

BOUNDS FOR THE ENTRIES OF MATRIX FUNCTIONS WITH APPLICATIONS TO PRECONDITIONING *

MICHELE BENZI^{1 †} and GENE H. GOLUB^{2 ‡}

¹*Scientific Computing Group, CIC-19, Los Alamos National Laboratory
MS B256, Los Alamos, NM 87545, USA. email: benzi@lanl.gov*

²*Department of Computer Science, Stanford University, Gates 2B MC 9025
Stanford, CA 94305-9025, USA. email: golub@sccm.stanford.edu*

Abstract.

Let A be a symmetric matrix and let f be a smooth function defined on an interval containing the spectrum of A . Generalizing a well-known result of Demko, Moss and Smith on the decay of the inverse we show that when A is banded, the entries of $f(A)$ are bounded in an exponentially decaying manner away from the main diagonal. Bounds obtained by representing the entries of $f(A)$ in terms of Riemann–Stieltjes integrals and by approximating such integrals by Gaussian quadrature rules are also considered. Applications of these bounds to preconditioning are suggested and illustrated by a few numerical examples.

AMS subject classification: 65F10, 65F30, 15A.

Key words: Matrix functions, quadrature rules, Lanczos process, band matrices, exponential decay, preconditioned conjugate gradients.

1 Introduction.

Many problems in numerical linear algebra can be formulated in terms of finding suitable approximations to certain matrix functions. For example, finding a good preconditioner to be used with an iterative method for the solution of $Ax = b$ can be viewed as the problem of approximating the matrix function $f(A) = A^{-1}$, subject to certain constraints that must be imposed in order for the preconditioner to be useful in practice. Another important problem is the approximation of the matrix exponential $f(A) = e^{-tA}$, $t \geq t_0$, which arises in the solution of semidiscretized parabolic equations. In other cases, it is necessary to approximate scalar functions of matrices: see [3, 4] for bounds on the trace of the inverse $\text{tr}(A^{-1})$ and the determinant $\det(A)$.

Methods for deriving bounds and estimates for smooth matrix functions $f(A)$ where A is a symmetric matrix have been described in a series of papers by

*Received February 1998. Revised November 1998. Communicated by Lothar Reichel.

[†]Research supported in part by the Department of Energy through grant W-7405-ENG-36 with Los Alamos National Laboratory.

[‡]Research supported in part by NSF grant CCR-9505393.

Golub and collaborators: see, e.g., [15, 22, 23, 24, 26, 28]. The approach in these papers is based on the integral representation of bilinear expressions of the type

$$u^T f(A)v$$

where f is defined on an interval containing the spectrum of A and u, v are n -vectors. Bounds and estimates on the bilinear form can be obtained by approximating the integral with Gaussian quadrature rules. In turn, these quadrature rules can be computed by means of the Lanczos process.

The choice $f(\lambda) = \lambda^{-1}$, $u = e_i$ and $v = e_j$ corresponds to the (i, j) entry of A^{-1} . Hence, quadrature rules and the Lanczos process can be used to obtain bounds and estimates on individual entries of the inverse [24]. In principle, these estimates can be used to determine explicit approximate inverse preconditioners for use with iterative methods. This is a possible application that has not been previously considered.

When A is a symmetric positive definite (SPD) band matrix, a result of Demko, Moss and Smith [18] states that the entries of A^{-1} are bounded in an exponentially decaying manner away from the main diagonal. Decay is usually fast if A is diagonally dominant. This result suggests that a good approximation to A^{-1} with a banded sparsity pattern may be feasible. In this paper we show that a similar exponential decay bound holds, more generally, for the entries of $f(A)$ when f is analytic and A is any symmetric band matrix. This means that banded approximations to $f(A)$ are justified in many cases.

This decay result, while interesting in itself, is qualitative in nature and provides bounds that are generally too pessimistic to be useful in practice. In contrast, the bounds and estimates from quadrature rules can be quite accurate. As we shall see, such bounds can be useful in the derivation of preconditioners.

The paper is organized as follows. The exponential decay bound for the entries of analytic functions of band matrices is given in Section 2. In Section 3 we summarize the results in [24] on the use of quadrature rules and the Lanczos process for obtaining bounds on the entries of matrix functions. In Section 4 we propose two sample applications to preconditioning, and in Section 5 we present our conclusions.

This paper is based on the technical report [5], where a more self-contained treatment and additional numerical examples are given.

2 Decay rates for the entries of band matrix functions.

In several applications it is important to know whether the entries of a matrix function $f(A)$ exhibit some kind of decay behavior away from the main diagonal, and to be able to estimate the rate of decay. For instance, in preconditioning [6, 12, 30] and in the study of the convergence of spline approximations [17] it is useful to obtain information on the decay behavior for the entries of the inverse. A well-known result due to Demko, Moss and Smith [18] states that if A is a banded SPD matrix, the entries of A^{-1} are bounded in an exponentially decaying manner along each row or column. The rate of decay is governed by the bandwidth and by the extreme eigenvalues of A : decay is fast if the matrix

is well-conditioned and has narrow bandwidth, otherwise it can be rather slow. The proof in [18] is based on a result of Chebyshev on the best polynomial approximation of $f(\lambda) = \lambda^{-1}$. See [35] for an excellent survey.

In this section we prove that a similar result holds for a wide class of (analytic) matrix functions. This generalization is obtained by applying a fundamental result in classical approximation theory, due to S. N. Bernstein, on the best polynomial approximation of analytic functions. Our method of proof closely follows the one by Demko *et al.*, with Bernstein's Theorem playing the role of Chebyshev's result. We begin by recalling Bernstein's Theorem. Our discussion is based on [34].

Let P_k be the set of all polynomials with real coefficients and degree less than or equal to k . For a continuous function F on $[-1, 1]$, the *best approximation error* is defined as

$$E_k(F) = \inf\{\|F - p\|_\infty : p \in P_k\}$$

where

$$\|F - p\|_\infty = \max_{-1 \leq x \leq 1} |F(x) - p(x)|.$$

Bernstein [7] investigated the asymptotic behavior of the quantity $E_k(F)$ for a function F analytic on a domain which contains the interval $[-1, 1]$. His result states that this error decays to zero exponentially as $k \rightarrow \infty$, and shows how to estimate the decay rate.

