A}\,p - 11\,\lambda\,p - 12\,\lambda\,p^2 - 4\,\lambda\,p^3}{3\,(1+n)},$$

and from the definition of $\mathcal{A}$

$$\mathcal{A} = \frac{(\mathcal{A} + (-1)^p\,\lambda)\,(1+p)}{1+n}.$$

(Note that for the following calculations, Identities 4.2 and 4.3 are invoked repeatedly.) Solving for the two unknowns, we get

$$\lambda = \frac{3\,(-1)^p\,(1+n)}{3 + 3\,n + 8\,n\,p - 8\,p^2 + 4\,n\,p^2 - 4\,p^3},$$

and

$$\mathcal{A} = \frac{3\,(1+n)\,(n-p)}{(1+p)\,(3 + 3\,n + 8\,n\,p - 8\,p^2 + 4\,n\,p^2 - 4\,p^3)}.$$

These values are now available to be used in c_ℓ in (4.5).

In order to get to the conjugate gradient error, we compute the expression K by way of the equivalent (4.4), for c_k given by (4.5)

$$K = \frac{n+1}{2}\left[(1-c_0)^2 + \sum_{k=1}^{p} c_k^2\right] - \frac{1}{2}\mathcal{A}^2.$$

Substituting for $\{c_k\}$, we obtain

$$K = \frac{3\,(1+n)\,(n-p)}{2\,(1+p)\,(3 + 3\,n + 8\,n\,p - 8\,p^2 + 4\,n\,p^2 - 4\,p^3)}.$$

Recall that for the relative error $\Theta(y^\star, p)$ we have (see (2.6) and Remark 3.1)

$$\Theta^2(y^\star, p) = \frac{4K}{||y^\star||_E^2},$$

where Definition 2.1 and Identity 4.4 supply the value for the denominator. Therefore, performing a series expansion in $\frac{1}{n}$ on the error of the $p^{\underline{\text{th}}}$ iteration, we have for large n

$$\Theta^2(y^\star, p) \approx \frac{3\,p\,(5 + 4\,p)}{n\,(1+p)\,(3 + 8\,p + 4\,p^2)^2} + \frac{3}{3 + 11\,p + 12\,p^2 + 4\,p^3},$$

i.e., it is, basically, independent of n. In Figure 1 we show what the relative error looks like. Note that the first iterations are the most effective.

If we use a stopping criterion of $\Theta(y^\star, p) \le \tau$, we get that the following number of iterations, $k(y^\star, \tau)$, are necessary

$$k(y^\star, \tau) \approx \frac{(3 + \frac{3}{n})^{\frac{1}{3}}}{2^{\frac{2}{3}}\,\tau^{\frac{2}{3}}}. \tag{4.6}$$

For large n we obtain (4.1). □

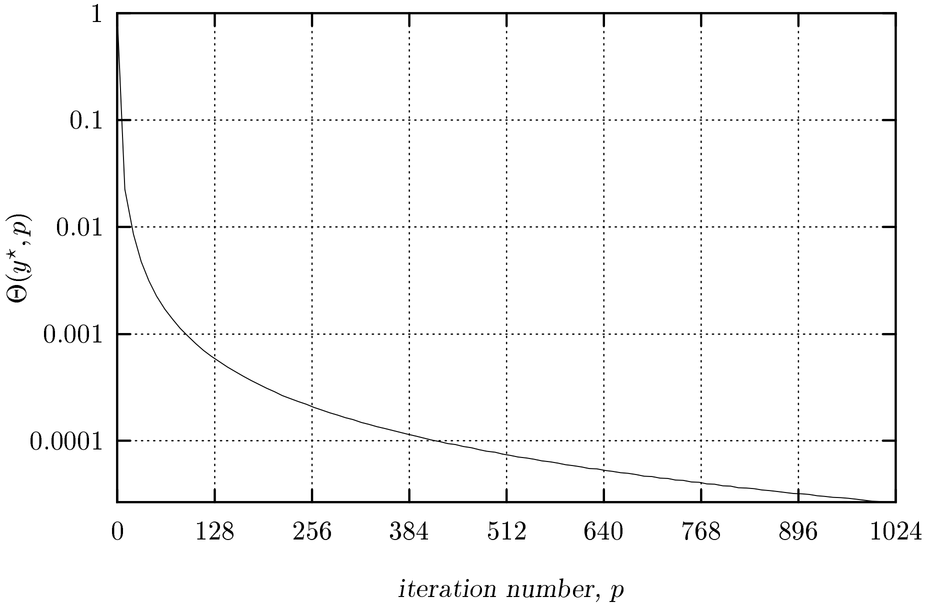

FIG. 1. *Relative error, $\Theta(y^\star, p)$, for $\beta = 0$ and $n = 1024$*

Let us compare this result with the standard bound (1.2). The conjugate gradient method is a direct method, so that it is guaranteed to compute the exact solution in at most n steps. Thus, (1.2) only has meaning if $\tau > 0.086$. However, use of either (1.2) or the finite termination property only guarantees that $k(y^\star, \tau) \leq O(n)$. In contrast, Theorem 4.5 shows that for any fixed τ, the limiting value of $k(y^\star, \tau)$ for $\beta = 0$ is a constant that is independent of n.

