Tridiagonal Matrices: Invertibility and Conditioning*

L. Brugnano and D. Trigiante
Dipartimento di Matematica
Campus Universitario
Trav. 200 Re David
I-70125 Bari, Italy

Submitted by Peter Lancaster

ABSTRACT

Tridiagonal matrices arise in a large variety of applications. Most of the time they are diagonally dominant, and this is indeed the case most extensively studied. In this paper we study, in a unified approach, the invertibility and the conditioning of such matrices. The results presented provide practical criteria for a tridiagonal and irreducible matrix to be both invertible and "well conditioned." An application to a singular perturbation boundary value problem is then presented.

1. INTRODUCTION

There are many applications in numerical analysis which lead to solving either tridiagonal systems or second order difference equations such as

$$\tau_i y_{i+1} + y_i + \sigma_{i-1} y_{i-1} = 0. \tag{1.1}$$

The two problems are related. For boundary value problems, the connection is evident. For initial value problems, the connection arises when we try to obtain the minimal solution of (1.1). More precisely, often (1.1) has a minimal and a maximal solution [6]: in many applications, one is interested in the former. If we try to solve (1.1) iteratively, small errors—either in the initial

*Work supported by the "Ministero della Ricerca Scientifica," project 40%.

655 Avenue of the Americas, New York, NY 10010 0024-3795/92/$5.00

conditions or arising from the use of finite arithmetic—are amplified, and a maximal solution is eventually obtained [6]. One then is obliged to solve an approximate boundary problem (Miller and Olver's algorithm [6, 10, 13]).

Often, especially for problems derived from the discretization of PDEs, the matrices are diagonally dominant, and much is known in this case. In this paper, we shall study the conditions needed for the tridiagonal matrix associated to (1.1) with boundary conditions to be nonsingular and well conditioned. If τ_i and σ_i are independent of i, such conditions are known [7], and essentially they amount to requiring that the characteristic polynomial of (1.1) have two real roots, r_1 and r_2: one inside and one outside the unit circle. In the case where τ_i and σ_i are not constant, there are partial results requiring essentially that τ_i and σ_i should be slowly varying in the neighborhood of an asymptotic value [6, 7]. We shall study the problem (1.1) from the matrix point of view, and we shall find new conditions for the problem to be well conditioned. The results presented here will provide a quite practical tool for testing the well-conditioning of irreducible, tridiagonal matrices.

2. SUFFICIENT CONDITION FOR NONSINGULARITY

Even though the problem has been extensively studied [5, 11, 15], the main result concerning the conditions on the invertibility will be restated in this section, since it turns out to be an easy consequence of the notation and of Lemma 2.1, which we need to introduce for further discussion. Let us consider, for convenience, the "normalized" matrices associated to the problem (1.1):

$$T = \begin{bmatrix} 1 & \tau_1 & & & & & \\ \sigma_1 & 1 & \tau_2 & & & & \\ & \sigma_2 & 1 & \cdot & & & \\ & & \cdot & \cdot & \cdot & \cdot & \\ & & & \cdot & \cdot & \cdot & \\ & & & & \cdot & \cdot & \tau_{n-1} \\ & & & & & \sigma_{n-1} & 1 \end{bmatrix}, \tag{2.1}$$

where $\sigma_i, \tau_i \neq 0$, $i = 1, \dots, n-1$. As a convention we assume $\sigma_0 = \sigma_n = \tau_0 = \tau_n = 0$.

First, we shall review sufficient conditions needed to obtain the factorization

$$T = LDU, \tag{2.2.1}$$

where

$$D = \operatorname{diag}(d_1, \ldots, d_n),$$

$$L = \begin{bmatrix} 1 & & & & & \\ \hat{\sigma}_1 & 1 & & & 0 & \\ & \hat{\sigma}_2 & 1 & & & \\ & & \cdot & \cdot & & \\ & & & \cdot & \cdot & \\ & & & & \cdot & \cdot \\ & & & & \hat{\sigma}_{n-1} & 1 \end{bmatrix},$$

$$U = \begin{bmatrix} 1 & \hat{\tau}_1 & & & & \\ & 1 & \hat{\tau}_2 & & & \\ & & 1 & \cdot & & \\ & & & \cdot & \cdot & \\ & & & & \cdot & \cdot \\ & 0 & & & \cdot & \hat{\tau}_{n-1} \\ & & & & & 1 \end{bmatrix}, \qquad (2.2.2)$$

$$\hat{\sigma}_i = \sigma_i d_i^{-1}, \quad \hat{\tau}_i = \tau_i d_i^{-1}, \qquad \begin{cases} d_{i+1} = 1 - \sigma_i \tau_i d_i^{-1}, & i = 1, \ldots, n-1, \\ d_1 = 1. \end{cases}$$

$$(2.2.3)$$

Obviously, the factorization (2.2.1) exists iff $d_i \neq 0$, $i = 1, \ldots, n-1$, while T is invertible iff $d_i \neq 0$, $i = 1, \ldots, n$.

Let us define the following functions:

$$x_+ = \begin{cases} x & \text{if} \quad x \geqslant 0, \\ 0 & \text{if} \quad x < 0, \end{cases} \qquad x_- = -(-x)_+.$$

Let us consider first the particular case of Toeplitz tridiagonal matrices, that is, the case in which $\sigma_i = \sigma$, $\tau_i = \tau$, $i = 1, \ldots, n-1$.