If F is analytic on a simply connected open region of the complex plane containing the interval $[-1, 1]$, there exist ellipses with foci in -1 and 1 such that F is analytic in their interiors. Let $\alpha > 1$ and $\beta > 0$ be the half axes of such an ellipse, $\alpha > \beta$; from the identity $\sqrt{\alpha^2 - \beta^2} = 1$ we find that

$$\alpha - \beta = \frac{1}{\alpha + \beta}$$

and the ellipse is completely specified once the number $\chi = \alpha + \beta$ is known, hence we may denote it by \mathcal{E}_χ . Furthermore, note that β is specified once α is, because $\beta = \sqrt{\alpha^2 - 1}$.

Bernstein's Theorem can be formulated as follows (see [34, p. 91] for a proof).

THEOREM 2.1. *Let the function F be analytic in the interior of the ellipse \mathcal{E}_χ , $\chi > 1$, and continuous on \mathcal{E}_χ . In addition, suppose $F(z)$ is real for real z . Then*

$$(2.1) \quad E_k(F) \leq \frac{2M(\chi)}{\chi^k(\chi - 1)}$$

where

$$M(\chi) = \max_{z \in \mathcal{E}_\chi} |F(z)|.$$

It is convenient to introduce now the concept of *regularity ellipse* of F , as in [34]. It is the ellipse $\mathcal{E}_{\bar{\chi}}$ where

$$\bar{\chi} = \bar{\chi}(F) = \sup\{\chi : F \text{ is analytic in the interior of } \mathcal{E}_\chi\}.$$

Evidently, it is important to study the behavior of the right-hand side of (2.1) as χ varies between 1 and $\bar{\chi}$. In particular, we see that the decay rate may become arbitrarily slow as $\chi \rightarrow 1$ (from the right). On the other hand, as χ increases, so does the rate of decay, as long as the quantity $M(\chi)$ remains bounded.

Now we translate Bernstein's Theorem in terms of matrices. Our treatment closely follows [18]. Let m be a nonnegative, even integer. A symmetric matrix $B = (b_{ij})$ is called m -banded if

$$b_{ij} = 0 \quad \text{when } |i - j| > \frac{m}{2}.$$

For example, a tridiagonal matrix is 2-banded.

Letting

$$K_0 = \frac{2\chi M(\chi)}{\chi - 1}, \quad q = \frac{1}{\chi} < 1,$$

we can rewrite the error bound (4.1) as

$$(2.2) \quad E_k(F) \leq K_0 q^{k+1},$$

and we have the following exponentially decaying bound for the entries of $F(B)$.

THEOREM 2.2. *Let B be symmetric, m -banded, and such that $[-1, 1]$ is the smallest interval containing $\sigma(B)$, the spectrum of B . Let $\rho = q^{\frac{2}{m}}$ and*

$$K = \max\{K_0, \|F(B)\|_2\}$$

with F and K_0 as above. Then we have

$$(2.3) \quad |(F(B))_{ij}| \leq K \rho^{|i-j|}.$$

PROOF. Observe first that B^k is km -banded for $k = 0, 1, \dots$ so that $p(B)$ is km -banded for all polynomials $p \in P_k$. From Bernstein's Theorem we know that there exists a sequence of polynomials p_k of degree k which satisfies

$$\|F - p_k\|_\infty = E_k(F) \leq K_0 q^{k+1}.$$

We have

$$\|F(B) - p_k(B)\|_2 = \max_{x \in \sigma(B)} |F(x) - p_k(x)| \leq \|F - p_k\|_\infty \leq K_0 q^{k+1}.$$

For $i \neq j$ write

$$|i - j| = \frac{km}{2} + l \quad \text{for } l = 1, \dots, \frac{m}{2};$$

then we have that

$$\frac{2|i - j|}{m} \leq k + 1$$

and hence, observing that $(p_k(B))_{ij} = 0$ for $|i - j| > mk/2$,

$$|(F(B))_{ij}| = |(F(B))_{ij} - (p_k(B))_{ij}| \leq \|F(B) - p_k(B)\|_2 \leq K_0 \rho^{|i-j|}.$$

Finally, if $i = j$ then $|(F(B))_{ii}| \leq \|F(B)\|_2$ and therefore (2.3) holds for all i, j . \square

Let A be a symmetric matrix and let $a = \lambda_{\min}(A)$ and $b = \lambda_{\max}(A)$, so that $[a, b]$ is the smallest interval containing $\sigma(A)$. If we introduce the linear affine function

$$\psi : \mathbb{C} \rightarrow \mathbb{C}, \quad \psi(\lambda) = \frac{2\lambda - (a + b)}{b - a}$$

then $\psi([a, b]) = [-1, 1]$, so that the spectrum of the symmetric matrix

$$B = \psi(A) = \frac{2}{b-a}A - \frac{a+b}{b-a}I$$

is contained in $[-1, 1]$. Furthermore, given a function f analytic on a simply connected region containing $[a, b]$ and such that $f(\lambda)$ is real when λ is real, then the function $F = f \circ \psi^{-1}$ satisfies the assumptions of Bernstein's Theorem. Here

$$\psi^{-1} : \mathbb{C} \rightarrow \mathbb{C}, \quad \psi^{-1}(x) = \frac{(b - a)x + a + b}{2}$$

is the inverse function of ψ . It is clear that the decay bound (2.3) for the entries of $F(B)$ leads to a similar bound for the entries of $f(A)$.

