An optimal bound for inverses of triangular matrices with monotone entries

Kenneth S. Berenhaut*, Richard T. Guy and Nathaniel G. Vish

Department of Mathematics, Wake Forest University, Winston-Salem, NC 27109, USA

Communicated by V. Mehrmann

(Received 9 June 2008; final version received 13 August 2009)

This article provides a new bound for 1-norms of inverses of positive triangular matrices with monotonic column entries. The main theorem refines a recent inequality established in Vecchio and Mallik [Bounds on the inverses of non-negative lower triangular Toeplitz matrices with monotonicity properties, Linear Multilinear Alg., 55 (2007), pp. 365–379]. The results are shown to be in a sense best possible under the given constraints.

**Keywords:** inverse matrix; monotone entries; triangular matrix; recurrence relations; optimal bound

**AMS Subject Classifications:** 15A09; 39A10; 15A57; 15A60

1. Introduction

This note provides a new bound for 1-norms of positive triangular matrices with monotonic column entries. The main theorem refines a recent inequality of Vecchio and Mallik [11]. We refer the reader to Vecchio [10] and Vecchio and Mallik [11] (and the references therein) for discussion of applications, particularly those to stability analysis of linear methods for solving Volterra integral equations (VIE). Other references of the topic include [4–7,9]. An application to VIE under Direct Quadrature is given in Example 1, below.

The matrices of interest here are $n \times n$ truncations of infinite lower triangular (real) matrices i.e.

$$A_{n} = \begin{bmatrix}
a_{1,1} & a_{2,1} & a_{2,2} \\
a_{3,1} & a_{3,2} & a_{3,3} & \cdots \\
\vdots & \vdots & \vdots & \ddots \\
a_{n,1} & \cdots & a_{n,n-1} & a_{n,n}
\end{bmatrix}. \quad (1)$$

The following result was proven in [11].

*Corresponding author. Email: berenhks@wfu.edu
THEOREM 1.1 Assume that

(i) \( a_{i,j} \geq a > 0, \ j = 1, \ldots, n, \ i = j, \ldots, n, \)
(ii) \( a_{j,j} \geq a_{j+1,j} \geq \cdots \geq a_{n,j}, \ j = 1, \ldots, n, \)

and let

\[
a_{\min} = \min_{i=1, \ldots, n} \{a_{i,i}\},
\]

and \( B_n = [b_{i,j}] \) be the inverse of the lower triangular matrix \( A_n. \) Then

\[
\|B_n\|_1 \leq \frac{1}{a_{\min}} + \frac{2}{a}.
\]

(3)

The result in (3) was first proven in the case of triangular Toeplitz matrices in [10] and improved, in this case, to the following in [3].

THEOREM 1.2 Suppose that the sequence \( \{a_i\}_{i \geq 0} \) satisfies

\[
a_0 \geq a_1 \geq a_2 \geq \cdots \geq a_n \geq a > 0,
\]

(4)

for some constant \( a \) and all \( n \) and

\[
C_n = \begin{bmatrix}
a_0 \\
a_1 & a_0 \\
a_2 & a_1 & a_0 \\
\vdots & \ddots & \ddots \\
a_n & \cdots & a_1 & a_0
\end{bmatrix}.
\]

(5)

Then

\[
\|C_n^{-1}\|_1 \leq \frac{2}{a} \left(1 - \rho(a, a_0)\left[\frac{a}{2}\right]\right)
\]

(6)

where \( \rho \) is the inverse ratio defined via

\[
\rho(x,y) = 1 - x/y,
\]

(7)

and, in particular

\[
\|C_n^{-1}\|_1 \leq \frac{2}{a},
\]

(8)

independent of \( a_0 \) and \( n. \)

The following result was recently proven in [2]. The theorem extends Theorem 1.2 to non-Toeplitz matrices and refines Theorem 1.1 in the case of constant diagonal.

THEOREM 1.3 Assume that the hypotheses of Theorem 1.1 are satisfied and in addition that

\[
a_{1,1} \leq a_{2,2} \leq \cdots \leq a_{n,n}.
\]

(9)

Then

\[
\|B_n\|_1 \leq \frac{2}{a} \left(\frac{a_{n,n}}{a_{1,1}}\right) \left(1 - \frac{\rho(a, a_{n,n})\left[\frac{a}{2}\right]}{2} + \frac{\rho(a, a_{n,n})\left[\frac{a}{2}\right]}{2}\right).
\]

(10)
In particular, if
\[ a_{1,1} = a_{2,2} = \cdots = a_{n,n} = a^*, \]
then
\[ \| B_n \|_1 \leq \frac{2}{a} \left( 1 - \frac{\rho(a, a^*) [\xi]}{2} + \frac{\rho(a, a^*) [\xi]}{2} \right) \]
and hence
\[ \| B_n \|_1 < \frac{2}{a}, \]
independent of \( a^* \).

