MATHEMATICS OF COMPUTATION, VOLUME 33, NUMBER 148
OCTOBER 1979, PAGES 1283–1288

On Bounding $\|A^{-1}\|_\infty$ for Banded A

By Stephen Demko

Abstract. Upper bounds for $\|A^{-1}\|_\infty$ in terms of inverses of certain submatrices are obtained for band matrices. An application to a problem in spline theory is made.

1. Introduction. Aside from arguments arising from Gerschgorin's theorem or from the Neumann series, there seems to be little known about how to bound the norm of the inverse of an $N \times N$ matrix. Varga [9] has obtained some extensions of these methods. In several papers, [2] and [3], for example, de Boor has used the fact that the inverse of a totally positive matrix is a "checkerboard" matrix to obtain bounds for specific matrices. In the case of Hermitian matrices, lower bounds on the moduli of the elements of the spectrum yield upper bounds on the spectral norm of the inverse. Generally speaking, however, if a matrix is not strictly diagonally dominant, it is not easy to bound its inverse (if it has one).

In this paper we show, essentially, that if A is a band matrix having nicely conditioned diagonal subblocks, then A itself is nicely conditioned. This result can be viewed as a generalization of the fact that a diagonal matrix is invertible if and only if each of its 1×1 diagonal submatrices is invertible. Although our result does not give as sharp bounds as the traditional methods, it can be used (with the aid of a computer) in cases where the diagonal dominance arguments do not apply. As an example, we shall bound the inverse of a specific seven-diagonal Toeplitz matrix of arbitrary size simply by knowing the inverse of a 9×9 subblock. We also apply our result to a problem in spline approximation and, thereby, extend some results of [6].

Unless stated otherwise all matrices will be $N \times N$ (i.e., finite). The usual coordinate vectors will be denoted by $e_i := (0, 0, \ldots, 0, 1, 0, \ldots, 0)^T$, so that $x \in \mathbf{C}^N$ has the expansion $x = \Sigma_i x_i e_i$. We use the usual inner product $\langle x, y\rangle := \Sigma x_i \bar{y}_i$ and the usual definitions for the vector norms $\|\cdot\|_p$ and the operator norms $\|\cdot\|_p$, $1 \leqslant p \leqslant \infty$, (cf. [7], for example). We use capital letters to denote matrices and lower case letters to denote the entries. If $Q = (q_{ij})$, then $Q^* = (\bar{q}_{ji})$ and $\langle Qx, y\rangle = \langle x, Q^*y\rangle$ for all $x, y \in \mathbf{C}^N$. If E is a nonempty subset of $\mathbf{N} := \{1, 2, \ldots, N\}$ and if A is an $N \times N$ matrix, then $A_E = (\tilde{a}_{i,j})$ will denote the $N \times N$ matrix with $\tilde{a}_{ij} = a_{ij}$, if i and j are in E and $\tilde{a}_{ij} = 0$ otherwise. I_E will be the diagonal matrix with diagonal entries $q_{ii} = 1$, if $i \in E$, $q_{ii} = 0$, if $i \notin E$. $A = (a_{ij})$ will be said to have *bandwidth* k, if $a_{ij} = 0$ for $|i - j| > k$.

Received October 12, 1977.
AMS (MOS) subject classifications (1970). Primary 15A09; Secondary 41A15.

2. Main Results. Our main result is

THEOREM 1. *Let $B = (b_{ij})$ be an $N \times N$ matrix with bandwidth k. Suppose that for each $i \in \mathbf{N}$ there exists an integer interval $F \subseteq \mathbf{N}$ containing i such that B_E is invertible in the sense that $(B_E)^{-1}B_E = I_E$, where $E := \{j \in \mathbf{N}: |j - f| \leqslant k \text{ for all } f \in F\}$. Let $v^{(i)}$ be the ith row of B_E^{-1} and assume that*

$$d_i := \sum_{n \in E} \left| \sum_{j \in E \setminus F} v_j^{(i)} b_{jn} \right| < 1.$$

Then, B is invertible and

$$\|B^{-1}\|_\infty \leqslant \max_i \sum_{j \in F} |v_j^{(i)}|/(1 - d_i).$$

Proof. Fix i, and let $x \in \mathbf{C}^N$. Note that for $j \in F$, $(Bx)_j = (B_E x)_j$; it is here that we use the bandedness. Now

$$x_i = (B_E^{-1}B_E x)_i = \sum_{j \in E} v_j^{(i)} (B_E x)_j = \sum_{j \in F} v_j^{(i)} (Bx)_j + \sum_{j \in E \setminus F} v_j^{(i)} \sum_{n \in E} b_{jn} x_n.$$