For instance, in the special case where A is SPD and $f(\lambda) = \lambda^{-\frac{1}{2}}$, we apply Bernstein's result to the function

$$F(z) = \frac{1}{\sqrt{\frac{(b-a)}{2}z + \frac{a+b}{2}}}$$

The regularity ellipse for this F is $\mathcal{E}_{\bar{\chi}}$ where

$$\bar{\chi} = \frac{b + a}{b - a} + \sqrt{\left(\frac{b + a}{b - a}\right)^2 - 1}$$

For $1 < \chi < \bar{\chi}$, the function F is analytic inside \mathcal{E}_{χ} and continuous on \mathcal{E}_{χ} . If we let $\kappa = \frac{b}{a}$ (the spectral condition number of A) we find that

$$\bar{\chi} = \frac{(\sqrt{\kappa} + 1)^2}{\kappa - 1}$$

From $1 < \chi < \bar{\chi}$ and recalling that $q = \frac{1}{\chi}$ we get

$$(2.4) \quad \frac{\kappa - 1}{(\sqrt{\kappa} + 1)^2} = \frac{1}{\bar{\chi}} < q < 1.$$

In this special case the constant $M(\chi)$ is easily determined. The function $|F|$ attains its maximum on the ellipse \mathcal{E}_{χ} , where $\chi = \alpha + \beta$, at the point $z = -\alpha$ on the real axis, so that

$$(2.5) \quad M(\chi) = \max_{z \in \mathcal{E}_{\chi}} |F(z)| = |F(-\alpha)| = \frac{\sqrt{2}}{\sqrt{(a-b)\alpha + a + b}} = \frac{\sqrt{2}}{\sqrt{\frac{(a-b)(\chi^2 + 1)}{2\chi} + a + b}}$$

This constant can be very large for a near zero, as is the case if A is nearly singular. Clearly, both α and χ approach 1 when $a \rightarrow 0^+$ and therefore the denominator in (2.5) tends to zero, independent of the value of b .

From (2.1)–(2.5) one can see that the decay rate of the bound on the entries of $A^{-\frac{1}{2}}$ is faster for well-conditioned A ($\kappa \approx 1$), as for $\kappa \rightarrow 1^+$ we have $\frac{1}{\chi} \rightarrow 0$ and thus we can take a very small value for q in (2.2); furthermore, if a remains bounded away from zero, we see that K_0 remains bounded as $\kappa \rightarrow 1^+$, and decay is fast away from the main diagonal. Conversely, the same formulas show that decay can be arbitrarily slow as the condition number of A increases to infinity, as in this case $q \rightarrow 1^-$ and K_0 grows without bound. It is also clear from the definition of ρ in Theorem 2.2 that decay is faster when the bandwidth of A is narrow (small m).

The following two examples illustrate the decay in smooth functions of band matrices. See [5] for additional examples.

EXAMPLE 2.1. Let $T_4 := \text{tridiag}(-1, 4, -1)$. For $n = 10$ the upper triangular part of $T_4^{-\frac{1}{2}}$ rounded to four places is

$$\begin{pmatrix} 0.5129 & 0.0681 & 0.0136 & 0.0030 & 0.0007 & 0.0001 & \cdots \\ & 0.5266 & 0.0711 & 0.0143 & 0.0032 & 0.0008 & \cdots \\ & & 0.5273 & 0.0713 & 0.0144 & 0.0032 & \cdots \\ & & & 0.5273 & 0.0713 & 0.0144 & \cdots \\ & & & & 0.5273 & 0.0713 & \cdots \\ & & & & & 0.5273 & \cdots \\ & & & & & & \ddots \\ & & & & & & \ddots \end{pmatrix}$$

showing the rapid decay away from the main diagonal, as predicted by Theorem 2.2. Notice that $T_4^{-\frac{1}{2}}$ is nonnegative [39, 20]. \square

EXAMPLE 2.2. Let $T_2 := \text{tridiag}(-1, 2, -1)$. Consider $f(T_2) = \cos T_2^{\frac{1}{2}}$, a matrix function which arises in the numerical solution of partial differential equations of hyperbolic type [19]. For $n = 10$, the upper triangular part is, rounded to four digits:

$$\begin{pmatrix} 0.1899 & 0.3516 & 0.0340 & 0.0012 & 0.0000 & 0.0000 & \cdots \\ & 0.2239 & 0.3528 & 0.0340 & 0.0012 & 0.0000 & \cdots \\ & & 0.2239 & 0.3528 & 0.0340 & 0.0012 & \cdots \\ & & & 0.2239 & 0.3528 & 0.0340 & \cdots \\ & & & & 0.2239 & 0.3528 & \cdots \\ & & & & & 0.2239 & \cdots \\ & & & & & & \ddots \\ & & & & & & \ddots \end{pmatrix}.$$

Note the very fast (although nonmonotonic) decay away from the main diagonal. This example shows that rapid decay of the entries of $f(A)$ is possible even if A is only weakly diagonally dominant. \square

where u and v are given vectors and f is a smooth (possibly C^∞) function on an interval containing the spectrum of the symmetric matrix A , was considered. In order to make the present paper self-contained, we summarize the approach in [24] below.

Let $A = Q\Lambda Q^T$ be the eigendecomposition of A , where Q is an orthogonal $n \times n$ matrix and Λ is a diagonal matrix consisting of the eigenvalues λ_i of A which we order as

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n.$$

Let $a = \lambda_1$ and $b = \lambda_n$. For a smooth function f defined on a set containing the closed, bounded interval $a \leq x \leq b$, define a matrix $f(A)$ by letting

$$f(A) = Qf(\Lambda)Q^T$$

where

$$f(\Lambda) = \begin{pmatrix} f(\lambda_1) & 0 & \cdots & 0 \\ 0 & f(\lambda_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & f(\lambda_n) \end{pmatrix}.$$

For each $u, v \in \mathbb{R}^n$ we have

$$u^T f(A)v = u^T Qf(\Lambda)Q^T v = \alpha^T f(\Lambda)\beta = \sum_{i=1}^n f(\lambda_i)\alpha_i\beta_i$$

where $\alpha = Q^T u$ and $\beta = Q^T v$. This sum can be interpreted as a Riemann–Stieltjes integral

$$(3.1) \quad u^T f(A)v = I[f] = \int_a^b f(\lambda)d\mu(\lambda),$$

where μ is the piecewise constant function

$$\mu(\lambda) = \begin{cases} 0 & \text{if } \lambda < a = \lambda_1, \\ \sum_{j=1}^i \alpha_j\beta_j & \text{if } \lambda_i \leq \lambda < \lambda_{i+1}, \\ \sum_{j=1}^n \alpha_j\beta_j & \text{if } b = \lambda_n \leq \lambda. \end{cases}$$

Notice that for $u = v$ the function $\mu(\lambda)$ is nonnegative and nondecreasing: it is a staircase function with steps of size α_i^2 at the eigenvalues λ_i .