Note that triangular matrices satisfying (11) arise in the study of linear groups (see e.g. [8]) and are particularly important in the theory of matrix decompositions.

Motivated by the above results, here we will improve on Theorem 1.1 by showing that the term \( 1/a_{\min} \) in (3) is not needed and in addition that the new bound is in a sense best possible. In particular we will prove the following.

**Theorem 1.4** Assume that the hypotheses of Theorem 1.1 are satisfied. Then
\[ \| B_n \|_1 \leq \frac{2}{a}. \]

In fact, setting
\[ A_n(a) = \{ A = [a_{ij}]_{n \times n} \mid A \text{ satisfies } (1), \ (i) \text{ and } (ii) \}, \]
we have
\[ \sup_{A \in \bigcup_{n=1}^\infty A_n(a)} \| A^{-1} \|_1 = 2/a. \]

The reader is referred to [1] for some discussion of bounds for inverses of matrices of the form in (1) when the condition of monotonicity within columns is replaced with that within rows.

**Example 1** A nice application of a priori bounds as in (3) and (14) to numerical solution of Volterra integral equations is given in Section 3.2 of [11]. In particular, consider a VIE of the second type given by
\[ y(t) = g(t) + \int_0^t k(t, s)y(s)ds, \quad t \in [0, T], \]
where \( y, g \) and \( k \) are real valued functions with \( y(t) \) and \( g(t) \) continuous in \( 0 \leq t \leq T \) and \( k(t, s) \) continuous for \( 0 \leq s \leq t \leq T \). The application of a Direct Quadrature (DQ) method leads to the discrete equation
\[ y_i = \tilde{g}_i + h \sum_{l=n_0}^i w_{i,l} k_{i,l} y_l, \quad i \geq n_0, \quad y_0, \ldots, y_{n_0-1} \text{ given}, \]
where
\[ t_i = ih, \quad h = T/N, \quad k_{i,l} = k(t_i, t_t), \quad y_i \approx y(t_i), \quad \tilde{g}_i = g(t_i) + h \sum_{l=0}^{n_0-1} w_{i,l} k_{i,l} y_l, \]
Table 1. Values of $c_1$ and $c_2$ for some DQ methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>$c_1$</th>
<th>$c_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backward Euler (order 1, $n_0 = 1$)</td>
<td>$\infty$</td>
<td>1</td>
</tr>
<tr>
<td>Trapezoidal (order 2, $n_0 = 1$)</td>
<td>2</td>
<td>1/2</td>
</tr>
<tr>
<td>Simpson + Trapezoidal (order 3, $n_0 = 2$)</td>
<td>6/5</td>
<td>1/3</td>
</tr>
<tr>
<td>($\rho,\sigma$) reducible (order 3, $n_0 = 2$)</td>
<td>3/2</td>
<td>5/12</td>
</tr>
</tbody>
</table>

and $[w_{i,j}]$ is the matrix of coefficients for the given quadrature method. The numerical solution (18) can be written in the form

$$y_n = A_n^{-1} g_n$$

where

$$y_n = [y_{n_0}, y_{n_0+1}, \ldots, y_{n-1+n_0}]^T \in \mathbb{R}^n, \quad g_n = [g_{n_0}, g_{n_0+1}, \ldots, g_{n-1+n_0}]^T \in \mathbb{R}^n,$$

and the lower triangular matrix $A_n$ has diagonal entries $a_{i,i} = 1 - hw_{i,i}k_{i,i}$ and off-diagonal entries $a_{i,j} = -hw_{i,j}k_{i,j}$ for $j < i$.

Now, set $c_1 = \max_{1 \leq i \leq n} |w_{j+1,i} - w_{j,i}|$ and $c_2 = \min_{1 \leq i \leq n} w_{i,i}$. Values for $c_1$ and $c_2$ for some common DQ methods are provided in Table 1. Specific values of $w_{i,j}$ can be found in [11].

The following theorem is proven in [11].

**Theorem 1.5** ([11]) Assume that

(i) $k(t,s) \leq 0, \quad t \geq 0, \quad 0 \leq s \leq t$,
(ii) $\partial k(t,s)/\partial s \geq 0, \quad t \geq 0, \quad 0 \leq s \leq t$,
(iii) $0 < K_m \leq -k(t,s) \leq K_M$,
(iv) $h \leq c_1/K_M$,
(v) $g(t) \leq g$.