LEMMA 2.1. *Let* $\Delta = 1 - 4(\sigma\tau)_+ \geqslant 0$, *and* $m > 0$ *be such that*

$$\frac{1 - \Delta^{1/2}}{2} \leqslant m \leqslant \frac{1 + \Delta^{1/2}}{2}.$$

Then for d_i defined by (2.2.3), *it results that $d_1 = 1 \geqslant m$, and for $i \geqslant 2$*

$$m \leqslant d_i \leqslant 1-(\sigma\tau)_- m^{-1}.$$

Proof. From the hypothesis on Δ it follows that the equation $x^2 - x + (\sigma\tau)_+ = 0$ has real roots. For such values of m it is true that $m^2 - m + (\sigma\tau)_+ \leqslant 0$, that is, $0 < m \leqslant 1-(\sigma\tau)_+ m^{-1}$. By setting $f(x) = 1-(\sigma\tau)x^{-1}$, we have $(\sigma\tau)f_x \geqslant 0$, and then the minimum of $f(x)$ in the domain $D_m = \{x : x \geqslant m\}$ is given by $1-(\sigma\tau)_+ m^{-1}$. It follows then that $m \leqslant f(x)$ for all $x \geqslant m$, and then $d_i \geqslant m$ for all $i \geqslant 2$. The proof ends on considering that $f(x) \leqslant 1-(\sigma\tau)_- x^{-1} \leqslant 1-(\sigma\tau)_- m^{-1}$. ■

We shall take the most favorable value of m (because it minimizes the interval of variation of d_i), that is, $m = (1+\Delta^{1/2})/2$.

Going back to the general case with σ_i and τ_i varying, the previous result can be generalized. Denoting

$$\Delta_i = 1 - 4(\sigma_i\tau_i)_+, \qquad (\sigma\tau)_- = \min_i \{(\sigma_i\tau_i)_-\}, \qquad m = \min_i \left\{ \frac{1+\Delta_i^{1/2}}{2} \right\},$$

one has (see also [4] for a more general version)

THEOREM 2.1. *If $\Delta_i \geqslant 0$ for $i = 1,\dots,n-1$, then*

$$m \leqslant d_i \leqslant 1-(\sigma\tau)_- m^{-1} \qquad \textit{for} \quad i = 1,\dots,n.$$

Proof. Let $f_i(x) = 1-(\sigma_i\tau_i)x^{-1}$, $D_m = \{x : x \geqslant m\}$. Since $\Delta_i \geqslant 0$ and

$$\frac{1-\Delta_i^{1/2}}{2} \leqslant m \leqslant \frac{1+\Delta_i^{1/2}}{2},$$

from Lemma 2.1 it follows that for $x \geqslant m$, $f_i(x) \geqslant m$. Moreover

$$f_i(x) \leqslant 1-(\sigma_i\tau_i)_- x^{-1} \leqslant 1-(\sigma_i\tau_i)_- m^{-1}. \qquad ■$$

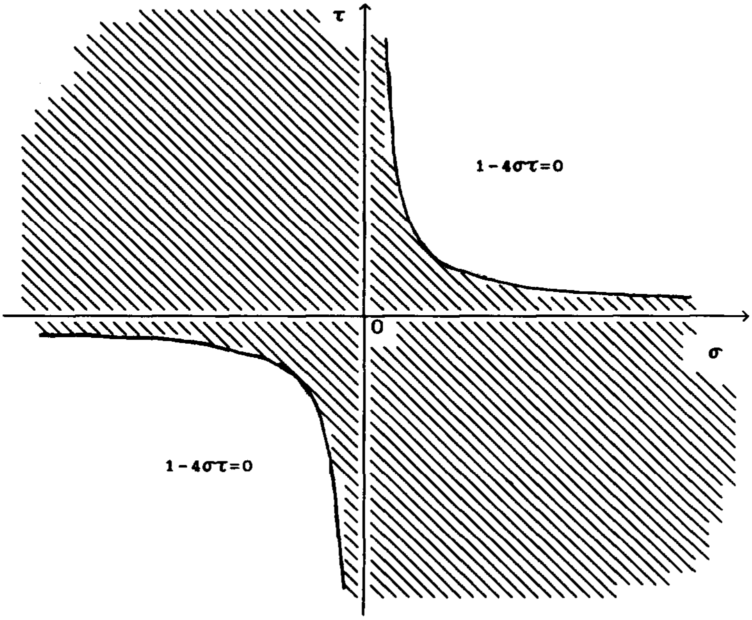

FIG. 1.

The condition $\Delta_i \geqslant 0$ implies that in the (σ, τ) plane, the points (σ_i, τ_i) must lie inside the region bounded by the hyperbola $1 - 4\sigma\tau = 0$ (see Figure 1), that is,

$$\sigma_i \tau_i \leqslant \tfrac{1}{4}, \qquad i = 1, \ldots, n-1. \tag{2.4}$$

Outside this region it may happen that the matrix T is singular (see for example [1, 15] for the case of Toeplitz matrices).

3. SUFFICIENT CONDITIONS FOR WELL-CONDITIONING

We shall now suppose that the conditions required by Theorem 2.1 are satisfied, that is, (σ_i, τ_i) are inside the region of Figure 1. This does not ensure that the matrix T is well conditioned. We say that the matrix T is *well conditioned* if its condition number $\kappa(T)$ is bounded from above by a quantity independent of the dimension n of the matrix. In the applications, however, this quantity can be allowed to grow like a small power of n, for example n or n^2 [7]. In this case, we shall say that T is *weakly well conditioned*. We shall derive now additional conditions to get $\kappa(T)$ independent of n, or growing, at most, as a power of n. In order to get the result, we need an estimate of $\|T^{-1}\|$. From (2.2.1) we have

$$\|T^{-1}\| \leqslant \|L^{-1}\| \cdot \|D^{-1}\| \cdot \|U^{-1}\|.$$

From Theorem 2.1 we have

$$\|D^{-1}\| = \left(\min_i \{d_i\}\right)^{-1} \leqslant m^{-1}.$$

The previous estimate is independent of the dimension of the matrix. Now we shall derive estimates for $\|L^{-1}\|$ and $\|U^{-1}\|$. Let us define the sequences

$$\ell_1 = 1,$$

$$\ell_{i+1} = 1 + \ell_i \|\hat{\sigma}_i\|, \qquad i = 1 \dots n-1; \tag{3.1.1}$$

$$\boldsymbol{m}_1 = 1,$$

$$\boldsymbol{m}_{i+1} = 1 + \boldsymbol{m}_i \|\hat{\sigma}_{n-i}\|, \qquad i = 1 \dots n-1; \tag{3.1.2}$$

$$\boldsymbol{u}_1 = 1,$$

$$\boldsymbol{u}_{i+1} = 1 + \boldsymbol{u}_i \|\hat{\tau}_{n-i}\|, \qquad i = 1 \dots n-1; \tag{3.1.3}$$

$$\boldsymbol{v}_1 = 1,$$

$$\boldsymbol{v}_{i+1} = 1 + \boldsymbol{v}_i \|\hat{\tau}_i\|, \qquad i = 1 \dots n-1, \tag{3.1.4}$$

where $\hat{\sigma}_i$ and $\hat{\tau}_i$ are defined in (2.2.2), (2.2.3). The following theorem is essentially due to Higham [9].