Setting $u = e_i, v = e_j$ in (3.1) we obtain an integral formula for the (i, j) entry of the matrix $f(A)$. Of special interest for applications are the cases $f(A) = A^{-1}$ and $f(A) = A^{-2}$, which have already been discussed in several papers (see [16, 24, 25, 26, 28]). However, applications to preconditioning were not investigated in those papers.

Formula (3.1) suggests that we use Gauss-type quadrature rules to obtain estimates for the quadratic form $u^T f(A)u$. The general Gauss-type quadrature

rule for the Riemann–Stieltjes integral is

$$(3.2) \quad \int_a^b f(\lambda)d\mu(\lambda) = \sum_{j=1}^N w_j f(t_j) + \sum_{k=1}^M v_k f(z_k) + R[f]$$

where the weights $\{w_j\}_{j=1}^N$, $\{v_k\}_{k=1}^M$ and the nodes $\{t_j\}_{j=1}^N$ are unknowns and the nodes $\{z_k\}_{k=1}^M$ are prescribed. For $M = 0$, formula (3.2) reduces to the Gauss rule with no prescribed nodes. For $M = 1$ and $z_1 = a$ or $z_1 = b$ we have the Gauss–Radau formula. If $M = 2$ and $z_1 = a, z_2 = b$ we obtain the Gauss–Lobatto formula. In (3.2), $R[f]$ denotes the error, for which a general formula is known and can be found in [14, 24].

In particular, the Gauss quadrature rule can be written as

$$I[f] = \int_a^b f(\lambda)d\mu(\lambda) = \sum_{j=1}^N w_j^G f(t_j^G) + R_G[f].$$

Recall that the nodes and weights in (3.2) are closely related to the sequence of polynomials $\{p_k(\lambda)\}_{k=0}^\infty$ that are orthonormal with respect to μ :

$$\int_a^b p_i(\lambda)p_j(\lambda)d\mu(\lambda) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

These polynomials satisfy a three-term recurrence relation

$$\begin{aligned} \gamma_j p_j(\lambda) &= (\lambda - \omega_j)p_{j-1}(\lambda) - \gamma_{j-1}p_{j-2}(\lambda), & j \geq 2, \\ p_{-1}(\lambda) &\equiv 0, & p_0(\lambda) \equiv 1. \end{aligned}$$

Each p_k is of exact degree k , its roots are distinct, real and lie in the interval $[a, b]$. The recurrence relation above can be rewritten in matrix form as

$$\lambda p(\lambda) = J_N p(\lambda) + \gamma_N p_N(\lambda) e_N,$$

where

$$\begin{aligned} p(\lambda)^T &= [p_0(\lambda) \quad p_1(\lambda) \quad \cdots \quad p_{N-1}(\lambda)], \\ e_N^T &= (0 \quad 0 \quad \cdots \quad 0 \quad 1), \\ J_N &= \begin{pmatrix} \omega_1 & \gamma_1 & & & \\ \gamma_1 & \omega_2 & \gamma_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \gamma_{N-2} & \omega_{N-1} & \gamma_{N-1} \\ & & & \gamma_{N-1} & \omega_N \end{pmatrix}. \end{aligned}$$

The eigenvalues of the symmetric tridiagonal matrix J_N , which are the zeroes of p_N , are the nodes of the Gauss quadrature rule. As shown in [29], the weights

are the squares of the first component of the normalized eigenvectors of J_N . The coefficients ω_i, γ_i can be obtained by the Lanczos process [24, 29]. It is important to emphasize that in some cases there is actually no need to compute the eigenvalues and eigenvectors of the tridiagonal matrix. Consider for example the Gauss rule; then it can be easily shown [23, 24] that

$$(3.3) \quad \sum_{j=1}^N w_j^G f(t_j^G) = e_1^T f(J_N) e_1.$$

It follows that there is no need to compute the eigenvalues and eigenvectors of J_N , provided that the (1,1) entry in $f(J_N)$ is easily computable. Similar statements hold for the Gauss–Radau and Gauss–Lobatto rules.

Thus, the evaluation of Gauss-type quadrature rules is reduced to the computation of orthonormal polynomials via a three-term recurrence or, equivalently, to finding a tridiagonal matrix and certain spectral information about it. A natural way to do this is to use the Lanczos process.

Bounds or estimates for the bilinear form $u^T f(A)v$ with $v \neq u$ can be obtained either by the unsymmetric Lanczos process for the evaluation of Gauss-type quadrature rules with respect to signed measures [24], or else using the *polarization identity*

$$u^T f(A)v = \frac{1}{4} [p^T f(A)p - q^T f(A)q], \quad p = u + v, \quad q = u - v,$$

which can be used to express a bilinear form in terms of the corresponding quadratic form.

Suppose now that we want to estimate the (i, i) entry of matrix $f(A)$, where f is a smooth function such that $f^{(2j)}(x) > 0$ for all $j \geq 0$ and for all $x \in (a, b)$. Then the Gauss rule gives a lower bound [24], which can be computed with the Lanczos algorithm with $x_0 = e_i$. Taking Nit steps in the Lanczos algorithm corresponds to computing the Gauss rule with $N = Nit$ nodes. As Nit increases, the bound approaches the exact value of $e_i^T f(A)e_i$. If f is a polynomial of degree $2N - 1$ or less, at most $Nit = N$ steps are required to find the exact value (in exact arithmetic). We recall here that any matrix function $f(A)$ can be expressed as a polynomial in A of degree $\leq n - 1$. In exact arithmetic, the number of Lanczos steps actually needed to compute $u^T f(A)u$ is determined by the degree of the minimal polynomial of A . For symmetric A , this equals the number of distinct eigenvalues.