Then the solution of the DQ methods under consideration satisfy

$$\|y_n\|_1 \leq \frac{\hat{g}}{1 + hc_2K_m} + \frac{2\hat{g}}{hc_2K_m}$$

with

$$\hat{g} = g + h(1 + 1/c_1)K_M \sum_{l=0}^{n-1} |y_l|.$$  \hspace{1cm} (21)

Note that there was a small typographical error in the equation corresponding to (20) in the statement of Theorem 3.1 in [11]. If Theorem 1.4 is used in place of Theorem 1.1, the sharper bound

$$\|y_n\|_1 \leq \frac{2\hat{g}}{hc_2K_m}$$

follows directly. For details regarding the proof of Theorem 1.5 and some added discussion see [11].

We now turn to a proof of Theorem 1.4.
2. Preliminaries and notation

In order to prove Theorem 1.4 we will need several results mentioned in [11]; for completeness we will prove the necessary preliminaries here.

First, define the sequence \( \{U_{i,j}\} \) via \( U_{i,j} = 0 \) if \( j > i \), \( U_{j,j} = 1/a_{j,j} \) and for \( i > j \),

\[
U_{i,j} = \frac{1}{a_{i,i}} - \sum_{l=j}^{i-1} \frac{a_{i,l}}{a_{i,i}} U_{l,j}. \tag{23}
\]

Note that for \( i \geq j \), \( U_{i,j} = u_{i,i} \), where \( [u_{i,j}] \) is the fundamental matrix as defined in Equation (2.10) in [11].

We have the following lemma (see also Theorem 2.1 in [11]).

**Lemma 2.1** Under the assumptions of Theorem 1.1,

\[
U_{m,j} \geq 0, \tag{24}
\]

for all \( m, j \geq 1 \).

**Proof** For fixed \( j \geq 1 \), we have \( U_{i,j} = 1/a_{i,j} > 0 \). Thus assume the result holds for \( m = j, \ldots, i-1 \), for some \( i \geq j + 1 \). Then, by Assumption (ii), the induction hypothesis, and the definition in (23),

\[
U_{i,j} = \frac{1}{a_{i,i}} \left( 1 - \sum_{l=j}^{i-1} \frac{a_{i,l}}{a_{i,i}} U_{l,j} \right) \geq \frac{1}{a_{i,i}} \left( 1 - \sum_{l=j}^{i-1} \frac{a_{i-1,l}}{a_{i-1,i-1}} U_{l,j} - \frac{1}{a_{i,i}} U_{i,i} - U_{i-1,i-j} \right) = 0 \tag{25}
\]

and the result follows. \( \blacksquare \)

Note that in the line corresponding to (25) in the proof of Theorem 2.1 in [11] there is a missing negative sign.

Now, note that the lower triangular matrix \( B_n = [b_{i,j}] = A_n^{-1} \) satisfies \( b_{j,j} = 1/a_{j,j} \) and for \( 1 \leq j < i \leq n \),

\[
b_{i,j} = \sum_{l=j}^{i-1} - \frac{a_{i,l}}{a_{i,i}} b_{l,j}, \tag{26}
\]

(see e.g. [1]).

The next lemma is essentially a restatement of Equation (2.13) in [11].

**Lemma 2.2** For all \( 1 \leq j \leq i \leq n \),

\[
b_{i,j} = U_{i,j} - U_{i,j+1}. \tag{27}
\]

**Proof** For fixed \( j \geq 1 \), we have \( U_{j,j} - U_{j,j+1} = U_{j,j} = 1/a_{j,j} \) and for \( i > j \),

\[
U_{i,j} - U_{i,j+1} = \frac{1}{a_{i,i}} - \sum_{l=j}^{i-1} \frac{a_{i,l}}{a_{i,i}} U_{l,j} - \left( \frac{1}{a_{i,i}} - \sum_{l=j+1}^{i-1} \frac{a_{i,l}}{a_{i,i}} U_{l,j+1} \right) = \sum_{l=j}^{i-1} - \frac{a_{i,l}}{a_{i,i}} (U_{l,j} - U_{l,j+1}). \tag{28}
\]
The last equality in (28) follows since $U_{j,j+1} = 0$ (contrast with the definition $u_{j+1,j} = 1$ in [11]).

The result follows upon comparing (26) and (28). ■

3. Proof of the main theorem

We are now in a position to prove Theorem 1.4 (contrast the proof below with that of Theorem 2.3 in [11]; see the remark following the proof).