Therefore,

$$|x_i| \leqslant \|Bx\|_\infty \sum_{j \in F} |v_j^{(i)}| + \|x\|_\infty \sum_{n \in E} \left| \sum_{j \in E \setminus F} v_j^{(i)} b_{jn} \right|.$$

Consequently, since i was arbitrary and by the assumption on the d_i's

$$\|x\|_\infty \leqslant \|B\|_\infty \max_i \sum_{j \in F} |v_j^{(i)}|/(1 - d_i). \quad \text{Q.E.D.}$$

Remarks. 1. Merely the invertibility of the B_E's is not enough to assure the invertibility of B. For example, if B is the tridiagonal skew-symmetric matrix with general row of the form $-1\ 0\ 1$, then B is invertible if and only if its order (i.e., N) is even. So B can have lots of invertible subblocks, B_E, of arbitrarily large size and still not be invertible itself. Of course, the invertible subblocks have bad condition numbers.

2. The result of Theorem 1 becomes an equality, if B is a diagonal matrix for in this case one can take $k = 0$, and $F = \{i\}, i \in \mathbf{N}$. For tridiagonal matrices with general row of the form $1\ \alpha\ 1$ and $|1 - \alpha| > \sqrt{5}$, Theorem 1 furnishes the bound $\|B^{-1}\|_\infty \leqslant (|\alpha| - 2 - 4|\alpha^{-1}|)^{-1}$ with $F = \{i\}$ and N arbitrary as simple calculations show. This is not as sharp as the (sharp) bound $\|B^{-1}\|_\infty \leqslant (|\alpha| - 2)^{-1}$ for $|\alpha| > 2$ obtained by diagonal arguments; but it is asymptotically correct as $|\alpha| \longrightarrow \infty$.

3. It is possible to get results in other l_p norms, but they do not appear to be very useful. For example, using

$$|x_i|^p = \left| \sum_{j \in F} v_j^{(i)} (Bx)_j + \sum_{j \in E \setminus F} v_j^{(i)} \sum_{n \in E} b_{jn} x_n \right|^p,$$

summing on i and applying the standard inequalities, we get

$$\|x\|_p \leqslant \|Bx\|_p \left\{ \sum_i \|v^{(i)}\|_q^p \right\}^{1/p} + \|x\|_p \left\{ \sum_i \left(\sum_{j \in E \setminus F} |v_j^{(i)} b_{jn}|^q \right)^{p/q} \right\}^{1/p},$$

where $p^{-1} + q^{-1} = 1$. Hence, in order to get anything useful we must assume that $\Sigma_i(\Sigma_{j\in E\setminus F} |v_j^{(i)} b_{jn}|^q)^{p/q} < 1$. The condition of the theorem seems easier to verify. In addition, from [7, Corollary 2.4] we know that for banded matrices $\|B^{-1}\|_p$ can be bounded in terms of $\|B^{-1}\|_\infty$ independently of N.

4. We have stated our result for $N \times N$ matrices, but it is not hard to see that it also holds for infinite matrices acting on l_p spaces if we assume that we have an onto mapping.

Example. Let B denote the seven-diagonal Toeplitz matrix of arbitrary size satisfying the following for all i, j:

$$a_{ii} = 18,$$
$$a_{ij} = 10, \quad \text{if } |i-j| = 1,$$
$$a_{ij} = 4, \quad \text{if } |i-j| = 2,$$
$$a_{ij} = 1, \quad \text{if } |i-j| = 3.$$

In this case the bandwidth is $k = 3$. Let B_E be the 9×9 matrix of this form. We take F to be of the form $\{i-1, i, i+1\}$ so that for fixed i, E has the form $\{i-4, i-3, \ldots, i+4\}$. Now, the fifth row of B_E^{-1} is symmetric about the middle entry, and its first five entries are (after rounding)

$$.00060, .00131, .00815, -.05633, .11440.$$

Therefore, with the notation of the theorem

$$\sum_{j\in F} |v_j^{(i)}| = .1144 + 2(.05633) = .2271.$$

Similarly, one computes that

$$\sum_{n\in E} \left| \sum_{j\in E\setminus F} v_j^{(i)} b_{jn} \right| = .91836.$$

Consequently, by Theorem 1

$$\|B^{-1}\|_\infty \leqslant \frac{.2271}{1 - .9177} \leqslant 2.782.$$

All rounding was done in a way to maximize the final result. It should be noted that the matrix B is not strictly diagonally dominant if its order is greater than 2.