So far we have used $|y_i^\star| = 1$, $i = 1, \cdots, n$. If instead we have

$$|y_i^\star| < a, i = 1, \cdots, n, \text{ and } \frac{\sum_{i=1}^n a^2 \sin^2 j_i}{\sum_{i=1}^n (y_i^\star)^2 \sin^2 j_i} \leq \rho^2, \tag{4.7}$$

we also have the following corollary:

COROLLARY 4.6. *The required number of iterations $k(y^\star, \tau)$ for the solution vector $y^\star$ given by (4.7), is independent of n, for large enough n.*

Proof. Since everything in (3.1) is positive, the same minimization is performed as before, generating the same parameters $\{c_k\}$ and minimal value K. There is now an extra a in the numerator of the relative error $\Theta(y^\star, p)$, which, using the second inequality of (4.7), evaluates to

$$\Theta(y^\star, p) \approx \frac{1}{p^{\frac{3}{2}}} \left(\frac{3}{4}\right)^{\frac{1}{2}} \rho.$$

This is also independent of n. □

We take here n to be a power of two, and in Figure 2 we show the required number of iterations for $\tau = 0.01$. The results show that behavior of the type predicted by the standard analysis is seen only for small n. These and all subsequent experimental results were obtained on a CM-2 computer.

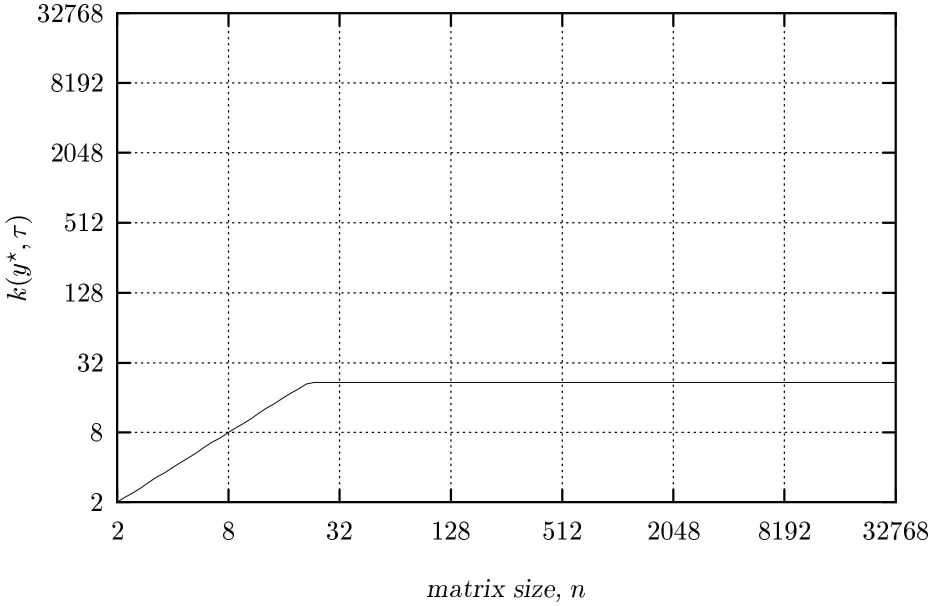

FIG. 2. *Required number of iterations, $k(y^\star, \tau)$, for $\beta = 0$ and $\tau = 0.01$*

5. Error for $\beta = 1$ and Analogous Problems. Once again we preface our theorem with the following identity which is an orthogonality property, this time of the cosine function:

IDENTITY 5.1.

$$\left[\sum_{i=1}^{n} \cos 2j_{\tilde{\ell}i}\right] + \frac{1}{2} + \frac{1}{2}\cos \tilde{\ell}\pi = 0,$$

for $\frac{\tilde{\ell}}{2(n+1)}$ not an integer.

Indeed any such $\tilde{\ell}$ which will concern us will always be smaller than $2(n+1)$ in absolute value, and therefore, we need only deal with $\tilde{\ell} = 0$ as a special case. This relationship can be proven in a straightforward manner using [3, 1.342.2], and expanding $\sin j_{n\ell}$ into $\sin \frac{\pi\ell}{2}\left(1 - \frac{1}{n+1}\right)$, as in [5, Appendix 3.B].

THEOREM 5.2. *For large enough n, the relative error and the required number of iterations $k(y^\star, \tau)$ for the solution vector $y^\star$, given by (2.7) with $\beta = 1$, are independent of n, with*

$$k(y^\star, \tau) \approx \frac{1}{2\,\tau^2}. \tag{5.1}$$

Proof. In this case, we have to minimize the expression (see (3.1))

$$\sum_{i=1}^{n}\left(1 - \sum_{k=1}^{p} a_k \sin^{2k} j_i\right)^2,$$

which, invoking [3, (1.320.1)], is equivalent to minimizing

$$K \overset{\text{def}}{=} \sum_{i=1}^{n}\left(1 - \sum_{k=0}^{p} c_k \cos 2kj_i\right)^2, \tag{5.2}$$

Once again, due to the one additional constant introduced, we need a constraint equation to interrelate the parameters $\{c_k\}$. The second summation in (5.2), evaluated at zero, produces the desired interdependence. The original form of the minimization tells us that this is equal to zero. Therefore, we have

$$\sum_{k=0}^{p} c_k = 0.$$

As before, using the Lagrange multiplier technique we seek the stationary point of

$$\begin{aligned}\mathcal{J} &= \sum_{i=1}^{n}\left(1-\sum_{k=0}^{p} c_k\cos 2kj_i\right)^2 + \frac{1}{2}\left(1-\sum_{k=0}^{p} c_k\right)^2 + \\ &\quad \frac{1}{2}\left(1-\sum_{k=0}^{p} c_k\cos k\pi\right)^2 + \lambda\sum_{k=0}^{p} c_k - \frac{1}{2}\mathcal{B}^2 - \frac{1}{2}\mathcal{A}^2,\end{aligned}$$

where $\mathcal{B} = 1-\sum_{k=0}^{p} c_k$ and $\mathcal{A} = 1-\sum_{k=0}^{p} c_k\cos k\pi$. We note immediately that because of the constraint equation, $\mathcal{B}=1$.