THEOREM 3.1. *For the sequences* (3.1.1)–(3.1.4),

$$\|L^{-1}\|_\infty = \max_i \{\ell_i\}, \qquad \|L^{-1}\|_1 = \max_i \{\boldsymbol{m}_i\},$$

$$\|U^{-1}\|_\infty = \max_i \{\boldsymbol{u}_i\}, \qquad \|U^{-1}\|_1 = \max_i \{\boldsymbol{v}_i\}.$$

Proof. See [9]. ■

We shall now consider the abovementioned sequences in more detail. It is easy to check that

$$\ell_i = 1 + \sum_{j=1}^{i-1} \prod_{k=j}^{i-1} |\hat{\sigma}_k|, \tag{3.2.1}$$

$$\mathscr{m}_i = 1 + \sum_{j=1}^{i-1} \prod_{k=j}^{i-1} |\hat{\sigma}_{n-k}|, \tag{3.2.2}$$

$$\mathscr{u}_i = 1 + \sum_{j=1}^{i-1} \prod_{k=j}^{i-1} |\hat{\tau}_{n-k}|, \tag{3.2.3}$$

$$\mathscr{v}_i = 1 + \sum_{j=1}^{i-1} \prod_{k=j}^{i-1} |\hat{\tau}_k|. \tag{3.2.4}$$

In order to evaluate the products in the previous expressions, let us define the sequence $\{\omega_k\}$ by setting

$$|\hat{\sigma}_k| = \frac{\omega_{k-1}}{\omega_k}, \qquad \omega_0 = 1, \quad \omega_{-1} = 0. \tag{3.3}$$

It follows that

$$\prod_{k=j}^{i-1} |\hat{\sigma}_k| = \frac{\omega_{j-1}}{\omega_{i-1}}, \qquad j < i. \tag{3.4}$$

The same can be done for the other products. Let us examine now the sequence $\{\omega_k\}$ in the two cases $\sigma_i \tau_i < 0$ and $\sigma_i \tau_i > 0$, $i = 1, \ldots, n-1$.

3.1. *The First Case: $\sigma_i \tau_i < 0$*

In this case, from (2.2.3) we have

$$\frac{\omega_k}{\omega_{k+1}} = |\hat{\sigma}_{k+1}| = \frac{|\sigma_{k+1}|}{|d_{k+1}|} = \frac{|\sigma_{k+1}|}{1 + |\tau_k| \cdot |\hat{\sigma}_k|} = \frac{|\sigma_{k+1}|}{1 + |\tau_k| \cdot \omega_{k-1} / \omega_k},$$

and thus the following iterative scheme is obtained:

$$\omega_{k+1}|\sigma_{k+1}| = \omega_k + |\tau_k|\omega_{k-1}.$$

By setting

$$a_k = |\sigma_k|^{-1}, \qquad b_k = |\tau_{k-1}|\cdot|\sigma_k|^{-1}, \tag{3.5}$$

we have, finally,

$$\omega_k = a_k\omega_{k-1} + b_k\omega_{k-2}, \qquad k \geqslant 1,$$
$$\omega_0 = 1, \qquad \omega_{-1} = 0. \tag{3.6}$$

This is a linear second order difference equation, with nonconstant and positive coefficients. From the positiveness of a_k, b_k, the positiveness of the sequence $\{\omega_k\}$ follows. But we need additional conditions for the products (3.4) to be bounded. In particular, we shall prove that it is sufficient that

$$a_k + b_k > 1, \qquad k \geqslant 1. \tag{3.7}$$

Let us define the following quantities:

$$\hat{\gamma} = \max_k\{a_k + b_k\}, \qquad \underline{\gamma} = \min_k\{a_k + b_k\}. \tag{3.8}$$

It is evident that, if the condition (3.7) is satisfied, then $\hat{\gamma} \geqslant \underline{\gamma} > 1$. Moreover, in the following, the notation $\lfloor x \rfloor$ means the integer part of x.

LEMMA 3.1. *With reference to the sequence defined by* (3.6), *let us define the following sequences*:

$$\underline{\omega}_j = \underline{\gamma}^{\lfloor (j-r)/2 \rfloor}\underline{\omega}, \quad j \geqslant r, \qquad \underline{\omega} = \min\{\omega_r, \omega_{r+1}\},$$

$$\hat{\omega}_j = \hat{\gamma}^{j-r}\hat{\omega}, \quad j \geqslant r, \qquad \hat{\omega} = \max\{\omega_r, \omega_{r+1}\}.$$

If the condition (3.7) *is satisfied and* $\omega_r, \omega_{r+1} > 0$, *then*

$$\underline{\omega}_j \leqslant \omega_j \leqslant \hat{\omega}_j, \qquad j \geqslant r.$$

Proof. By induction on j. Starting from the first inequality, we have, obviously, $\underline{\omega}_j \leqslant \omega_j$ for $j = r$ and $j = r + 1$. Suppose it holds true for $j \leqslant k$. One has for $j = k + 1$

$$\omega_{k+1} = a_{k+1}\omega_k + b_{k+1}\omega_{k-1} \geqslant (a_{k+1} + b_{k+1})\min\{\omega_k, \omega_{k-1}\}$$
$$\geqslant \underline{\gamma}\cdot\underline{\gamma}^{\lfloor (k-r-1)/2 \rfloor}\underline{\omega} = \underline{\gamma}^{\lfloor (k-r+1)/2 \rfloor}\underline{\omega} = \underline{\omega}_{k+1}.$$