Using Theorem 1 to obtain bounds for inverses of non-Toeplitz matrices is clearly more difficult than the procedure outlined above. One result that might be useful is Theorem 2.2 of [7]. This result asserts that the entries of the inverse of a band matrix decay exponentially to zero, the rate depending on the condition number of the given matrix. So if we could bound the quantities $\|B_E^{-1}\|_\infty$ sufficiently well, we could apply Theorem 1. This result was sharpened in [4]; but even the rate of decay given in that paper is probably too conservative to be of much use, for example, it is always bigger than ½. The fact of the matter is that we really do not know much beyond diagonal dominance as far as our ability to say anything about the invertibility of matrices is concerned.

3. A Problem in Spline Theory. In this section we apply Theorem 1 to a problem in spline theory. Following [5], we let $\{t_i\}_{i=1}^{n+k}$ be a nondecreasing knot sequence with $t_i < t_{i+k}$ for all i. The L_p normalized B-splines of order k for this sequence are defined by

$$N_{i,p}(t) = k^{1/p}(t_{i+k} - t_i)^{1-1/p}[t_i, \ldots, t_{i+k}](\cdot - t)_+^{k-1}, \quad 1 \leqslant i \leqslant n,$$

where $[t_i, \ldots, t_{i+k}]\, f(\cdot)$ denotes the kth divided difference of f at the points t_i, $\ldots, t_{i+k}$ and

$$(s-t)_+^{k-1} = \begin{cases} (s-t)^{k-1} & \text{if } s \geqslant t, \\ 0 & \text{otherwise.} \end{cases}$$

We recall that there exist constants $D_{k,p}$ depending only on k and p such that for all sequences $\{a_i\} := a$

$$\|a\|_p D_{k,p}^{-1} \leqslant \left\| \sum a_i N_{i,p} \right\|_p \leqslant \|a\|_p, \tag{3.1}$$

where the function norm is the usual L_p norm: $\int_{t_1}^{t_{n+k}} |f(t)|^p\, dt = \|f\|_p^p$ (cf. [5]). We are interested in the following problem: let $Q\colon L_p \xrightarrow{\text{onto}} \operatorname{span}\{N_{i,p}\}_{i=1}^n$ be a bounded projection, find a bound for $\|Q\|_\infty$ in terms of $\|Q\|_p := \sup\{\|Qf\|_p\colon \|f\|_p = 1\}$. The results we present here extend those of [5], [8], and [6, Corollary 4.5].

We assume that Q satisfies the following: there exist functions $\phi_i \in L_q$, where $p^{-1} + q^{-1} = 1$ and $1 \leqslant p \leqslant \infty$, such that

(a) $\int \phi_i g = 0$ if $g(x) = 0$ a.e. in (t_i, t_{i+k}),

(b) $\|\phi_i\|_q \leqslant 1$, $1 \leqslant i \leqslant n$,

(c) $\int \phi_i(Qf - f) = 0$, $1 \leqslant i \leqslant n$, and

(d) there is $\gamma > 0$ such that for all $\{a_i\} = a$,

$$\|a\|_q \leqslant \gamma \left\| \sum a_i \phi_i \right\|_q.$$

Now let $G := (\int \phi_i N_{j,p})$ be the Gramian matrix for Q. It follows from a result of de Boor [1, pp. 537–538] and from the easily verified fact that for all sequences $\{a_i\}$

$$\left\| \sum a_i \phi_i \right\|_q \leqslant k \|a\|_q$$

that $\gamma^{-1} D_{k,p}^{-1} \|G^{-1}\|_p \leqslant \|Q\|_p \leqslant k \|G^{-1}\|_p$.

Now, let $H = \operatorname{diag}\{(t_{1+k} - t_1)^{1/p}, \ldots, (t_{n+k} - t_n)^{1/p}\}$, and let $\tilde{G} = (\int \hat{\phi}_i N_{j,\infty})$ where $\hat{\phi}_i = k^{1/p}(t_{i+k} - t_i)^{-1/p}\phi_i$. As above, we have that $\|Q\|_\infty \leqslant k\|\tilde{G}^{-1}\|_\infty$. To use Theorem 1, we must work with certain submatrices of $\tilde{G}$. To insure that these have nice inverses, we make one final assumption:

(e) There is an $r \geqslant 0$ and a $0 < \Gamma$ depending on r such that for every set E of the form $E = E_i = \{j\colon |j - i| \leqslant r + k\}$, we have $\|G_E^{-1}\|_\infty \leqslant \Gamma$.

By the results of [7] and the inequalities proved above we see that (e) will hold if the projections $Q_E\colon L_p \longrightarrow \operatorname{span}\{N_{j,p}\colon j \in E\}$ are uniformly bounded on L_p; here $Q_E f = s$ if and only if $\int \phi_j(f - s) = 0$, $j \in E$. Note that r in general will depend on γ.