We have

$$\begin{aligned}\frac{\partial\mathcal{J}}{\partial c_\ell} &= \sum_{i=1}^{n}\left[-2\cos 2j_{\ell i}\left(1-\sum_{k=0}^{p} c_k\cos 2kj_i\right)\right] - \\ &\quad \left(1-\sum_{k=0}^{p} c_k\right) - \cos\ell\pi\left(1-\sum_{k=0}^{p} c_k\cos k\pi\right) + \lambda + 1 + \cos\ell\pi\mathcal{A}.\end{aligned}$$

Rewriting the cosine product terms (see [3, 1.314.3]), we get

$$\begin{aligned}\frac{\partial\mathcal{J}}{\partial c_\ell} &= \sum_{i=1}^{n}\left\{-2\cos 2j_{\ell i} + \sum_{k=0}^{p} c_k\left[\cos 2j_{(k+\ell)i} + \cos 2j_{(k-\ell)i}\right]\right\} - \\ &\quad 1 + \sum_{k=0}^{p} c_k - \cos\ell\pi + \frac{1}{2}\sum_{k=0}^{p} c_k[\cos(k+\ell)\pi + \cos(k-\ell)\pi] + \\ &\quad \lambda + 1 + (-1)^\ell\mathcal{A}.\end{aligned}$$

and rearranging

$$\begin{aligned}\frac{\partial\mathcal{J}}{\partial c_\ell} &= -2\left\{\left[\sum_{i=1}^{n}\cos 2j_{\ell i}\right] + \frac{1}{2}\cos\ell\pi\right\} + \\ &\quad \sum_{k=0}^{p} c_k\left\{\left[\sum_{i=1}^{n}\cos 2j_{(k+\ell)i}\right] + \frac{1}{2}\cos(k+\ell)\pi\right\} + \\ &\quad \sum_{k=0}^{p} c_k\left\{\left[\sum_{i=1}^{n}\cos 2j_{(k-\ell)i}\right] + \frac{1}{2}\cos(k-\ell)\pi\right\} - \\ &\quad 1 + \sum_{k=0}^{p} c_k + \lambda + 1 + (-1)^\ell\mathcal{A}.\end{aligned}$$

Therefore, for $\ell = 0$ we have

$$\begin{aligned}
\frac{\partial \mathcal{J}}{\partial c_0} &= -2\left\{n+\frac{1}{2}\right\} + \underbrace{2c_0\left\{n+\frac{1}{2}\right\}}_{k=0} + \underbrace{2\sum_{k=1}^{p} c_k\left\{-\frac{1}{2}\right\}}_{k\neq 0} - \\
&\qquad 1 + c_0 + \sum_{k=1}^{p} c_k + \lambda + 1 + \mathcal{A} \\
&= -2(1-c_0)\,(n+1) + \lambda + 1 + \mathcal{A},
\end{aligned}$$

and for non-zero ℓ, we have

$$\begin{aligned}
\frac{\partial \mathcal{J}}{\partial c_\ell} &= 1 + \sum_{k=0}^{p} c_k\left\{-\frac{1}{2}\right\} + \underbrace{c_\ell\left(n+\frac{1}{2}\right)}_{k=\ell} + \sum_{\substack{k=0\\ k\neq \ell}}^{p} c_k\left\{-\frac{1}{2}\right\} - \\
&\qquad 1 + c_\ell + \sum_{\substack{k=0\\ k\neq \ell}}^{p} c_k + \lambda + 1 + (-1)^{\ell}\mathcal{A} \\
&= c_\ell(n+1) + \lambda + 1 + (-1)^{\ell}\mathcal{A}.
\end{aligned}$$

After setting these two equations equal to zero, we write $\{c_\ell\}$ in terms of our two unknowns λ and $\mathcal{A}$

$$c_0 = 1 - \frac{\lambda+1+\mathcal{A}}{2(n+1)}, \text{ and } c_\ell = -\frac{\lambda+1+(-1)^{\ell}\mathcal{A}}{n+1}, \quad \ell = 1,\cdots,p. \tag{5.3}$$

With the combination of the constraint equation and the definition of $\mathcal{A}$, we have two equations with two unknowns

$$0 = 1 - \frac{1+\mathcal{A}+\lambda}{2\,(1+n)} + \frac{\mathcal{A}-(-1)^{p}\,\mathcal{A}-2\,p-2\,\lambda\,p}{2\,(1+n)},$$

and

$$\mathcal{A} = \frac{1+\mathcal{A}+\lambda}{2\,(1+n)} - \frac{1-(-1)^{p}+\lambda-(-1)^{p}\,\lambda-2\,\mathcal{A}\,p}{2\,(1+n)}.$$

Solving, we get the following values for the unknowns:

$$\lambda = \frac{2\,\left(n+n^2-p-2\,n\,p+p^2\right)}{1+n+2\,n\,p-2\,p^2},$$

and

$$\mathcal{A} = \frac{(-1)^{p}\,(1+n)}{1+n+2\,n\,p-2\,p^2}.$$

These values can be substituted into the quantities c_ℓ of (5.3).