As shown in [24], similar results hold for the Gauss–Radau and Gauss–Lobatto rules. The Gauss–Radau rules give lower and upper bounds which can be computed by means of the Lanczos process; the bounds become increasingly tighter as the iteration proceeds. The same is true for the upper bound from the Gauss–Lobatto rule. Notice that N Lanczos steps correspond to N nodes for the Gauss–Radau rule and to $N - 1$ nodes for the Gauss–Lobatto rule.

Explicit bounds for the entries of A^{-1} can be obtained by taking a single Lanczos step. Such bounds were obtained in [24] and are recalled below.

THEOREM 3.1. Let $s_i^2 := \sum_{j \neq i} a_{ji}^2$. Then

$$(3.4) \quad \frac{\sum_{k \neq i} \sum_{l \neq i} a_{ki} a_{kl} a_{li}}{a_{ii} \sum_{k \neq i} \sum_{l \neq i} a_{ki} a_{kl} a_{li} - (\sum_{k \neq i} a_{ki}^2)^2} \leq (A^{-1})_{ii},$$

$$(3.5) \quad \frac{a_{ii} - b + s_i^2/b}{a_{ii}^2 - a_{ii}b + s_i^2} \leq (A^{-1})_{ii} \leq \frac{a_{ii} - a + s_i^2/a}{a_{ii}^2 - a_{ii}a + s_i^2},$$

$$(3.6) \quad (A^{-1})_{ii} \leq \frac{a + b - a_{ii}}{ab}.$$

Notice that the lower bound (3.4) from the Gauss rule does not depend on a and b . Also, notice that for a matrix A with a clustered spectrum the lower and upper bounds (3.5) from the Gauss–Radau rules are tight. Conversely, the bounds may be inaccurate if a and b are far apart from each other. However, if A has only two distinct eigenvalues a and b , all the bounds in Theorem 3.1 give the exact value of $(A^{-1})_{ii}$. Indeed, since in this case the minimal polynomial of A has second degree, A^{-1} can be expressed as a first degree polynomial in A and the Gaussian quadrature rules with $N = 1$ are exact.

Estimates for the off-diagonal entries of A^{-1} can be found using the unsymmetric Lanczos method with the Gauss–Radau rules. The results are summarized in the following theorem from [24].

THEOREM 3.2. Let

$$t_{ij} := \sum_{k \neq i} a_{ki}(a_{ki} + a_{kj}) - a_{ij}(a_{ij} + a_{ii}).$$

For $(A^{-1})_{ij} + (A^{-1})_{ii}$ we have the two following estimates:

$$\frac{a_{ii} + a_{ij} - a + t_{ij}/a}{(a_{ii} + a_{ij})^2 - a(a_{ii} + a_{ij}) + t_{ij}}, \quad \frac{a_{ii} + a_{ij} - b + t_{ij}/b}{(a_{ii} + a_{ij})^2 - b(a_{ii} + a_{ij}) + t_{ij}}.$$

If $t_{ij} \geq 0$, the first expression gives an upper bound and the second one a lower bound.

Subtracting the bounds for the diagonal entries from the estimates in the previous theorem we obtain estimates for the off-diagonal entries. Notice that the same bounds as in Theorems 3.1–3.2 have been obtained with different methods in [36]. We also mention here that bounds for the entries of the inverse of a nonsymmetric matrix A can be obtained by computing bounds for bilinear expressions of the form $e_i^T (A^T A)^{-1} v$ where $v = A^T e_j$; see [3].

We now turn to approximating the entries of $f(A)$, where f denotes any sufficiently smooth function on an interval containing the eigenvalues of A . These estimates actually provide lower and upper bounds if f is strictly completely monotonic on an open interval containing the spectrum of A . Recall that f is

strictly completely monotonic on an interval I if $f^{(2j)}(x) > 0$ for all $x \in I$ and all $j \geq 0$, and $f^{(2j+1)}(x) < 0$ for all $x \in I$ and all $j \geq 0$; see [20, 39]. The generalization is straightforward. As before, the symmetric Lanczos process can be used to compute bounds for the diagonal entries $f_{ii} := (f(A))_{ii}$, whereas the unsymmetric Lanczos process gives estimates for $f_{ij} := (f(A))_{ij}$ with $i \neq j$ in the form of bounds for $f_{ij} + f_{ii}$.

We omit the details of the computations, which run as in the case of A^{-1} . The only difference is that it is now required to compute the (1,1) entry of a function of a 2×2 matrix. This entry can be computed numerically without difficulty. Here we derive explicit expressions (not found in [24]) with the help of the Lagrange interpolation formula for the evaluation of matrix functions (see [21], Chapter 5): if $G = (g_{ij})$ is a 2×2 matrix with distinct eigenvalues μ_1, μ_2 then

$$f(G) = \frac{f(\mu_1)}{\mu_1 - \mu_2}(G - \mu_2 I_2) + \frac{f(\mu_2)}{\mu_2 - \mu_1}(G - \mu_1 I_2).$$

It follows that

$$(3.7) \quad (f(G))_{11} = \frac{g_{11}[f(\mu_1) - f(\mu_2)] + \mu_1 f(\mu_2) - \mu_2 f(\mu_1)}{\mu_1 - \mu_2}.$$

Assume now that f is strictly completely monotonic. For the Gauss rule, formulas (3.3) and (3.7) applied to the matrix

$$G = J_2 = \begin{pmatrix} \omega_1 & \gamma_1 \\ \gamma_1 & \omega_2 \end{pmatrix}$$

yield the lower bound

$$(3.8) \quad \frac{a_{11}[f(\mu_1) - f(\mu_2)] + \mu_1 f(\mu_2) - \mu_2 f(\mu_1)}{\mu_1 - \mu_2} \leq f_{ii}$$

where

$$\mu_1 = \frac{1}{2}[\omega_1 + \omega_2 - \delta], \quad \mu_2 = \frac{1}{2}[\omega_1 + \omega_2 + \delta], \quad \delta = \sqrt{(\omega_1 - \omega_2)^2 + 4\gamma_1^2}$$

are the eigenvalues of J_2 . Here

$$\omega_1 = a_{ii}, \quad \gamma_1^2 = \sum_{j \neq i} a_{ji}^2, \quad \omega_2 = \frac{1}{\gamma_1^2} \sum_{k \neq i} \sum_{l \neq i} a_{ki} a_{kl} a_{li}.$$

Again, this lower bound does not depend on a and b .