Proof of Theorem 1.4 First note that by employing the definition in (23), (i), (ii) and Lemma 2.1, we have

$$U_{j,j} \leq \frac{a_{j+1,j+1}}{a} \left( \frac{a_{j+1,j}}{a_{j+1,j+1}} U_{j,j} \right) = \frac{a_{j+1,j+1}}{a} \left( \frac{1}{a_{j+1,j+1}} - U_{j+1,j} \right) \leq \frac{1}{a}. \quad (29)$$

Similarly

$$U_{j,j} + U_{j+1,j} \leq \frac{a_{j+2,j+2}}{a} \left( \frac{a_{j+2,j}}{a_{j+2,j+2}} U_{j,j} + \frac{a_{j+2,j+1}}{a_{j+2,j+2}} U_{j+1,j} \right)$$

$$= \frac{a_{j+2,j+2}}{a} \left( \frac{1}{a_{j+2,j+2}} - U_{j+2,j} \right) \leq \frac{1}{a}, \quad (30)$$

and in general for $m \geq j$,

$$\sum_{i=j}^{m} U_{i,j} \leq \frac{a_{m+1,m+1}}{a} \left( \sum_{i=j}^{m} \frac{a_{m+1,i}}{a_{m+1,m+1}} U_{i,j} \right)$$

$$= \frac{a_{m+1,m+1}}{a} \left( \frac{1}{a_{m+1,m+1}} - U_{m+1,j} \right) \leq \frac{1}{a}. \quad (31)$$

Employing Lemma 2.2, we have

$$\|B_n\| = \max_{j=1,\ldots,n} \sum_{i=j}^{n} |b_{i,j}| = \max_{j=1,\ldots,n} \sum_{i=j}^{n} |U_{i,j} - U_{i,j+1}|$$

$$\leq \max_{j=1,\ldots,n} \left( \sum_{i=j}^{n} U_{i,j} + \sum_{i=j}^{n} U_{i,j+1} \right). \quad (32)$$

Now, noting that $U_{j,j+1} = 0$, (31) gives

$$\sum_{i=j}^{n} U_{i,j} \leq \frac{1}{a} \quad \text{and} \quad \sum_{i=j}^{n} U_{i,j+1} \leq \frac{1}{a}. \quad (33)$$

The bound in (14) then follows upon applying (33) in (32).

Considering the apparent looseness in the inequality in (32), it is perhaps surprising that for large $n$, (14) is in fact optimal. In order to prove (16) we need to show that, in the limit, the bound in (14) is attained. To that end, suppose $a_{i,j} = a > 0$ for $i-j \in \{0, 1\}$ and $a_{i,j} = a$ otherwise. It is easy to verify in this case, that for $1 \leq j \leq i \leq n$,

$$b_{i,j} = (-1)^{i-j} \frac{1}{a^*} \left( 1 - \frac{a}{a^*} \right)^{\left\lfloor \frac{i-j}{2} \right\rfloor}, \quad (34)$$
and hence,

\[
\|A_n^{-1}\|_1 = \sum_{i=1}^{n} |b_{i,1}|
\]

\[
= \sum_{i=1}^{n} \frac{1}{a^*} (1 - \frac{a}{a^*})^{\frac{|i|}{2}} = \sum_{i=0}^{n-1} \frac{1}{a^*} \left(1 - \frac{a}{a^*}\right)^{\frac{i+1}{2}}
\]

\[
= \frac{1}{a^*} \left(\sum_{i=0}^{\frac{n-1}{2}} \left(1 - \frac{a}{a^*}\right)^i + \sum_{i=0}^{\frac{n-1}{2}} \left(1 - \frac{a}{a^*}\right)^i\right)
\]

\[
= \frac{2}{a} \left(1 - \frac{\rho(a, a^*)^{\frac{n}{2}} + \rho(a, a^*)^{\frac{n}{2}}}{2}\right).
\]

(35)

The result follows upon letting \( n \) tend to infinity in (35).

\[\blacksquare\]

\textbf{Remark}  Note that the essential subtle difference between the proof of (14) above and (3) in [11] lies in the definition \( U_{j+1,j}^- := 0 \). In contrast, in [11], following the conventional definition of the fundamental matrix, \( u_{j,j+1} := 1 \). The approach here allows for the second equality in (32) to hold (via Lemma 2.2), whereas in the proof of Theorem 2.3 in [11], there is a need to separate off the term corresponding to \( b_{j,j} = 1/a_{j,j} \).

\textbf{References}