Now that the c_ℓ coefficients are known, the next step is to compute K. We start by expanding the squared terms, giving

$$\begin{aligned} K \;=\; & \sum_{i=1}^{n}\left(1-2\sum_{k=0}^{p} c_k \cos 2j_{ki} + \sum_{k=0}^{p}\sum_{\ell=0}^{p} c_k c_\ell \cos 2j_{ki}\cos 2j_{\ell i}\right) + \\ & \frac{1}{2}\left(1-2\sum_{k=0}^{p} c_k + \sum_{k=0}^{p}\sum_{\ell=0}^{p} c_k c_\ell\right) + \\ & \frac{1}{2}\left(1-2\sum_{k=0}^{p} c_k \cos k\pi + \sum_{k=0}^{p}\sum_{\ell=0}^{p} c_k c_\ell \cos k\pi \cos \ell\pi\right) - \frac{1}{2}\big(1+\mathcal{A}^2\big). \end{aligned}$$

Invoking [3, 1.314.3] we have

$$\begin{aligned} K \;=\; & n+\frac{1}{2}+\frac{1}{2}-2\sum_{k=0}^{p} c_k\left[\left(\sum_{i=1}^{n}\cos 2j_{ki}\right)+\frac{1}{2}+\frac{1}{2}\cos k\pi\right] + \\ & \frac{1}{2}\sum_{k=0}^{p}\sum_{\ell=0}^{p} c_k c_\ell \left\{\left[\left(\sum_{i=1}^{n}\cos 2j_{(k+\ell)i}\right)+\frac{1}{2}+\frac{1}{2}\cos{(k+\ell)\pi}\right] + \right. \\ & \left.\left[\left(\sum_{i=1}^{n}\cos 2j_{(k-\ell)i}\right)+\frac{1}{2}+\frac{1}{2}\cos{(k-\ell)\pi}\right]\right\} - \frac{1}{2}\big(1+\mathcal{A}^2\big) \\ =\; & (n+1)-2[c_0(n+1)] + \underbrace{\frac{1}{2}c_0^2(n+1)2}_{k=\ell=0} + \underbrace{\frac{1}{2}\sum_{k=1}^{p} c_k^2(n+1)}_{k=\ell\neq 0} - \\ & \frac{1}{2}\big(1+\mathcal{A}^2\big) \\ =\; & (n+1)\left(1-2c_0+c_0^2+\frac{1}{2}\sum_{k=1}^{p} c_k^2\right)-\frac{1}{2}\big(1+\mathcal{A}^2\big), \end{aligned}$$

and we get

$$K=\frac{n+n^2-p-2\,n\,p+p^2}{1+n+2\,n\,p-2\,p^2}.$$

Recall that for the relative error squared: $\Theta^2(y^\star,p)=\frac{4K}{||y^\star||_E^2}$. By Definition 2.1 the denominator is just $4n$. In Figure 3 we graph the relative error as a function of the iteration number. Dividing through and retaining the first two terms (the $O(1)$ and $O\big(\frac{1}{n}\big)$ terms), we get

$$\Theta^2(y^\star,p)\approx\frac{-2\,p^2}{n\,(1+2\,p)^2}+\frac{1}{1+2\,p}.$$

To make $\Theta(y^\star,p)$ equal to a tolerance τ,

$$k(y^\star,\tau)\approx\frac{n}{1+2\,n\,\tau^2} \tag{5.4}$$

iterations are needed. For large n, we obtain (5.1). □

In Figure 4 we show the required number of iterations for $\tau=0.01$. Again, the "classical behavior" is seen for small enough n, but as $n\to\infty$, the iteration counts tend to a constant. Note that the number of iterations is large compared with the