For the Gauss–Radau and Gauss–Lobatto rules the bounds are given by similar expressions. To simplify the formulas we introduce the function

$$\phi(\mu_1, \mu_2) = \frac{\omega_1[f(\mu_1) - f(\mu_2)] + \mu_1 f(\mu_2) - \mu_2 f(\mu_1)}{\mu_1 - \mu_2}.$$

Then we can reformulate (3.8) as

$$\phi(\mu_1, \mu_2) \leq f_{ii}$$

where μ_1 and μ_2 are the eigenvalues of J_2 as expressed above. The bounds from

If the approximate inverses satisfy (4.1), it is not difficult to see that all the Δ_i will be strictly diagonally dominant. It was shown in [12] that all the pivot blocks Σ_i in the block Cholesky factorization of A are strictly diagonally dominant. Now, recall that Theorem 3.2 in [2] insures that both $\Sigma_2 = D_2 - A_2 D_1^{-1} A_2^T$ and its tridiagonal approximation $\Delta_2 = D_2 - A_2 \Lambda_1 A_2^T$ are M-matrices. Observing that the last term in the right-hand side of the identity

$$D_2 - A_2 \Lambda_1 A_2^T = D_2 - A_2 D_1^{-1} A_2^T + A_2 (D_1^{-1} - \Lambda_1) A_2^T$$

is a nonnegative matrix, it follows that Δ_2 is at least as strictly diagonally dominant as Σ_2 , since the diagonal entries cannot be smaller and the off-diagonal entries cannot increase in absolute value. A simple inductive argument can be used to show that all the Δ_i remain strictly diagonally dominant, and they are at least as strictly diagonally dominant as the Σ_i . For Poisson's equation with Dirichlet boundary conditions discretized with a five-point finite difference scheme on a uniform mesh, all the Δ_i exhibit a good degree of diagonal dominance and the entries in Δ_i^{-1} decay rapidly away from the main diagonal. As we shall see, the bounds in Section 3 can be used to compute accurate and inexpensive tridiagonal approximations to the inverse of strictly diagonally dominant matrices.

As a simple illustration we consider $T_4 = \text{tridiag}(-1, 4, -1)$, as in [12]. The diagonal blocks of the five-point finite difference matrix discretization of the Laplace operator on a rectangular region with Dirichlet boundary conditions (the so-called model problem) are all equal to T_4 . Note that the spectral condition number of T_4 is ≤ 3 (independent of n). An approximate inverse of T_4 is needed in the initial step of an incomplete Cholesky factorization for the two-dimensional model problem. The approximation Λ_1 to $\Delta_1^{-1} = D_1^{-1} = T_4^{-1}$ obtained from the Gauss rules satisfies the important property (4.1), with $p = 2$. As already noted, this in turn guarantees that $\Delta_2 = D_2 - A_2 \Lambda_1 A_2^T$ is a strictly diagonally dominant M-matrix.

Thus, we are interested to see whether the Gauss rules give a good approximation to T_4^{-1} . For $n = 10$ the upper triangular part of T_4^{-1} rounded to four places is

$$(4.2) \quad \begin{pmatrix} 0.2679 & 0.0718 & 0.0192 & 0.0052 & 0.0014 & 0.0004 & \cdots \\ & 0.2872 & 0.0770 & 0.0206 & 0.0055 & 0.0015 & \cdots \\ & & 0.2886 & 0.0773 & 0.0207 & 0.0056 & \cdots \\ & & & 0.2887 & 0.0773 & 0.0207 & \cdots \\ & & & & 0.2887 & 0.0773 & \cdots \\ & & & & & 0.2887 & \cdots \\ & & & & & & \ddots \\ & & & & & & \ddots \end{pmatrix}$$

showing the rapid decay away from the main diagonal. Rounded to four places, the eigenvalues of T_4 are

$$\{2.0810, 2.3175, 2.6903, 3.1692, 3.7154, 4.2846, 4.8308, 5.3097, 5.6825, 5.9190\}$$

imations does not significantly deteriorate with increasing problem size. This is due to the fact that $\kappa(T_4)$ remains bounded independent of n .

Approximations to T_4^{-1} were computed also using the Gauss–Radau and Gauss–Lobatto bounds. Because these rules use more information (in the form of estimates for the extreme eigenvalues), one could expect better results than those obtained using the Gauss rules. However, this is not always true. The banded approximate inverses based on the Gauss–Radau and Gauss–Lobatto bounds were often found to be slightly less accurate than those obtained from the Gauss bounds. Furthermore, condition (4.1), which is very important in the context of block incomplete factorizations, was not always fulfilled by these approximate inverses.

4.2 A banded inverse preconditioner for Toeplitz matrices.

Here we propose using certain banded approximate inverses as preconditioners for dense Toeplitz systems. Following Strang [38], we observe that in many practical applications the main diagonal of a Toeplitz matrix $A = (a_{|i-j|})$ and its immediate neighbors are strongly dominant. The idea is then to consider, for some k , the k -banded approximation $A^{(k)}$ to A defined by

$$(A^{(k)})_{ij} = \begin{cases} a_{|i-j|} & \text{if } |i-j| \leq k/2, \\ 0 & \text{otherwise.} \end{cases}$$

Next, a banded approximate inverse of $A^{(k)}$ is computed using any of the methods described in this paper. If A is not too ill-conditioned, the explicit formulas corresponding to a single Lanczos step should yield a reasonable approximation. For a fixed bandwidth, the cost of this construction is $\mathcal{O}(1)$. The resulting banded approximate inverse is used as a preconditioner for the conjugate gradient method (PCG) applied to $Ax = b$. The cost of using this preconditioner at each PCG iteration is only $\mathcal{O}(n)$, as compared to $\mathcal{O}(n \log n)$ for the widely used circulant preconditioners. Note that the set-up costs for circulant preconditioners are also $\mathcal{O}(n \log n)$ if the circulant is inverted before iterating, as is customary. Also note that the banded inverse preconditioner is trivially vectorized and parallelized. An alternative would be to use $A^{(k)}$ itself as preconditioner, but its application would require a band matrix solve at each step of PCG rather than a matrix–vector multiply, an operation that is not easily parallelized.