In a similar manner, for the second inequality one has $\omega_j \leqslant \hat{\omega}_j$ for $j = r$ and $j = r + 1$. Let us suppose it true for $j \leqslant k$. For $j = k + 1$ we have

$$\omega_{k+1} = a_{k+1}\omega_k + b_{k+1}\omega_{k-1} \leqslant (a_{k+1} + b_{k+1})\max\{\omega_k, \omega_{k-1}\}$$
$$\leqslant \hat{\gamma}\cdot\hat{\gamma}^{k-r}\hat{\omega} = \hat{\gamma}^{k-r+1}\hat{\omega} = \hat{\omega}_{k+1}. \qquad \blacksquare$$

From this result, the following theorem follows.

THEOREM 3.2. *With reference to the sequence defined by* (3.6), *if the condition* (3.7) *is satisfied, then there exists* $k \in \mathbb{N}$ *such that*

$$\frac{\omega_i}{\omega_j} < 1 \quad \text{if } j \geqslant i + k, \qquad \text{for all} \quad i \geqslant 1.$$

Proof. From Lemma 3.1, choosing $r = 1$, it follows that $\lim_{j\to\infty}\omega_j = +\infty$. Let us consider then the generic element ω_i. We define $\eta = \min\{\omega_i, \omega_{i+1}\}$. It may happen that:

(a) $\eta = \omega_{i+1}$. In this case, from Lemma 3.1, we find $\omega_{i+k} > \underline{\gamma}^{\lfloor k/2 \rfloor}\omega_{i+1}$. It follows that the thesis will be true if $\underline{\gamma}^{\lfloor k/2 \rfloor}\omega_{i+1} \geqslant \omega_i$, that is,

$$k \geqslant \left\lfloor 2\left(\frac{\log \max_i\{\omega_i / \omega_{i+1}\}}{\log \underline{\gamma}}\right) + 1 \right\rfloor = \delta, \tag{3.9}$$

where [see (3.3) and Theorem 2.1]

$$\max_i\left\{\frac{\omega_i}{\omega_{i+1}}\right\} = \max_i\{|\hat{\sigma}_i|\} \leqslant \frac{\max_i\{|\sigma_i|\}}{\min_i\{|d_i|\}} \leqslant \max_i\{|\sigma_i|\}\cdot m^{-1}.$$

(b) $\eta = \omega_i$. In this case $\omega_{i+1} > \omega_i$; moreover $\omega_{i+2} = a_{i+2}\omega_{i+1} + b_{i+2}\omega_i > \omega_i$. It follows that $\omega_{i+k} > \omega_i$ for all $k \geqslant 1$. ■

It remains to be seen what happens for $0 < j - i < k$.

Theorem 3.3. *With reference to the sequence defined by* (3.6), *if the condition* (3.7) *is satisfied, it follows that*

$$\frac{\omega_i}{\omega_j} < \alpha\hat{\gamma}^k \qquad \textit{for} \quad 0 < j - i < k.$$

Proof. From Theorem 3.2 it follows that $\omega_{i+k} > \omega_i$. Moreover, from Lemma 3.1 we have $\omega_{i+k}\hat{\gamma}^{(j-i-k)} \leqslant \max\{\omega_j, \omega_{j+1}\}$. Two cases are possible:

(a) $\max\{\omega_j, \omega_{j+1}\} = \omega_j$. It follows that

$$\frac{\omega_i}{\omega_j} \leqslant \frac{\omega_i}{\omega_{i+k}}\hat{\gamma}^{k-(j-i)} < \hat{\gamma}^k.$$

(b) $\max\{\omega_j, \omega_{j+1}\} = \omega_{j+1}$. It follows that

$$\frac{\omega_i}{\omega_j} \leqslant \frac{\omega_i}{\omega_{i+k}}\hat{\gamma}^{k-(j-i)}\frac{\omega_{j+1}}{\omega_j} < \hat{\gamma}^k \max_j\left\{\frac{\omega_{j+1}}{\omega_j}\right\}.$$

The thesis follows by observing that (see Theorem 2.1)

$$\max_j\left\{\frac{\omega_{j+1}}{\omega_j}\right\} = \max_j\left\{\left|\hat{\sigma}_j\right|^{-1}\right\} \leqslant \frac{\max_j\{|d_j|\}}{\min_j\{|\sigma_j|\}} \leqslant \frac{1-(\sigma\tau)_- m^{-1}}{\min_j\{|\sigma_j|\}}.$$ ■

From these results, considering (3.1.1), (3.1.2), (3.2.1), (3.2.2), and Theorem 3.1, it follows that

$$\|L^{-1}\|_\infty = \max_i\{l_i\} = \max_i\left\{1 + \sum_{j=1}^{i-1}\prod_{k=j}^{i-1}|\hat{\sigma}_k|\right\}$$

$$= \max_i\left\{1 + \sum_{j=1}^{i-1}\frac{\omega_{j-1}}{\omega_{i-1}}\right\} = 1 + \max_i\left\{\sum_{j=0}^{i-k-1}\frac{\omega_j}{\omega_i} + \sum_{j=i-k}^{i}\frac{\omega_j}{\omega_i}\right\}$$

$$< 1 + \alpha k\hat{\gamma}^k + \frac{2}{\underline{\gamma} - 1}. \tag{3.10}$$

In a similar manner, it can be shown that $\|L^{-1}\|_1$ also is bounded from above by the same quantity.