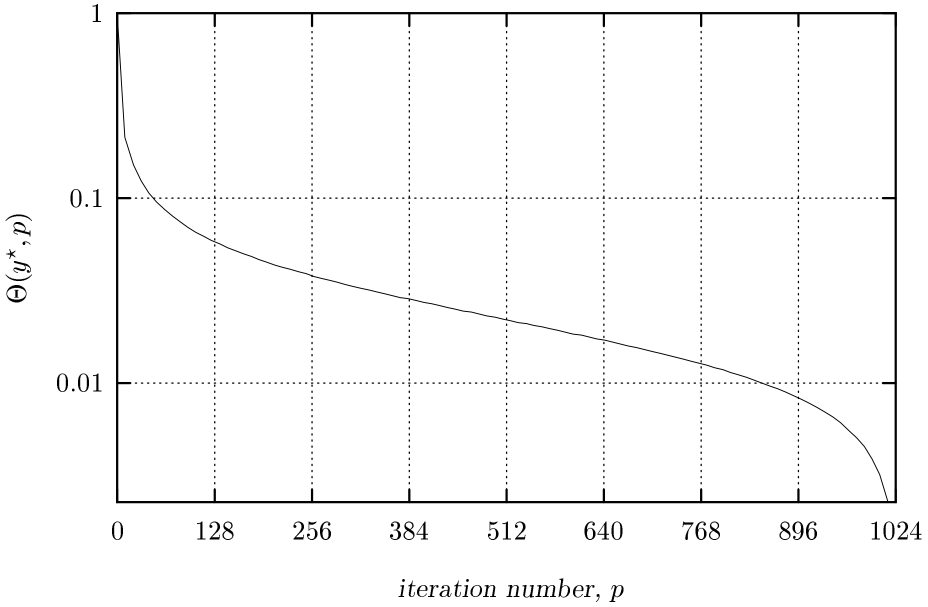

FIG. 3. *Relative error, $\Theta(y^\star, p)$, for $\beta = 1$ and $n = 1024$*

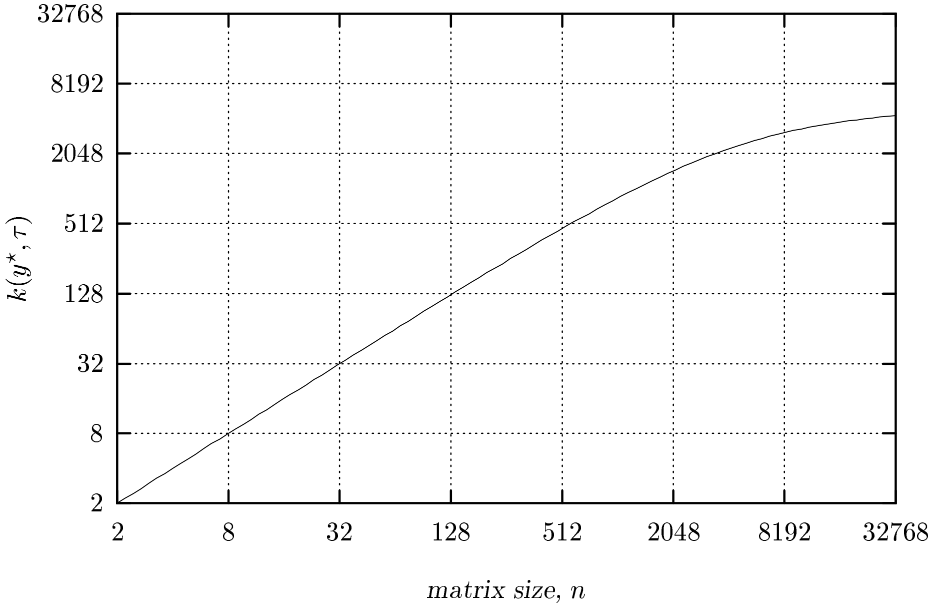

FIG. 4. *Required number of iterations, $k(y^\star, \tau)$, for $\beta = 1$ and $\tau = 0.01$*

case $\beta = 0$ (Figure 2). This is due to the less effective initial iterations, as seen in the difference between Figures 1 and 3.

Some remarks are in place:

1) We see the importance of retaining the $O\left(\frac{1}{n}\right)$ term in the calculation of $\Theta^2(y^\star, p)$, because it is not (necessarily) negligibly small compared with the first term.
2) There is a qualitative difference between this result and the one for $\beta = 0$.

Here, the iteration count (5.4) does not approach a constant until n is large relative to $\frac{1}{\tau^2}$. In contrast, for $\beta = 0$, constant iteration counts are obtained in (4.6) for values of n that are independent of τ. This explains the slower development of the "knee" in Figure 4 than in Figure 2. (In fact, exactly n steps were performed for small n in Figure 4.)

As in the previous section, we can generalize the form of $y^\star$

$$|y_i^\star| < a \csc j_i, i = 1, \cdots, n, , \text{ and } \frac{na^2}{\sum_{i=1}^n (y_i^\star)^2 \sin^2 j_i} \leq \rho^2, \tag{5.5}$$

and obtain the following corollary:

Corollary 5.3. *The required number of iterations $k(y^\star, \tau)$ for the solution vector $y^\star$ given by 5.5, is independent of n, for large enough n.*

6. Empirical Results. In the previous sections we have *analytically* derived the required number of iterations $k(y^\star, \tau)$ for $y^\star$ given in (2.7) and $\beta = 0, 1$. Here, we show empirically computed values of $k(y^\star, \tau)$ for a large collection of values of β. In particular, Figure 5 plots k for values of β between -3 and 10 and $n \leq 32768$, for $\tau = 10^{-2}$, and Figure 6 shows analogous results for $\tau = 10^{-5}$. Figures 7 and 8 give detailed pictures of these results for $n = 32768$.

The results for $\beta = 0$ and $\beta = 1$ are exactly as predicted by the analysis of Sections 4 and 5. We summarize this with the following theorem:

Theorem 6.1. *For β equal 0 or 1, for any τ, there exists $\bar{k}$ (depending on τ) such that $k(y^\star, \tau) < \bar{k}$ for all n.*

It also seems that similar behavior is seen for most other values of β, that is, as n increases, the required number of iterations stops increasing and becomes independent of n. However, scrutiny of Figures 7 and 8 suggests that this may not be the case for β near 2. In the next section we derive a closed for expression for $k(y^\star, \tau)$ when $\beta = 2$ showing that in this case the iteration counts are proportional to n.

7. Error for $\beta = 2$. As usual, we start with a few identities. The first one is the same as Identity 4.1, just shifted:

Identity 7.1.

$$\left[\sum_{i=1}^n \sin{(2k-1)j_i} \sin{(2\ell-1)j_i}\right] + \frac{1}{2} \sin \frac{\pi(2k-1)}{2} \sin \frac{\pi(2\ell-1)}{2} = \frac{n+1}{2}\delta_{k\ell},$$

for $1 \leq k, \ell \leq n+1$.

Note that the second term on the left hand side is equal to $\frac{1}{2}(-1)^{k+\ell}$. The following two identities come up in the derivation below. These identities were proven with the aid of *Mathematica* [6]. Recall that $j_i = \frac{i\pi}{2(n+1)}$.

Identity 7.2.

$$\sum_{i=1}^n \frac{\sin{(2k-1)j_i}}{\sin j_i} = n + 2\left(1 - \left\lceil \frac{k}{2} \right\rceil\right),$$

for $k \leq n$.