We illustrate this procedure on three simple examples from [38] and [9]. In these examples the symmetric Toeplitz matrix $A = (a_{|i-j|})$ is defined by $a_k = 1/(k+1)^p$ where $0 \leq k \leq n-1$ and p takes the values 2, 1 and 1/2, respectively. For each problem we consider the tridiagonal approximation $A^{(2)}$ to A formed with the three central diagonals of A , and construct a tridiagonal approximation to the inverse of $A^{(2)}$ using the Gauss rules. The resulting approximate inverse is used as a preconditioner for the conjugate gradient method applied to the original Toeplitz system $Ax = b$. For the sake of comparison, we also present results obtained with T. Chan’s optimal circulant preconditioner [10]. The results for different values of n are illustrated in Tables 4.1–4.3 in terms of PCG iterations (above) and number of floating point operations (below, in millions). The latter

includes the work required to form the preconditioner. The iteration was stopped when the Euclidean norm of the initial residual had been reduced by at least seven orders of magnitude. The initial guess used was $x_0 = 0$, and the right-hand side was a random vector with entries uniformly distributed over $(0, 1)$.

Table 4.1: Number of iterations and work for the case $p = 2$.

Precond.	Problem size n					
	32	64	128	256	512	1024
None	12	13	14	14	14	15
	.03	.12	.30	.67	1.5	3.5
Gauss	7	8	8	8	8	8
	.02	.08	.17	.39	.86	1.9
Optimal	8	8	8	8	8	8
	.05	.11	.25	.57	1.3	2.8

Table 4.2: Number of iterations and work for the case $p = 1$.

Precond.	Problem size n					
	32	64	128	256	512	1024
None	21	27	35	37	43	49
	.05	.25	.75	1.8	4.6	11.4
Gauss	9	11	13	15	17	18
	.02	.11	.28	.73	1.8	4.2
Optimal	9	8	9	9	10	10
	.05	.11	.28	.64	1.6	3.5

Table 4.3: Number of iterations and work for the case $p = 1/2$.

Precond.	Problem size n					
	32	64	128	256	512	1024
None	30	37	49	67	82	101
	.07	.35	1.0	3.2	8.7	23.5
Gauss	10	14	17	20	25	29
	.02	.13	.37	.97	2.7	6.8
Optimal	9	9	9	10	10	11
	.05	.12	.28	.70	1.6	3.8

It can be seen from these results that the Gauss rule preconditioner outperforms the optimal one for all values of n on the first of the three problems, for which the decay away from the main diagonal is fastest. On the second problem the optimal preconditioner is better for $n \geq 256$, and on the third one already for $n \geq 64$. Given the fact that the Gauss rule preconditioner, by construction, uses information contained in the three central diagonals only, it is not surprising that it works really well only for problems with rapid decay. On the other hand the optimal preconditioner is computed using information from the

entire matrix, and works almost equally well on all three problems. For the Gauss rule preconditioner, the number of iterations can be further reduced by increasing the bandwidth of the preconditioner. For the first of the three problems, for example, an approximate inverse with five nonzero diagonals results in six PCG iterations, independent of problem size. Adding more diagonals or using more than three diagonals of A to construct the preconditioner does not lead to a noticeable improvement, and taking more than one Lanczos step is not cost-effective.

5 Conclusions.

The entries of a matrix function $f(A)$, where A is a banded symmetric matrix and f is a smooth function, are bounded in an exponentially decaying manner away from the main diagonal. When the actual decay is rapid, banded approximations to $f(A)$ are justified and may be useful in applications. Inexpensive bounds and estimates can be obtained by means of Gauss quadrature rules combined with one step of the Lanczos process.

In this paper we showed that banded approximations to the inverse can be used for deriving preconditioners. The most natural application of the techniques considered in this paper is perhaps the construction of approximate inverses in the context of block incomplete factorizations. It was shown experimentally that for strongly diagonally dominant matrices, the approximate inverses obtained from the explicit bounds corresponding to one Lanczos step are very good and inexpensive approximations to the true inverse.

These same principles can be used to construct banded approximate inverse preconditioners for certain dense Toeplitz systems. If the entries in the coefficient matrix decay fast enough away from the main diagonal, this preconditioning strategy can be very effective. On the other hand, other methods should be preferred for problems which do not exhibit rapid decay.

A further application that has not been considered here is the derivation of preconditioners for solving linear systems of the form $f(A)x = b$ where f is a smooth function and A is a symmetric matrix. Iterative methods for this type of problems which only use information from the matrix A have been developed (e.g., in [40]), but little work has been done on preconditioning for such iterative methods. An exception is [8], for the special case of the matrix exponential. It would be interesting to investigate the use of Gauss quadrature rules in this context.

Acknowledgements.

The first author would like to thank Stanford University for the support and hospitality offered in November 1997, when part of this work was completed. We would like to thank the anonymous referees, whose comments made possible the deflation of the report [5] into this paper. Thanks also to Daniele Bertaccini for performing the runs with T. Chan's optimal preconditioner and to Jim Nagy for useful discussions.