REMARK 3.1. By considering (3.5), it is easy to verify that condition (3.7) is equivalent to the following set:

$$\left.\begin{array}{lll} \sigma_k + \tau_{k-1} < 1 & \text{if} & \sigma_k > 0, \quad \tau_{k-1} < 0, \\ \sigma_k - \tau_{k-1} < 1 & \text{if} & \sigma_k > 0, \quad \tau_{k-1} > 0, \\ \sigma_k + \tau_{k-1} > -1 & \text{if} & \sigma_k < 0, \quad \tau_{k-1} > 0, \\ \sigma_k - \tau_{k-1} > -1 & \text{if} & \sigma_k < 0, \quad \tau_{k-1} < 0, \end{array}\right\} \quad k \geqslant 1. \tag{3.11}$$

The geometric interpretation of this set of conditions is the following: the *off-diagonal* elements in the *columns* of the matrix T [that is, the pair (σ_k, τ_{k-1})] must lie in the region of the (σ, τ) plane shaded in Figure 2.

A similar result holds for $\|U^{-1}\|$. By defining

$$c_k = |\tau_k|^{-1}, \qquad e_k = |\sigma_{k-1}| \cdot |\tau_k|^{-1}, \tag{3.12}$$

the following scheme is derived:

$$\phi_k = c_k \phi_{k-1} + e_k \phi_{k-2}, \qquad k \geqslant 1,$$
$$\phi_0 = 1, \qquad \phi_{-1} = 0.$$

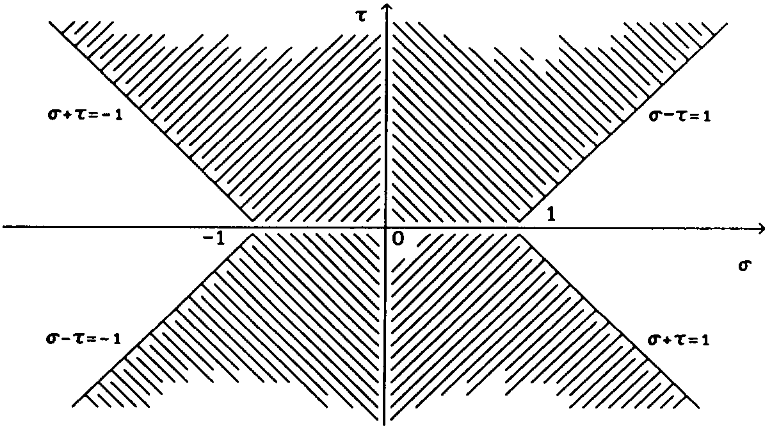

FIG. 2.

In order to bound $\|U^{-1}\|$, one needs to bound the products in (3.2.3) and (3.2.4). The condition corresponding to (3.7) is now

$$c_k + e_k > 1, \qquad k \geqslant 1, \tag{3.13}$$

From the above the equivalent set of conditions [corresponding to (3.11)] is obtained [see (3.12)]:

$$\left.\begin{array}{lll} \tau_k + \sigma_{k-1} < 1 & \text{if} & \tau_k > 0, \quad \sigma_{k-1} < 0, \\ \tau_k - \sigma_{k-1} < 1 & \text{if} & \tau_k > 0, \quad \sigma_{k-1} > 0, \\ \tau_k + \sigma_{k-1} > -1 & \text{if} & \tau_k < 0, \quad \sigma_{k-1} > 0, \\ \tau_k - \sigma_{k-1} > -1 & \text{if} & \tau_k < 0, \quad \sigma_{k-1} < 0, \end{array}\right\} k \geqslant 1. \tag{3.14}$$

that is, the *off-diagonal* elements in the *rows* of the matrix T [that is, the pair (σ_{k-1}, τ_k)] must stay in the region of the (σ, τ) plane shaded in Figure 3.

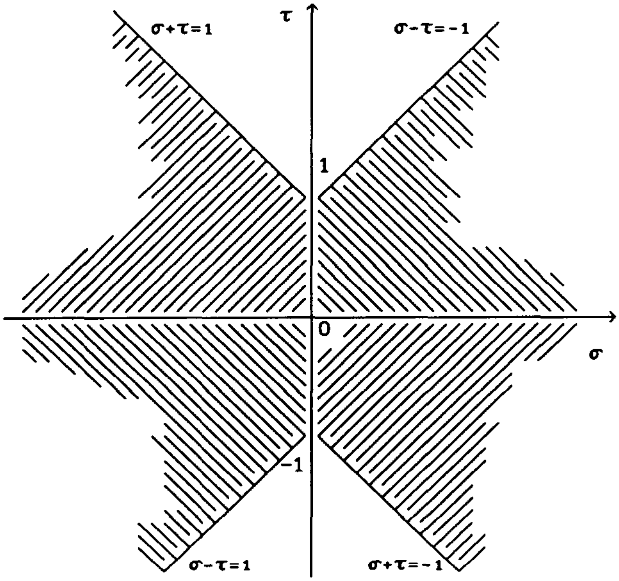

FIG. 3.

REMARK 3.2. Let us examine, more in detail, what happens when the inequalities (3.7) and (3.13) become equalities. We get

$$a_k + b_k = 1, \qquad k \geqslant 1, \tag{3.15}$$

$$c_k + e_k = 1, \qquad k \geqslant 1. \tag{3.16}$$

We shall study the former: a similar result holds for the latter. If (3.15) holds, the sequence (3.6) is still positive, and ω_i approaches a limit point as $i \to \infty$. In fact, by setting

$$\delta_i = \omega_i - \omega_{i-1},$$

we can obtain, from (3.6) and (3.15),

$$\delta_i = (a_i - 1)\omega_{i-1} + b_i\omega_{i-2} = -b_i\delta_{i-1}.$$

Since $0 < b_i = 1 - a_i < 1$, it follows that $\delta_i \to_{i\to\infty} 0$. This implies that

$$\omega_i \xrightarrow[i\to\infty]{} \omega_\infty = \text{constant}.$$

It turns out that $a_1 < \omega_i < 1$, and then, from (3.1.1), Theorem 3.1, and (3.2.1), we obtain the following result:

$$\|L^{-1}\|_\infty \leqslant 1 + (n-1)a_1^{-1} < na_1^{-1} = n|\sigma_1|.$$

This means that $\|L^{-1}\|_\infty \leqslant O(n)$. The same result holds obviously for $\|L^{-1}\|_1$. In a similar manner, from (3.16), it follows that $\|U^{-1}\| \leqslant O(n)$. In the worst case, when both (3.15) and (3.16) occur, it follows that $\kappa(T) \leqslant O(n^2)$.