Identity 7.3.

$$\sum_{i=1}^n \csc^2 j_i = \frac{2}{3}n^2 + \frac{4}{3}n.$$

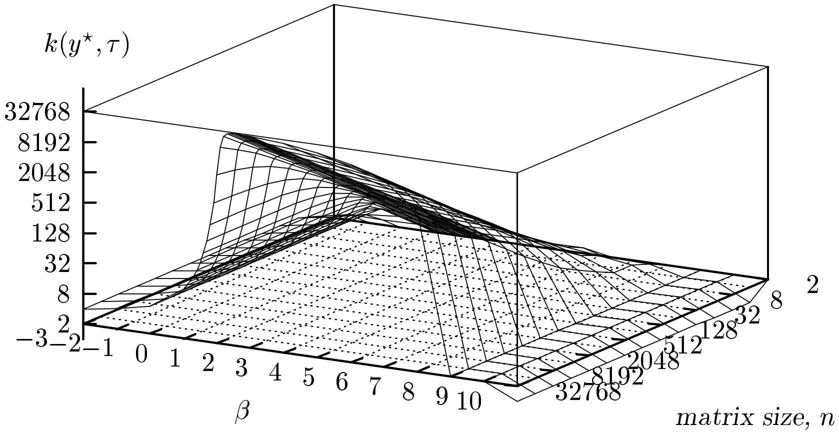

FIG. 5. *Required number of iterations for* $\left|y_i^\star\right| = [\csc j_i]^\beta$ *and* $\tau = 10^{-2}$

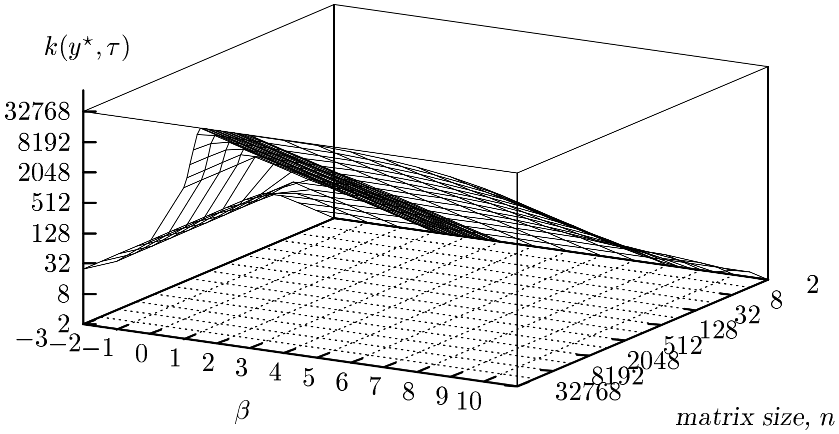

FIG. 6. *Required number of iterations for* $\left|y_i^\star\right| = [\csc j_i]^\beta$ *and* $\tau = 10^{-5}$

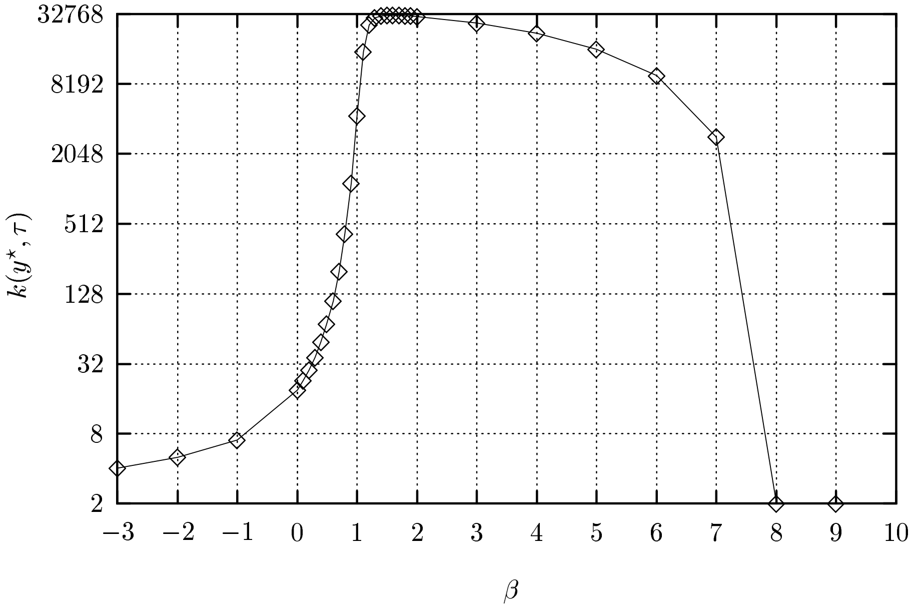

FIG. 7. *Req. num. of iter. for* $\left|y_i^\star\right| = [\csc j_i]^\beta$*, n = 32768 and* $\tau = 10^{-2}$

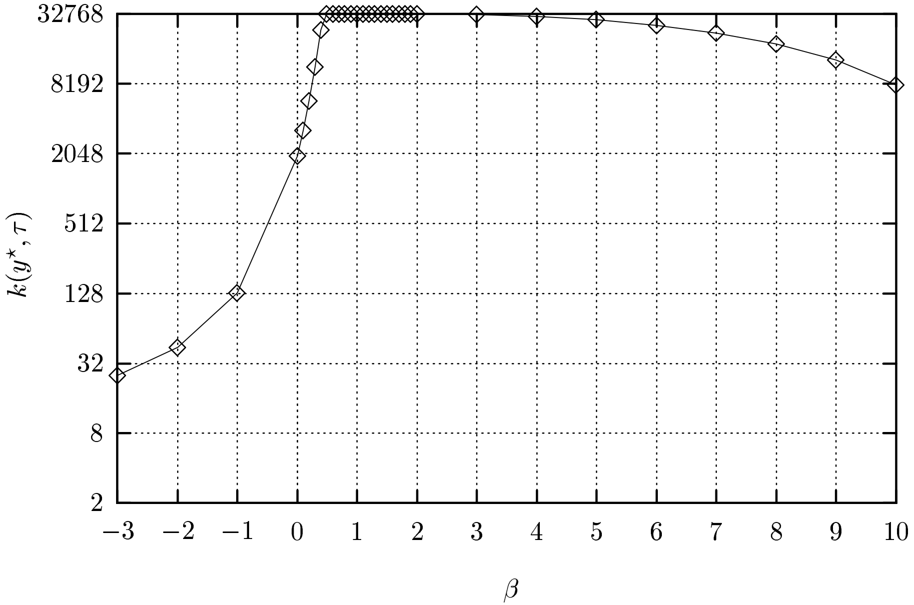

FIG. 8. *Req. num. of iter. for* $\left|y_i^\star\right| = [\csc j_i]^\beta$*, n = 32768 and* $\tau = 10^{-5}$

We now derive an expression for $k(y^\star, \tau)$ for $\beta = 2$.