REFERENCES

1. O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, Cambridge, 1994.
2. O. Axelsson and B. Polman, *On approximate factorization methods for block matrices suitable for vector and parallel processors*, Linear Algebra Appl., 77 (1986), pp. 3–26.
3. Z. Bai, M. Fahey, and G. H. Golub, *Some large-scale matrix computation problems*, J. Comput. Appl. Math., 74 (1996), pp. 71–89.
4. Z. Bai and G. H. Golub, *Bounds for the trace of the inverse and the determinant of symmetric positive definite matrices*, Ann. Numer. Math., 4 (1997), pp. 29–38.
5. M. Benzi and G. H. Golub, *Bounds for the entries of matrix functions with applications to preconditioning*, Stanford University Report SCCM-98-04, February 1998.
6. M. Benzi, C. D. Meyer, Jr., and M. Tũma, *A sparse approximate inverse preconditioner for the conjugate gradient method*, SIAM J. Sci. Comput., 17 (1996), pp. 1135–1149.
7. S. N. Bernstein, *Leçons sur les Propriétés Extrêmes et la Meilleure Approximation des Fonctions Analytiques d'une Variable Réelle*, Gauthier-Villars, Paris, 1926.
8. P. Castillo and Y. Saad, *Preconditioning the matrix exponential operator with applications*, University of Minnesota Supercomputing Institute Technical Report UMSI 97/142, Minneapolis, MN, 1997.
9. R. H. Chan and G. Strang, *Toeplitz equations by conjugate gradients with circulant preconditioner*, SIAM J. Sci. Stat. Comput., 10 (1989), pp. 104–119.
10. T. Chan, *An optimal circulant preconditioner for Toeplitz systems*, SIAM J. Sci. Stat. Comput., 9 (1998), pp. 766–771.
11. E. Chow and Y. Saad, *Approximate inverse preconditioners via sparse-sparse iterations*, SIAM J. Sci. Comput., 19 (1998), pp. 995–1023.
12. P. Concus, G. H. Golub, and G. Meurant, *Block preconditioning for the conjugate gradient method*, SIAM J. Sci. Stat. Comput., 6 (1985), pp. 220–252.
13. P. Concus and G. Meurant, *On computing INV block preconditionings for the conjugate gradient method*, BIT, 26 (1986), pp. 493–504.
14. P. J. Davis and P. Rabinowitz, *Methods of Numerical Integration*, 2nd ed., Academic Press, New York, 1984.
15. G. Dahlquist, S. C. Eisenstat and G. H. Golub, *Bounds for the error of linear systems using the theory of moments*, J. Math. Anal. Appl., 37 (1972), pp. 151–166.
16. G. Dahlquist, G. H. Golub, and S. G. Nash, *Bounds for the error in linear systems*, in Proceedings of the Workshop on Semi-Infinite Programming, R. Hettich, ed., Springer-Verlag, New York, 1978, pp. 154–172.
17. S. Demko, *Inverses of band matrices and local convergence of spline projections*, SIAM J. Numer. Anal., 14 (1977), pp. 616–619.
18. S. Demko, W. F. Moss, and P. W. Smith, *Decay rates for inverses of band matrices*, Math. Comp., 43 (1984), pp. 491–499.
19. V. L. Druskin and L. A. Knizhnerman, *Two polynomial methods of calculating functions of symmetric matrices*, USSR Comput. Maths. Math. Phys., 29 (1989), pp. 112–121.

20. M. Fiedler and H. Schneider, *Analytic functions of M -matrices and generalizations*, Linear and Multilinear Algebra, 13 (1983), pp. 185–201.
21. F. R. Gantmacher, *The Theory of Matrices*, Vol. I, Chelsea, New York, 1959.
22. G. H. Golub, *Some uses of the Lanczos algorithm in numerical linear algebra*, in Topics in Numerical Analysis, J. H. Miller, ed., Academic Press, New York, 1973, pp. 173–184.
23. G. H. Golub, *Bounds for matrix moments*, Rocky Mountain J. Math., 4 (1974), pp. 207–211.
24. G. H. Golub and G. Meurant, *Matrices, moments and quadratures*, in Numerical Analysis 1993, D. F. Griffiths and G. A. Watson, eds., Pitman Research Notes in Mathematics, Vol. 303, Essex, England, 1994, pp. 105–156.
25. G. H. Golub and G. Meurant, *Matrices, moments and quadratures II; how to compute the norm of the error in iterative methods*, BIT, 37 (1997), pp. 687–705.
26. G. H. Golub and Z. Strakoš, *Estimates in quadratic formulas*, Numer. Algorithms, 8 (1994), pp. 241–268.
27. G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, 1996.
28. G. H. Golub and U. von Matt, *Quadratically constrained least squares and quadratic problems*, Numer. Math., 59 (1991), pp. 561–580.
29. G. H. Golub and J. H. Welsh, *Calculation of Gauss quadrature rules*, Math. Comp., 23 (1969), pp. 221–230.
30. M. Grote and T. Huckle, *Parallel preconditioning with sparse approximate inverses*, SIAM J. Sci. Comput., 18 (1997), pp. 838–853.
31. L. A. Knizhnerman, *The simple Lanczos procedure: estimates of the error of the Gauss quadrature formula and their applications*, Comput. Math. Math. Phys., 36 (1996), pp. 1481–1492.
32. L. Yu. Kolotilina and A. Yu. Yeremin, *Factorized sparse approximate inverse preconditionings I. Theory*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 45–58.
33. L. Yu. Kolotilina and A. Yu. Yeremin, *On a family of two-level preconditionings of the incomplete block factorization type*, Sov. J. Numer. Anal. Math. Modelling, 1 (1986), pp. 293–320.
34. G. Meinardus, *Approximation of Functions: Theory and Numerical Methods*, Springer-Verlag, New York, 1967.
35. G. Meurant, *A review on the inverse of symmetric tridiagonal and block tridiagonal matrices*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 707–728.
36. P. D. Robinson and A. Wathen, *Variational bounds on the entries of the inverse of a matrix*, IMA J. Numer. Anal., 12 (1992), pp. 463–486.
37. Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS Publishing, Boston, 1996.
38. G. Strang, *A proposal for Toeplitz matrix calculations*, Stud. Appl. Math., 74 (1986), pp. 171–176.
39. R. S. Varga, *Nonnegatively posed problems and completely monotonic functions*, Linear Algebra Appl., 1 (1968), pp. 329–347.
40. H. A. van der Vorst, *An iterative solution method for solving $f(A)x = b$, using Krylov subspace information obtained for the symmetric positive definite matrix A* , J. Comput. Appl. Math., 18 (1987), pp. 249–263.