REMARK 3.3. We have seen that (3.7) and (3.13) are sufficient conditions for the matrix T to be well conditioned. This means, following the definition, that $\kappa(T)$ is bounded from above by a quantity that is independent of its dimension n. We observe that this does not mean that such bound is small. In fact, if we consider the bound (3.10) (or the equivalent bound obtainable for$\|U^{-1}\|$), it is proportional to the quantity $2/(\underline{\gamma} - 1)$. From (3.5) and (3.8), it follows that

$$\underline{\gamma} = \min_i \{a_i + b_i\} = \min_i \left\{ \frac{1 + |\tau_{i-1}|}{|\sigma_i|} \right\}.$$

Even if the condition (3.7) [or equivalently (3.11)] is satisfied, nevertheless one has $\underset{\sim}{\gamma} \to 1$ if $|\sigma_i|, |\tau_{i-1}| \to \infty$. In this case, it follows that the integer k defined in Theorem 3.2 tends to ∞ [see (3.9)]. If the matrix T is normalized so that its entries are in the interval $[-1, 1]$, then $|\sigma_i|, |\tau_{i-1}| \to \infty$ is equivalent to saying that $T \to \mathscr{S}$, where, denoting nonzero entries by $*$,

$$\mathscr{S} = \begin{bmatrix} 0 & * & & & & & 0 \\ * & 0 & * & & & & \\ & * & 0 & \cdot & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & \cdot & \cdot & \cdot & \\ & & & & \cdot & \cdot & * \\ 0 & & & & & * & 0 \end{bmatrix}.$$

3.2. The Second Case: $\sigma_i \tau_i > 0$

This case is quite similar to the first one. For this reason we shall omit the proofs of the statements (see [1] for details). With the same notation of (3.5), from (2.2.3) the following second order equation is obtained:

$$\omega_k = a_k \omega_{k-1} - b_k \omega_{k-2}, \qquad k \geqslant 1,$$
$$\omega_0 = 1, \qquad \omega_{-1} = 0. \tag{3.17}$$

In this case a sufficient condition for the condition number of matrix L to be bounded is the following:

$$a_k - b_k > 1, \qquad k \geqslant 1. \tag{3.18}$$

LEMMA 3.2. *If the condition* (3.18) *is satisfied, the sequence defined by* (3.17) *is positive and monotone increasing.*

Define now $\underset{\sim}{\gamma} = \min_k \{a_k - b_k\}$. Observe that if (3.18) is satisfied, then $\underset{\sim}{\gamma} > 1$.

THEOREM 3.4. *With reference to the sequence defined by* (3.17), *if the condition* (3.18) *is satisfied, then*

$$\frac{\omega_i}{\omega_j} < \underset{\sim}{\gamma}^{i-j+1} \qquad \textit{for} \quad j > i.$$

From this result, it follows that $\|L^{-1}\| < \underset{\sim}{\gamma}/(\underset{\sim}{\gamma} - 1)$.

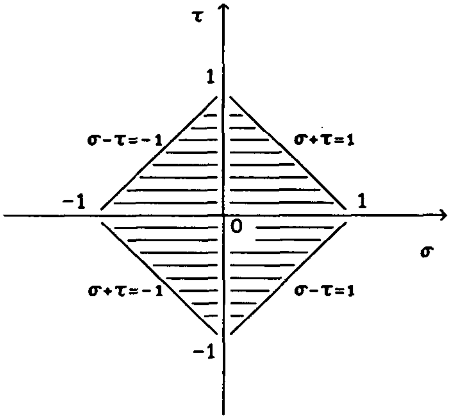

FIG. 4.

REMARK 3.4. One verifies that the condition (3.18) is equivalent to the following set:

$$\left.\begin{array}{lll} \sigma_k + \tau_{k-1} < 1 & \text{if} & \sigma_k > 0, \quad \tau_{k-1} > 0, \\ \sigma_k - \tau_{k-1} < 1 & \text{if} & \sigma_k > 0, \quad \tau_{k-1} < 0, \\ \sigma_k + \tau_{k-1} > -1 & \text{if} & \sigma_k < 0, \quad \tau_{k-1} < 0, \\ \sigma_k - \tau_{k-1} > -1 & \text{if} & \sigma_k < 0, \quad \tau_{k-1} > 0, \end{array}\right\} \quad k \geqslant 1. \tag{3.19}$$

These are equivalent to asking that the *off-diagonal* elements in the *columns* of the matrix T [that is, the pair (σ_k, τ_{k-1})] should be inside the region shaded in Figure 4.

A similar result holds for $\|U^{-1}\|$. One obtains that $\|U^{-1}\|$ is bounded from above by a quantity independent of its dimension if the following conditions are satisfied:

$$\left.\begin{array}{lll} \tau_k + \sigma_{k-1} < 1 & \text{if} & \tau_k > 0, \quad \sigma_{k-1} > 0 \\ \tau_k - \sigma_{k-1} < 1 & \text{if} & \tau_k > 0, \quad \sigma_{k-1} < 0 \\ \tau_k + \sigma_{k-1} > -1 & \text{if} & \tau_k < 0, \quad \sigma_{k-1} < 0 \\ \tau_k - \sigma_{k-1} > -1 & \text{if} & \tau_k < 0, \quad \sigma_{k-1} > 0 \end{array}\right\} \quad k \geqslant 1. \tag{3.20}$$

These are equivalent to the requirement that the *off-diagonal* elements in the *rows* of the matrix T [that is, the pair (σ_{k-1}, τ_k)] should be inside the region shaded in Figure 4.