THEOREM 7.4. *For all n, the relative error and the required number of iterations $k(y^\star, \tau)$ for the solution vector $y^\star$, given by (2.7) with $\beta = 2$, are dependent on n, with, for large n*

$$k(y^\star, \tau) \approx \frac{1-\tau^2}{3} n. \tag{7.1}$$

Proof. In this case, we have the following minimization problem:

$$K \overset{\text{def}}{=} \min_{a_k} \sum_{i=1}^{n} \left(\csc j_i - \sum_{k=1}^{p} a_k \sin^{2k-1} j_i \right)^2,$$

which is equivalent to (using the multiple angle formulae)

$$K = \min_{c_k} \sum_{i=1}^{n} \left(\csc j_i - \sum_{k=1}^{p} c_k \sin{(2k-1)} j_i \right)^2. \tag{7.2}$$

Once again realize that solving for K will exactly determine the conjugate gradient iterate errors. Note that no additional constants were introduced, and therefore no constraint equation is necessary.

This case can be handled using a straightforward minimization, setting derivatives equal to zero. In order to invoke Identity 7.1, we add and subtract half of the summand term, for $i := n+1$, i.e.

$$\pm\frac{1}{2}\left(\csc \frac{\pi}{2} - \sum_{k=1}^{p} c_k \sin{(2k-1)}\frac{\pi}{2} \right)^2 = \pm\frac{1}{2}\underbrace{\left(1 - \sum_{k=1}^{p} c_k (-1)^{k+1} \right)^2}_{\overset{\text{def}}{=} \mathcal{A}}.$$

Differentiating, we get

$$\begin{aligned}
\frac{\partial K}{\partial c_\ell} &= \left[-2\sum_{i=1}^{n} \left(\csc j_i - \sum_{k=1}^{p} c_k \sin{(2k-1)} j_i \right) \sin{(2\ell-1)} j_i \right] - \\
&\quad \left(1 - \sum_{k=1}^{p} c_k (-1)^{k+1} \right) (-1)^{\ell+1} + \mathcal{A}(-1)^{\ell+1} \\
&= -2\left\{ \left[\sum_{i=1}^{n} \csc j_i \sin{(2\ell-1)} j_i \right] + \frac{1}{2}(-1)^{\ell+1} \right\} + \\
&\quad 2\sum_{k=1}^{p} c_k \left\{ \left[\sum_{i=1}^{n} \sin{(2k-1)} j_i \sin{(2\ell-1)} j_i \right] + \frac{1}{2}(-1)^{k+\ell} \right\} + \\
&\quad (-1)^{\ell+1}\mathcal{A}.
\end{aligned}$$

Breaking down to different values of ℓ, and invoking Identity 7.1

$$\begin{aligned}
\ell = 1: \quad \frac{\partial K}{\partial c_1} &= -2\left\{n + \frac{1}{2}\right\} + 2\sum_{k=1}^{p} c_k \left\{\frac{n+1}{2}\delta_{k1}\right\} + (-1)^{\ell+1}\mathcal{A} \\
&= -2n + c_1(n+1) + \mathcal{A} - 1, \\
\ell \neq 1: \quad \frac{\partial K}{\partial c_\ell} &= -2\underbrace{\left[\sum_{i=1}^{n} \frac{\sin{(2\ell-1)j_i}}{\sin j_i}\right]}_{\equiv B(n,\ell)} - (-1)^{\ell+1} + \\
&\quad 2\sum_{k=1}^{p} c_k \left\{\frac{n+1}{2}\delta_{k1}\right\} + (-1)^{\ell+1}\mathcal{A} \\
&= -2B(n,\ell) + c_\ell(n+1) + (-1)^{\ell+1}(\mathcal{A}-1).
\end{aligned}$$

Note that $B(n,1) = n$, therefore we can combine the two possibilities above, producing

$$c_\ell = \frac{1}{n+1}\Big[2B(n,\ell) + (-1)^{\ell}(\mathcal{A}-1)\Big], \quad \ell = 1, \cdots, p.$$

Now, getting back to our definition of $\mathcal{A}$, we have with some manipulation

$$\begin{aligned}
\mathcal{A} &\equiv 1 - \sum_{k=1}^{p} c_k(-1)^{k+1} \\
&= 1 - \frac{1}{n+1}\sum_{k=1}^{p}\Big[2B(n,k)(-1)^{k+1} - (\mathcal{A}-1)\Big] \\
&= 1 - \frac{2}{n-p+1}\sum_{k=1}^{p}(-1)^{k+1}B(n,k)
\end{aligned}$$

Invoking Identity 7.2, we see that for p even: $\mathcal{A} = 1$, and for p odd:

$$\mathcal{A} = 1 - \frac{2}{n-p+1}B(n,p) = -1.$$

Therefore, we have

$$\mathcal{A} = (-1)^p, \text{ and } c_\ell = \frac{1}{n+1}\Big\{2B(n,\ell) + (-1)^{\ell}[(-1)^p - 1]\Big\}.$$