With all of these assumptions, let $\sigma \geqslant 1$. By [7], there is an $r \geqslant 0$ such that for all l, $G_{E_l}^{-1} = D_l + C_l$, where $c_{ij} = 0$, if $|i-j| \leqslant r$ and $\|G_{E_l} C_l\|_\infty < \frac{1}{2}\sigma^{-1/p}$. Consequently, if

$$\max_{|i-j|\leqslant r+k} \frac{t_{i+k} - t_i}{t_{j+k} - t_j} \leqslant \sigma,$$

then $\|\tilde{G}_{E_l}\tilde{C}_l\|_\infty = \|H^{-1}G_{E_l}C_lH\|_\infty < \frac{1}{2}$. This implies the condition of Theorem 1 that $d_l < 1$. Again, interpreting Theorem 1 in operator form, we have the inequality

$$\|G^{-1}\|_\infty \leqslant 2 \max_l \sigma^{1/p} \|G_{E_l}^{-1}\|_\infty \leqslant 2\sigma^{1/p}\Gamma.$$

We summarize this in

THEOREM 2. *Let Q: $L_p \longrightarrow \operatorname{span}\{N_{j,p}\}$ be a projection satisfying* (a)–(e). *Then, given $\sigma \geqslant 1$, there is an $r = r(\sigma) \geqslant 0$ such that for all knot sequences $\{t_i\}_{i=1}^{n+k}$ satisfying*

$$\max_{|i-j|\leqslant r+k} \frac{t_{i+k} - t_i}{t_{j+k} - t_j} \leqslant \sigma, \qquad \|Q\|_\infty \leqslant 2\sigma^{1/p}\Gamma.$$

In particular, the least squares projection L satisfies $\|L\|_\infty \leqslant \text{const}\ \sigma^{1/2}$.

It is known that the L_∞ norm of the least squares projection can be bounded in terms of the quantity $\tau := \max_{i,j}\{(t_{i+k} - t_i)/(t_{j+k} - t_j)\}$, [5]. It is also known that there is a constant $\lambda_k > 1$ with the property that: given $1 \leqslant \lambda < \lambda_k$ there exists $K = K(\lambda)$ such that $\max_{|i-j|\leqslant 1}\{(t_{i+k} - t_i)/(t_{j+k} - t_j)\} \leqslant \lambda$ implies that the L_∞ norm of the least squares projection is bounded by K, [4]. Theorem 2 is sort of a compromise between these results: it allows partitions irregular in the sense that the global mesh ratios can become arbitrarily large; and it allows ratios of adjacent mesh intervals to be fairly large, as long as "on the average" they are not too big. Here, "on the average" is determined by the parameter σ in the theorem.

Theorem 2 also says that if the least squares projection is bounded sufficiently nicely for all partitions of a fixed cardinality, then it is bounded for all partitions. So that, theoretically, one might be able to solve the general problem of bounding the least squares projection independently of the partition by showing that its Gramian is sufficiently well behaved for all partitions of fixed size.

School of Mathematics
Georgia Institute of Technology
Atlanta, Georgia 30332

1. C. DE BOOR, "Bounding the error in spline interpolation," *SIAM Rev.*, v. 16, 1974, pp. 531–544.

2. C. DE BOOR, "On the convergence of odd-degree spline interpolation," *J. Approximation Theory*, v. 1, 1968, pp. 452–463.

3. C. DE BOOR, "On bounding spline interpolation," *J. Approximation Theory*, v. 14, 1975, pp. 191–203.

4. C. DE BOOR, *Odd-Degree Spline Interpolation at a Bi-Infinite Knot Sequence*, MRC TSR #1666, August, 1976.

5. C. DE BOOR, "A bound on the L_∞-norm of the L_2-approximation by splines in terms of a global mesh ratio," *Math. Comp.*, v. 30, 1976, pp. 765–771.

6. S. DEMKO, "Local approximation properties of spline projections," *J. Approximation Theory*, v. 19, 1977, pp. 176–185.

7. S. DEMKO, "Inverses of band matrices and local convergence of spline projections," *SIAM J. Numer. Anal.*, v. 14, 1977, pp. 616–619.

8. J. DOUGLAS, JR., T. DUPONT & L. WAHLBIN, "Optimal L_∞ error estimates for Galerkin approximations to solutions of two-point boundary value problems," *Math. Comp.*, v. 29, 1975, pp. 475–483.

9. R. S. VARGA, "On diagonal dominance arguments for bounding $\|A^{-1}\|_\infty$," *Linear Algebra and Appl.*, v. 14, 1976, pp. 211–217.