REMARK 3.5. As seen in Remark 3.2, if instead of (3.18) we have

$$a_k - b_k = 1, \tag{3.21}$$

then $\|L^{-1}\| \leqslant O(n)$. A similar argument can be made for $\|U^{-1}\|$. Observe that if (3.21) is satisfied and T is symmetric, then necessarily we have $\|U^{-1}\| \leqslant O(n)$ and consequently $\kappa(T) \leqslant O(n^2)$.

4. A PARTICULAR CASE

A relevant case, common to many applications, is the one in which σ_i and τ_i have constant sign in the submatrices in which the products $\sigma_i \tau_i$ have constant sign. In this case we have seen that the matrix T defined by (2.1) is invertible if the condition (2.4) is satisfied. Observe that this condition is obviously satisfied if the following holds true:

$$|\sigma_i + \tau_i| < 1. \tag{4.1}$$

Moreover, the conditions (3.11), (3.14), (3.19), (3.20), which ensure the well-conditioning of the matrix, will be satisfied if the following two conditions hold true:

$$|\sigma_i + \tau_{i-1}| < 1, \qquad |\sigma_{i-1} + \tau_i| < 1. \tag{4.2}$$

This means that *all* the pairs $(\sigma_{i-1}, \tau_i), (\sigma_i, \tau_i), (\sigma_i, \tau_{i-1})$ are inside the strip in the (σ, τ) plane shaded in Figure 5. The three conditions (4.1), (4.2) are very simple to test, and provide a practical tool to test both invertibility and conditioning for the matrices examined in this section.

5. EXAMPLES

We report two examples of application of the results presented here. The former one is relative to a matrix which is not diagonally dominant, but nevertheless well conditioned. The latter derives from the discretization of a

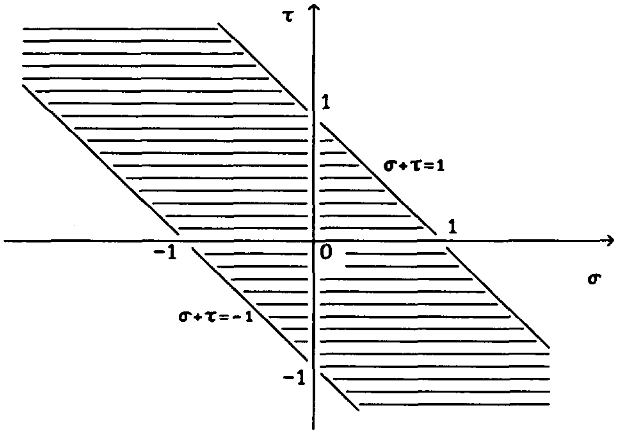

FIG. 5.

differential boundary value problem. By imposing the conditions presented here for the resulting tridiagonal matrix, a good numerical solution can be obtained.

5.1. *The First Example*

Consider the following tridiagonal matrix:

$$\begin{bmatrix}
1 & -0.9 \\
0.9 & 1 & -1.8 \\
 & 1.8 & 1 & -2.7 & & & & & & & & & & \text{\Large 0} \\
 & & 2.7 & 1 & -3.6 \\
 & & & 3.6 & 1 & -4.5 \\
 & & & & 4.5 & 1 & -5 \\
 & & & & & 5 & 1 & 5 \\
 & & & & & & -5 & 1 & -5 \\
 & & & & & & & 5 & 1 & 5 \\
 & & & & & & & & -5 & \ddots & \ddots \\
 & & & & & & & & & \ddots & \ddots & \ddots \\
 & & & & & & & & & & \ddots & \ddots & -5 \\
 & & & & & & & & & & & 5 & 1 & 5 \\
 & & \text{\Large 0} & & & & & & & & & & -5 & 1 & 4.5 \\
 & & & & & & & & & & & & & -4.5 & 1 & 3.6 \\
 & & & & & & & & & & & & & & -3.6 & 1 & 2.7 \\
 & & & & & & & & & & & & & & & -2.7 & 1 & 1.8 \\
 & & & & & & & & & & & & & & & & -1.8 & 1 & 0.9 \\
 & & & & & & & & & & & & & & & & & -0.9 & 1
\end{bmatrix}$$

This matrix is not diagonally dominant. Nevertheless, it satisfies the condi-

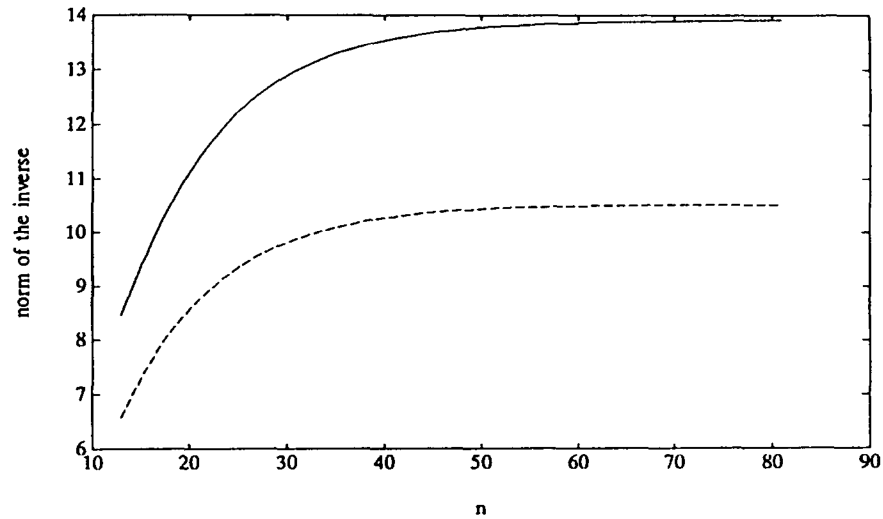

FIG. 6.

tion (2.4). It follows that it is invertible, and its factorization is bounded (see Theorem 2.1). Moreover, it is well conditioned: its *LDU* decomposition [see (2.2.1)–(2.2.3)] has the norms $\|L^{-1}\|$ and $\|U^{-1}\|$ bounded, because both conditions (3.11) and (3.14) are satisfied. So the boundedness of the condition number follows. In Figure 6 the $\|L^{-1}\|_1$ (solid line) and the $\|U^{-1}\|_1$ (dashed line) are plotted against the dimension n of the matrix.