We now have $\mathcal{A}$ and c_ℓ in order to plug back into (7.2) and again invoking Identity 7.1

$$\begin{aligned} K &= \sum_{i=1}^{n} \left[\csc^2 j_i - 2\csc j_i \sum_{k=1}^{p} c_k \sin(2k-1) + \right. \\ &\quad \left. \sum_{k=1}^{p}\sum_{\ell=1}^{p} c_k c_\ell \sin(2k-1)\sin(2\ell-1) \right] + \\ &\quad \frac{1}{2} - \sum_{k=1}^{p} c_k(-1)^{k+1} + \frac{1}{2}\sum_{k=1}^{p}\sum_{\ell=1}^{p} c_k c_\ell (-1)^{k+\ell} - \frac{1}{2}\underbrace{\mathcal{A}^2}_{=1} \\ &= \sum_{i=1}^{n} \csc^2 j_i - 2\sum_{k=1}^{p} c_k \underbrace{\sum_{i=1}^{n} \frac{\sin(2k-1)j_i}{\sin j_i}}_{=B(n,k)} + \\ &\quad \sum_{k=1}^{p}\sum_{\ell=1}^{p} c_k c_\ell \frac{n+1}{2}\delta_{k\ell} - \sum_{k=1}^{p} c_k(-1)^{k+1} \\ &= \frac{2}{3}n^2 + \frac{4}{3}n + \sum_{k=1}^{p}\left[c_k^2\frac{n+1}{2} - 2c_k B(n,k) + (-1)^k c_k \right], \end{aligned}$$

where the first two terms of the last expression are from Identity 7.3.

Now, let us take p to be even. Therefore, $c_\ell = \frac{2B(n,\ell)}{n+1}$, and we have

$$\begin{aligned} K &= \frac{2}{3}n^2 + \frac{4}{3}n + \\ &\quad \sum_{k=1}^{p}\left[\frac{2}{n+1}B^2(n,k) - 2\frac{2}{n+1}B^2(n,k) + (-1)^k\frac{2}{n+1}B(n,k) \right] \\ &= \frac{2}{3}n^2 + \frac{4}{3}n - \frac{2}{n+1} \times \\ &\quad \left\{ \sum_{k=1,\text{odd}}^{p-1} \left[B^2(n,k) + B(n,k)\right] + \sum_{k=2,\text{even}}^{p} \left[B^2(n,k) - B(n,k)\right] \right\}. \end{aligned}$$

Using Identity 7.2 we have

$$\begin{aligned} K &= \frac{2}{3}n^2 + \frac{4}{3}n - \frac{2}{n+1} \times \\ &\quad \left\{ \sum_{k=1,\text{odd}}^{p-1} \left[(n-k+1)^2 + (n-k+1)\right] + \right. \\ &\quad \left. \sum_{k=2,\text{even}}^{p} \left[(n-k+2)^2 - (n-k+2)\right] \right\}, \end{aligned}$$

and renumbering the summation index k

$$\begin{aligned}
K &= \frac{2}{3}n^2 + \frac{4}{3}n - \frac{2}{n+1} \times \\
&\quad \sum_{k=1}^{p/2} \Big\{ [n-2(k-1)]^2 + [n-2(k-1)] + (n-2k+2)^2 - (n-2k+2) \Big\} \\
&= \frac{2}{3}n^2 + \frac{4}{3}n - \frac{4}{n+1} \sum_{k=1}^{p/2} (n-2k+2)^2 \\
&\ \vdots \\
&= \frac{4\,n}{3} + \frac{2\,n^2}{3} - \frac{2\,p\,\left(2 + 6\,n + 3\,n^2 - 3\,p - 3\,n\,p + p^2\right)}{3\,(1+n)} \qquad (7.3)
\end{aligned}$$

(We have once again used *Mathematica* to obtain to the last step.) Note that the choice $p = n$ gives $K = 0$, i.e., finite termination is included in this derivation.

For the relative error squared, we have $\Theta^2(y^\star, p) = \frac{4K}{||y^\star||_E^2}$, and by Definition 2.1 and Identity 7.3 the denominator is $4\big(\frac{2}{3}n^2 + \frac{4}{3}n\big)$. Dividing this into (7.3) and retaining only the high order terms, we get

$$\Theta^2(y^\star, p) \approx 1 - \frac{3\,p}{n}.$$

Thus

$$k(y^\star, \tau) \approx \frac{1-\tau^2}{3} n$$

iterations are needed to make $\Theta(y^\star, p)$ equal to a tolerance τ, as in (7.1). □

We remark that the above estimates are more accurate than the classical estimate of (1.2). It is not clear whether there is a β such that (1.2) is achieved.

8. Summary. In this paper we have analyzed the number of iterations required by the conjugate gradient method for solving linear systems which stem from discretized, one-dimensional boundary value problems of second order. For a specific, parameterized family of initial guesses, exact, analytic expressions were derived for three values of this parameter. Two values lead to iteration counts which, while very different from each other, are independent of the size of the system, for large enough systems. The third value displays the well-known linear dependence, while providing a more accurate estimate than the classical one. Numerical computations were included.

REFERENCES

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

[2] G. H. Golub and C. F. V. Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, second ed., 1989.

[3] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products*, Academic Press, New York, NY, forth ed., 1980.

[4] A. Greenbaum, *Comparison of splittings used with the conjugate gradient algorithm*, Numerische Mathematik, 33 (1979), pp. 181–194.

[5] A. E. Naiman, *Computer Solutions of Finite Element Linear Systems*, PhD thesis, University of Maryland, College Park, MD, 1994.

[6] S. Wolfram, *Mathematica: A System for Doing Mathematics by Computer*, Addison-Wesley, Readwood City, CA, second ed., 1991.