5.2. *The Second Example*

This example derives from the discretization of the following singular perturbation differential boundary value problem:

$$\varepsilon y''(t) + y'(t) = 0, \qquad t \in [0, \alpha]$$

$$y(0) = 0, \qquad y(\alpha) = 1, \tag{5.1}$$

where $\alpha, \varepsilon > 0$. When ε is small, we have a stiff problem. From the discretization of the problem (5.1) with variable step size, a discrete boundary value problem derives, where

$$\sigma_i = \frac{-h_{i+1}(2\varepsilon - h_{i+1})}{(h_i + h_{i+1})(2\varepsilon + h_i - h_{i+1})}, \qquad \tau_i = \frac{-h_{i-1}(2\varepsilon + h_{i-1})}{(h_i + h_{i-1})(2\varepsilon + h_{i-1} - h_i)},$$

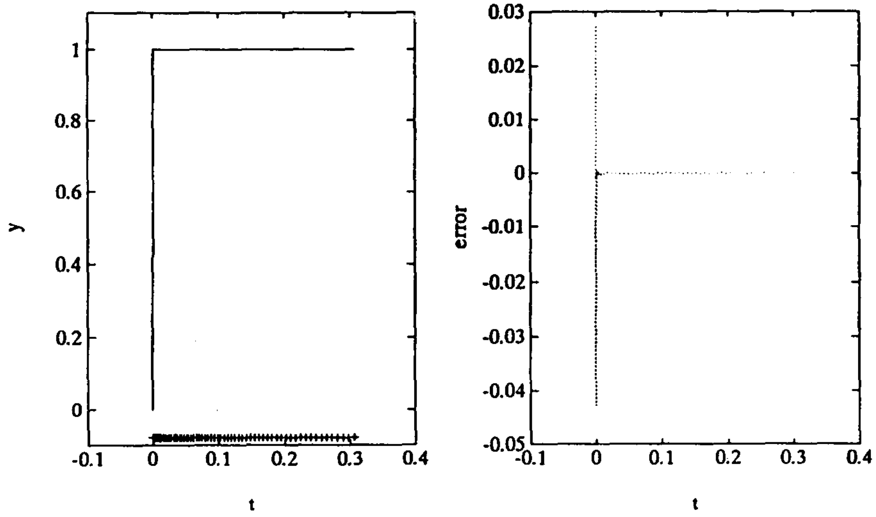

FIG. 7.

where h_i is the ith step of discretization. From the consistency we have $|\sigma_{i-1} + \tau_i| = 1$. We increase the step of discretization monotonically by at most 2ε at each step. In this way, the quantities in the denominator are positive. It follows that τ_i is negative, while σ_i is negative if $h_{i+1} < 2\varepsilon$, positive if $h_{i+1} > 2\varepsilon$. The resulting matrix is invertible, because the condition (2.4) turns out always to be satisfied. Moreover, if we increase the discretization step so that also

$$|\tau_{i-1} + \sigma_i| = 1,$$

then it follows, in accordance with Section 4, that the resulting tridiagonal matrix, far from being diagonally dominant, is at least weakly well conditioned. The resulting computed solution, as well as the error are shown in Figure 7 ($\varepsilon = 10^{-4}$, $\alpha = 0.3$).

REFERENCES

1 L. Brugnano and D. Trigiante, Tridiagonal Matrices: Invertibility and Conditioning, Rapporto Interno 1/91, Dipartimento di Matematica, Univ. Bari, Italy.

2 B. Buckberger and G. A. Emel'yanenko, Methods for inverting tridiagonal matrices, *U.S.S.R. Comput. Math. and Math. Phys.* 13:10–20 (1973).

3 M. Capovani, Sulla determinazione dell' inversa delle matrici tridiagonali e tridiagonali a blocchi, *Calcolo* 7:295–303 (1970).

4 G. Di Lena and D. Trigiante, Stability and spectral properties of incomplete factorization, *Japan J. Appl. Math.* 7(1):145–163 (1990).

5 C. F. Fisher and R. A. Usmani, Properties of some tridiagonal matrices and their application to boundary value problems, *SIAM J. Numer. Anal.* 6(1):127–142 (1969).

6 W. Gautschi, Computational aspects of three-term recurrence relations, *SIAM Rev.* 9(1):24–82 (1967).

7 S. Godounov and V. Riabenki, in *Schemas aux Differences*, MIR, Moscow, 1977, Chapter 2.

8 D. E. Heller, D. K. Stevenson, and J. F. Traub, Accelerated iterative methods for the solution of tridiagonal systems on parallel computers, *J. Assoc. Comput. Mach.* 23(4):636–654 (1976).

9 N. J. Higham, Efficient algorithms for computing the condition number of a tridiagonal matrix, *SIAM J. Sci. Statist. Comput.* 7(1):150–165 (1986).

10 V. Lakshmikantham and D. Trigiante, *Theory of Difference Equations: Numerical Methods and Applications*, Academic, New York, 1988.

11 J. W. Lewis, Inversion of tridiagonal matrices, *Numer. Math.* 38:333–345 (1982).

12 R. M. M. Mattheij and M. D. Smooke, Estimates for the inverse of tridiagonal matrices arising in boundary-value problems, *Linear Algebra Appl.* 73:33–57 (1986).

13 F. W. J. Olver, Numerical solutions of second-order linear difference equations. *J. Res. Nat. Bur. Standards Math. and Math. Phys.* 71B(2 and 3):111–129 (1967).

14 G. Piazza and D. Trigiante, Limiti per il numero di condizione di una classe di matrici tridiagonali e tridiagonali a blocchi, *Calcolo*, 1989.

15 R. A. Usmani, A note on the inversion of a band-matrix, *Indian J. Math.* 22(1):23–32 (1980).

Received 20 August 1990; final manuscript accepted 22 February 1991